An meine Eltern und meine Schwester

Abstract

Despite sophisticated present day automatic speech recognition (ASR) techniques, a single recognizer is usually incapable of accounting for the varying conditions in a typical natural environment. Higher robustness to a range of noise cases can potentially be achieved by combining the results of several recognizers operating in parallel.

One such approach is multi-band processing, mimicking parallel processing of frequency subbands in human speech recognition as had been claimed by Fletcher. However, recent findings in both human and automatic speech recognition have revealed insufficiencies, such as the assumption of independence between frequency subbands, of the original multi-band ASR approach which often leads to reduced performance in the case of clean speech and wide-band noise.

To overcome this problem, we propose and investigate a new set of "full combination" rules which integrate acoustic models trained on all possible combinations of subbands, preserving correlation information and leading to higher performance in all noise conditions. In this development, particular attention was given to the theoretical basis for all of the rules developed in terms of statistical theory, so that the assumptions that were necessary in each model become clear. The new combination strategies are developed for both posterior- and likelihood-based systems.

These new combination strategies are then also applied to the combination of diverse feature streams, for example derived from multi-time scale analysis, which results in better exploitation of the often used instantaneous and time difference features.

While combination may give the same weight to each expert, robustness of a multiple stream system can be further enhanced when each stream expert is assigned a weight reflecting its reliability. The new combination techniques are tested with several fixed and adaptive weighting strategies, including relative frequency of correct classification, least mean squared error, local signal-to-noise ratio, and maximum-likelihood based weights.

We will see how the new multi-band approaches, which are consistently trained in clean speech, outperform original multi-band ASR models in both clean and noisy speech. Multi-band processing improves over the baseline fullband recognizer only in the case of narrow-band noise. However, combining multiple data streams from different time scales, using the same "full combination" rules, has also shown to significantly improve over the baseline in wide-band factory noise.

Version Abrégée

Même si les techniques actuelles de reconnaissance automatique de la parole (RAP) sont très sophistiquées, un reconnaisseur isolé n'est en général pas capable de tenir compte des conditions très variables d'un environnement naturel typique. Un moyen d'obtenir une meilleure robustesse à une certaine gamme de bruits consiste à combiner les résultats de plusieurs reconnaisseurs travaillant en parallèle.

Le traitement multi-bandes se base sur cette idée. Il simule le traitement en parallèle de sous-bandes de fréquences qui, selon Fletcher, est effectué lors de la reconnaissance vocale humaine. Cependant, comme l'ont révélé de récentes découvertes dans le domaine de la reconnaissance humaine et automatique de la parole, cette approche présente certains défauts, notamment celui de considérer les sous-bandes de fréquences comme indépendantes les unes des autres. Cette erreur a souvent entraîné une déterioration des performances dans le cas de parole claire ou de bruit à large bande.

Pour remédier à ce problème, nous proposons un nouvel ensemble d'approches de "full combination" qui intègrent des modèles acoustiques entraînés sur toutes les combinaisons possibles de sous-bandes, préservant ainsi l'information de corrélation et conduisant à de meilleurs résultats dans toutes les conditions de bruit. Ces règles ont été élaborées dans le souci constant de respecter la théorie des approches statistiques. De cette façon, les hypothèses utilisées dans chaque modèle sont également bien mises en évidence. Les nouvelles stratégies de combinaison sont développées à la fois pour différents systèmes à base de probabilités a posteriori, et pour ceux qui s'appuient sur la vraisemblance. Par la suite, ces stratégies sont également appliquées à la combinaison de différents flux de paramètres caractéristiques dérivés, par exemple, d'une analyse à échelles temporelles multiples. Cette technique permet une meilleure exploitation des paramètres caractéristiques instantanés et différentiels couramment utilisés.

Il est possible de combiner les experts en donnant à chacun le même poids, mais on peut encore améliorer la robustesse d'un système multi-flux en attribuant à chaque expert un poids qui reflète sa fiabilité. Différentes techniques de combinaison sont donc testées en appliquant plusieurs stratégies de pondération fixes ou adaptatives, qui comprennent la fréquence relative de bonne classification, le critère de l'erreur quadratique moyenne, le rapport signal sur bruit local, et le maximum de vraisemblance des poids de combinaison.

Nous verrons comment les nouvelles approches multi-bandes, entraînées exclusivement sur de la parole claire, conduisent à des performances supérieures aux modèles multi-bandes classiques, et ceci aussi bien pour la parole claire que pour la parole bruitée. Toutefois, ce n'est que dans le cas de bruit à bande limitée que le traitement multi-bandes est plus performant que le système de référence travaillant sur le spectre entier. D'un autre côté, le fait de combiner des flux de données multiples extraits à partir de différentes échelles temporelles en utilisant les mêmes règles de "full combination" a également apporté un gain significatif par rapport à notre système de référence dans le cas de bruit d'usine à large bande.

Abstrakt

Trotz der heutzutage hochentwickelten automatischen Spracherkennungssysteme ist ein einzelner Spracherkenner oft nicht in der Lage, die störenden Einflüsse einer beständig wechselnden akustischen Umgebung zufriedenstellend zu kompensieren. Höhere Robustheit gegen eine Vielzahl von Störungen unterschiedlicher Charakteristik kann hingegen erzielt werden, wenn die Ergebnisse mehrerer, parallel arbeitender Erkenner kombiniert werden.

Ein Ansatz hierzu ist die Mehrband ("multi-band")-Verarbeitung, die die vermutete Eigenschaft der menschlichen Wahrnehmung, einzelne Frequenzbänder getrennt zu erkennen, simuliert. Jüngere Arbeiten auf den Gebieten der menschlichen und maschinellen Spracherkennung haben die Unzulänglichkeiten dieser ursprünglichen Mehrband-Verfahren aufgezeigt, so etwa die inkorrekte Annahme einer unabhängigen Informationsverarbeitung in verschiedenen Frequenzbändern. Es wird angenommen, dass dies der Grund ist für die oft verringerte Erkennungsleistung dieser Systeme sowohl bei der Verarbeitung ungestörter als auch verrauschter Sprachsignale.

Zur Überwindung dieses Problems wird eine Gruppe neuer Kombinationsregeln eingeführt, die auf dem "full combination"-Ansatz basieren. Dieser integriert die akustischen Modelle aller verfügbaren Kombinationen von Frequenzbändern, so dass sämtliche Korrelationsinformation zwischen den Frequenzbändern ausgeschöpft wird. Dies führt zu höheren Erkennungsraten in allen hier getesteten Rauschvorkommen. In der mathematischen Entwicklung dieser Kombinationsregeln wurde insbesondere acht gegeben auf die wahrscheinlichkeitstheoretische Grundlage aller entwickelten Regeln, so dass die mathematischen Bedingungen, die für jede respektive Regel zutreffen, hervortreten. Die neuen Kombinationsregeln werden sowohl für "a posteriori"-Wahrscheinlichkeiten als auch für Likelihoods entwickelt.

Die neuen "full combination"-Kombinationsregeln werden dann ebenfalls im Rahmen der Mehrkanal ("multi-stream")-Erkennung auf mehrere Merkmalsströme angewandt. Diese werden zum Beispiel durch die Verwendung mehrerer unterschiedlicher Analysezeitspannen ("multiple time scale analysis") gewonnen, was in Verbindung mit den neuen Regeln zu einer besseren Ausschöpfung der Kurzzeitmerkmale und Langzeit-Ableitungsmerkmale führt.

Kombinationsregeln können jedem akustischen Model die gleiche Gewichtung geben. Die Robustheit eines Mehrkanal-Systems kann jedoch gesteigert werden, wenn die Gewichtung dem Mass an Zuverlässigkeit entspricht, das einem Merkmalsstrom zugeordnet werden kann. Verschiedene statische sowie adaptive Gewichtungsverfahren werden daher entwickelt und im Rahmen der neuen Kombinationsregeln ausgewertet. Sie basieren auf relativen Häufigkeiten, dem kleinsten mittleren quadratischen Fehlerkriterium, lokaler Signal/Rausch-Schätzung, und dem "maximum likelihood"-Kriterium.

Es wird gezeigt, dass die neuen ("multi-band") Mehrband-Verfahren, die nur auf rauschfreien Sprachdaten trainiert werden, die früheren Mehrband-Verfahren an Erkennungsrate übertreffen auf rauschfreien sowohl als auch auf verrauschten Testsignalen. Nichtsdestotrotz ist Mehrband-Verarbeitung nur in bandbegrenztem Rauschen von Vorteil. Die Kombination mehrerer Datenkanäle ("multi-stream"), auf der anderen Seite, erweist sich auch auf Sprachsignalen mit (breitbandigem) Fabrikhallenlärm dem einzel-angewandten Erkenner überlegen, wenn das "full combination"-Verfahren eingesetzt wird.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Automatic speech recognition (ASR) promises to be a powerful means to ease human interaction with computers. For this to become true, automatic speech recognizers need to provide performance which is (almost) comparable to human performance for any application domain. Unfortunately, up to now, automatic speech recognizers are highly sensitive to both speaker and noise characteristics, degrading fast under mismatched training and testing conditions. Here, the measure of "training and testing mismatch" is not the physical difference between the training and testing data sets but the mismatch in how well the knowledge gained from the training set applies to the unknown testing set.

If the testing or application domain and with that the expected speaker or noise characteristics are known, an automatic speech recognizer can be trained for this application. Unfortunately, due to the nowadays increased use of mobile phones, which are often used as the interface to automated telecommunication services, and laptops, which provide for speech in- and output, the application environment is difficult to foresee and usually continuously changing. Moreover, although for some applications such as dictation systems and voice dialing, the system could be trained for a specific user and acoustic environment, such a solution is not satisfactory for most other applications. A lot of research is therefore dedicated to investigate how automatic speech recognizers can be rendered more robust to noise.

1.1 Our search for increased noise robustness in automatic speech recognition (ASR)

The different approaches to increase the robustness of an automatic speech recognizer comprise, among others, more appropriate feature extraction, better acoustic modeling and advanced decoding schemes. In this framework the goal of this thesis is to investigate and develop new paradigms for noise robust ASR based on multi-band and multi-stream processing. Although the latter is a generic term of the first, we distinguish these two approaches due to historical reasons: multi-band processing was – as far as our research is concerned – investigated first and its principle was then generalized to multi-stream processing.

In multi-band processing, the entire frequency domain is split into frequency subbands which are processed independently up to a certain point where the information from each band is recombined. In multi-stream processing, either the entire frequency domain is considered several times, each time employing different processing strategies, or other modalities, such as visual representations, of speech production are included. The information from each of these streams is correspondingly recombined later in the process. Both approaches try to better utilize the inherent redundancy in the speech signal either by processing different parts of the signal separately or by different processing of the same signal stream. If the streams are correlated, it can be assumed that combination is best carried out on the feature level so that dependencies between the streams can be modeled. In case when the streams are corrupted by noise, the correlation between the streams is decreased. It can thus be assumed that the streams are better modeled independently, as this is likely to result in independent errors conducted by each stream recognizer due to train/test mismatch. Nothing can be done about these errors when dealing with a single-stream (fullband) recognizer only. However, when combining the outputs of two or more recognizers, independent errors coming from any one of them can be dampened. Thus, the multi-band and multi-stream systems are expected to provide higher noise robustness to any kind of noise than a single-stream system, without any knowledge of the noise or the necessity of different training databases and noise adaptation phase.

In this thesis, we investigated several frame-level combination approaches, some of which employ a reliability term for each subband or stream (possibly different for each speech unit). Different ways to estimate these reliability factors will be proposed. The multi-band and multi-stream strategies are developed on clean speech data and their noise robustness is tested and evaluated on noise-corrupted speech with the noise stemming from various additive noise environments. The different multi-band and multi-stream recognizers are compared amongst each other as well as to the baseline fullband recognizers.

Our research is carried out in the framework of Hidden Markov Model (HMM) based speech recognizers, where HMM emission probabilities are estimated through either Gaussian Mixture Models (GMMs) or Artificial Neural Networks (ANNs). The former system will be referred to as HMM-GMM recognizer, the latter as HMM/ANN hybrid. An HMM-GMM system outputs likelihoods so that combination of different stream HMM-GMMs is carried out on these, whereas the ANN in an HMM/ANN system outputs posterior probabilities which are used for recombination in this case. After recombination, the posteriors are divided by the prior probabilities to obtain (scaled) likelihoods for the Viterbi decoder.

1.2 Goals of this thesis

In the framework of this thesis, we discuss two principle approaches to enhance noise robustness in an automatic speech recognizer. These are multi-band and multi-stream processing both relying on independent processing and recombination of individual streams. It will be shown how, in both approaches, the different streams can account for diverse mismatched conditions due to their inherently different processing strategies, and how the streams can complement each other at the combination stage.

Improvement of multi-band processing In *multi-band* processing, the speech signal in the spectral domain is split into several subbands which are processed independently for feature extraction and possibly probability estimation before they are recombined for further processing. In the case when noise only occurs in one frequency subband, it does therefore not mix with the other clean feature coefficients which allow for reliable decoding of the clean part of the speech. Similarly, in *missing data* (MD) processing as applied to robust ASR, it is tried to segregate the different sound sources, such as speech and noise, in the input signal, and then to recognize at each time frame the clean speech part only. This includes the necessity for a noise detection algorithm and for the processing of continuously varying combinations of (clean) feature coefficients. Moreover, only one fixed decomposition into clean and noisy data (a so-called "MD mask") is considered at each time frame.

Original subband processing misses important frequency correlation information among subbands. We develop in this thesis, an approach to subband processing which provides a solution to the problem of both loss of frequency correlation in multi-band processing and fixed MD masks through a revised decomposition of the frequency band into an exhaustive and mutual exclusive set of frequency subbands. This induces new combination strategies as described below.

Investigating multi-stream processing In *multi-stream* processing, different possibilities exist to incorporate additional knowledge sources. They can stem, among others, from different data recordings (such as audio and visual streams), pre-processing, feature extraction, or from a different choice, structure and training of the classifiers. In this thesis, we concentrate on the use of different *feature* streams, from either different feature extraction techniques or the same technique but employing different parameters and/or pre- or post-processing strategies. Thus, the same (fullband) frequency domain undergoes different processing strategies leading to different feature representations which are used in individual recognizers, the errors of which are hoped to be complementary. The streams are recombined, just as in the multi-band approach, later in the process to dampen the errors.

Improved classifier combination strategies For both approaches, multi-band and multi-stream, the correct choice of features as well as *combination strategy* play an essential role for the performance and robustness of the system.

We therefore investigate the advantages and disadvantages of several combination strategies. One set of new combination strategies arises from the extension of the original subband approach as described above, based on the integration of all possible subband combinations. These are equally applicable to both multi-band and multi-stream processing. Other combination strategies are motivated from research on human speech recognition, such as the "product of errors rule" and "error correction in posteriors combination". Where appropriate, all approaches are derived for both posterior-based and likelihood-based systems.

Stream weighting according to stream reliability Due to the inherent characteristics of each (subband or fullband) input stream and the changing environmental conditions during the application of a speech recognizer, at a given point in time, some streams are more reliable than others. *Reliability* in this context signifies the trust we can have in the stream recognizer's output when it is applied under unknown and possibly mismatched condition. Depending on its reliability, a stream recognizer should receive more or less *weight* in the combination procedure in order to render the overall result as reliable as possible.

The combination strategies either naturally include weights due to their mathematical derivation or can artificially be enriched by appropriate weights which account for the reliability of each constituent. We will see that the weights can depend on (i) the stream index, (ii) the speech unit (in our case the phoneme), (iii) the local data and/or (iv) any combination of the former.

Different strategies for their estimation can be found in the literature which either employ offline training of the weights or online adaptation during recognition. In the case of matched training and testing conditions (of both the stream recognizers and the weights), the former approach is capable of sufficiently reflecting the performance of each recognizer. In the case of mismatched application, the latter approach is expected to more appropriately account for unknown conditions, but usually also demands more complicated algorithms with possibly slower performance due to increase in computational needs.

In this framework, new offline and online weighting schemes will be presented. The former are based on least mean square error and relative frequency estimations. The latter employ signal-to-noise ratio measurement and maximum likelihood estimation.

Experimental evaluations The proposed algorithms for combining multiple subband or fullband streams, together with the different weighting strategies, are tested on a continuously spoken digits database under noise-free (matched) conditions and under noise-corruption by artificial band-limited and natural wide-band noise (mismatch). As our goal was to develop systems which can easily generalize and adapt to unseen data, training is only carried out on clean speech.

1.3 Structure of thesis

In Chapter 1 the main goals of this thesis were presented, namely investigation and improvement of multi-band and multi-stream processing together with the search for improved classifier combination and reliability weighting strategies.

In this thesis, we pursue the task of speaker independent, spontaneous telephone speech recognition in clean *and* noisy environments through the use of multiple streams which are processed in parallel, laying specific emphasis on *training of the systems in clean speech only* and *unknown application to speech corrupted by various different additive noise cases*.

In Chapter 2 a short introduction to human speech processing is given. Early psychoacoustic findings which motivated multi-band processing in ASR are presented, some of which have meanwhile been revised. The more recent models are presented next and their influence on multi-band processing are discussed.

In Chapter 3 the necessary background for ASR will be given for the general case of single-stream fullband processing. This involves presentation of statistical pattern classification before coming to statistical sequence recognition by Hidden Markov Models (HMMs) as applied to ASR. The recognizer which is mainly used in this thesis is a combination of HMMs and Artificial Neural Networks (ANNs), which is refered to as HMM/ANN hybrid or, in the specific case when the neural net is a Multi-Layer Perceptron (MLP), as in this thesis, as HMM/MLP hybrid. It is described in the last section of Chapter 3.

A wide range of approaches to enhance noise robustness in automatic speech recognizers exist. In Chapter 4 we discuss the most widely used and most promising strategies, involving robust feature processing, robust modeling, and the missing data approach. Multi-band and multi-stream processing constitute another set of approaches to enhance noise robustness through parallel processing of multiple complementary information streams.

In Chapter 5, the paradigm of multi-band processing is presented in its general form, before we come to the description of previous multi-band research and its limitations. As early multi-band processing did not take into account correlation information between subbands we present, in the framework of this thesis, new approaches which circumvent this short-coming.

An essential part in stream combination is the recombination module. Different strategies to probability combination have been investigated and are presented in Chapter 6. Where appropriate, their mathematical formulae are developed for both posterior- and likelihood-based systems.

Through the use of multiple complementary streams and an appropriate combination scheme, multi-band and multi-stream approaches can achieve improved noise robustness as compared to a single fullband-based recognizer. This can even be enhanced through the use of reliability weights in the combination process. In Chapter 7 different reliability weighting strategies are presented which have been investigated in this thesis.

In Chapter 8 the new approaches to multi-band processing are evaluated experimentally on clean and noise corrupted data sets of continuously spoken digits. The newly developed combination strategies as well as reliability weights are compared to standard combination schemes and to fullband baseline systems. Most systems are HMM/MLP hybrids, however, for the implementation of one of the weighting schemes we had to use HMM-GMMs.

Chapter 9 addresses multi-stream ASR. After illustration of the general approach to multi-stream processing, its specific realization in this thesis is described, motivated from both recent psychoacoustic findings and engineering reasons. Within this framework, the different streams employed in this thesis consist of single- and multi-time scale feature streams. State of the art research to multi-stream processing is also discussed.

In Chapter 10 the experiments to our work on multi-stream processing using HMM/MLP hybrid systems are presented, employing the different proposed feature streams, combination strategies and reliability weights. The same test bed as for the multi-band experiments is used.

Chapter 11 summarizes the research pursued in this thesis, presents the conclusions which can be drawn, and gives some ideas for future directions in this account.

Chapter 2

Background to human speech processing (HSP)

Several developments presented in this thesis are inspired from human hearing properties.

In this chapter, we therefore give a short introduction to human speech processing. A well-known model for assessing the functions of the human speech production apparatus by time-invariant filters is introduced. We then turn to human speech perception, discussing some of the most prevailing characteristics of the human auditory system which influenced, to a certain extent, the development of specific automatic pre-processing units.

We then discuss the psychoacoustic motivations which led to multi-band processing. These stem from Fletcher's investigations on human auditory processing and his assumption of independent auditory bands. We then come to newer models which actually demonstrate the insufficiency of Fletcher's approach and show the necessity to revise both Fletcher's model and the multi-band approach based on it.

2.1 Human speech production modeling

The articulatory mechanism of speech production is often modeled with the so-called source-filter-model as introduced by Fant (1960). In this model, the vocal tract is represented by a time-varying filter, the source energy of which is the excitation signal. The glottis produces a sound of many frequencies, and the vocal tract filters a subset of these frequencies for radiation from the mouth. In the model, the excitation signal is approximated by either a generator of periodic impulses for the creation of voiced sounds, or by a generator for white noise for the production of unvoiced sounds in the excitation signal. By adjusting the degree the two generators are involved in the production of a certain sound, mixed sources of excitation or no excitation at all (for a speech pause) can also be modeled. In the notation of the "z-transform" (Kunt, 1996), the signal produced in this way by one or more generators $U(z)$ is then passed on to a linear filter that represents the filter function of the vocal tract $V(z)$. It enhances the respective resonance frequencies according to whatever sound is about to be generated and, with radiation at the lips $R(z)$, gives out the speech signal $F(z)$. The discrete signal can be represented in the z-domain (Schukat-Talamzzini, 1995, p. 33) simply by

$$F(z) = U(z) \cdot V(z) \cdot R(z) \tag{2.1}$$

This complete transfer function of speech production[1] is for practical reasons often approximated by only one filter $H(z) \simeq \frac{1}{A(z)}$, with the polynomial in z^{-1} $A(z) = \sum_{i=0}^{M} a_i z^{-i}$ and $a_0 = 1$, (which converges to a *finite impulse response* (FIR) filter). The speech production model can then be written as:

$$F(z) = E(z) \cdot H(z) = E(z) \cdot \frac{1}{A(z)} \tag{2.2}$$

with $E(z)$ the z-transform of the excitation signal and $F(z)$ the z-transform of the produced speech signal in the frequency domain. Speech signal $F(z)$ represents the output at the speaker's mouth, directly after production. Linear systems of this form are referred to as *all-pole* or *autoregressive models*. In Chapter 4, we will see various factors which act upon the signal $F(z)$ before it reaches the listener or other (non-human) receivers.

The source-filter model forms the basis for many of the most standard feature extractors, such as e.g. linear prediction (LP) modeling and cepstral analysis (Rabiner and Juang, 1993).

2.2 Human speech perception

In this section, we present some important characteristics of human speech perception, such as the concept of critical bands, subjective pitch and the perception of loudness. Moreover, the parallel processing which is conducted in human perception will be described. These phenomena found in human speech processing were incorporated in to ASR processing units in order to approximate human performance. Feature extraction based on knowledge drawn from the human auditory system will be discussed in Chapter 4.

Critical bands In the inner ear, inside the cochlea, sound waves cause the basilar membrane to vibrate up and down. A certain sound gives rise to a traveling wave on the basilar membrane (von Bekesy, 1960). The distance the wave travels before it reaches its peak amplitude is a direct function of the frequency of the sound. Because tones of different frequency give rise to maximal vibration amplitudes at different locations along the basilar membrane, the spectral components of a complex sound are separated along the basilar membrane according to frequency. The basilar membrane can therefore be seen as a spectrum analyzer characterized by *critical bands*. A critical band represents a certain frequency range, outside of which subjective responses, such as loudness, change abruptly[2] (Rabiner and Juang, 1993). Based on these human critical bands, a perceptually based frequency unit was created, the *Bark* scale, to link the absolute frequency of a sound and the frequency resolution of the ear in terms of critical bands. A Bark covers the frequency range of a critical band, increasing logarithmically with frequency.

[1] This actually only applies to voiced sounds (Schukat-Talamzzini, 1995).

[2] The loudness of a band of noise at a constant sound pressure remains constant as the bandwidth of the noise increases up to the width of the critical band; after the range of the critical band has been surpassed, increased loudness is perceived.

Subjective pitch Another important subjective criterion of the perception of frequency content is the fact that frequency is not perceived linearly. These findings, stemming from psychophysical experiments, have led to the introduction of another but similar scale, the *Mel* scale, which defines, for each tone with an actual frequency f (measured in Hz) its *subjective pitch*. As a point of reference, the pitch of a 1 kHz tone is thereby defined as 1000 Mels. By adjusting the frequency of a tone such that it is half or twice the perceived pitch of a reference tone, other subjective pitch values were determined (Rabiner and Juang, 1993).

Both perceptually based frequency scales, the *Bark* and the *Mel* scale, are widely used in pre-processing of speech signals as we see in Chapter 4. They provide higher noise robustness than linearly spaced frequency scales.

Power law of hearing Hearing sensation increases logarithmically as the intensity of the stimulus increases. The perception of intensity is usually referred to as *loudness*. With the help of auditory experiments it was found that loudness L is approximately proportional to the cube root of intensity I: $L = I^{0.3}$, with a doubling of loudness being observed for a 6 to 10 dB increase in sound pressure levels (SPL) (Moore, 1997). This rule is called the *power law of hearing*.

Equal loudness contours The human ear is not equally sensitive at all frequencies, being more sensitive to sounds between 2 and 5 kHz and less sensitive at higher and lower frequencies. (Thus, the range of 0 to 8 kHz is usually assumed sufficient for human speech perception, whereas for music a wider range of up to 16 kHz is required). Thus, although the human ear collects sounds ranging from 16-20 kHz (Schukat-Talamzzini, 1995, p. 38), it amplifies the 2-5 kHz frequency range where much of the important speech information registers. The sensitivity to different frequencies is more pronounced at low SPLs than at high SPLs which can be illustrated by so-called *equal loudness contours* (Moore, 1997). The contours tend to become flatter for high loudness levels.

Parallel and hierarchical organization The central auditory system is organized hierarchically, with acoustic information being transferred and processed progressively from one center to the next. This hierarchical organization is superimposed upon a parallel organization in the form of separate, parallel channels connecting the various levels. Auditory information is segregated across several parallel pathways which are represented by distinct populations of neurons, exhibiting different response properties to acoustic stimuli. It is argued that the different neurons process different attributes of the acoustic information.

We only briefly illustrated some characteristics of human hearing but they already give insight into the possibilities which exist to model human characteristics in automatic processing. We see in later chapters how they were brought to bear in feature extraction techniques in automatic speech recognition systems, such as MFCC, PLP, and J-RASTA-PLP processing, and significantly increase their levels of performance. Moreover, the description of human perception as processing heterogeneous information in parallel streams can be used to motivate similar approaches in ASR, such as multi-stream processing employing diverse information streams.

2.3 Psychoacoustic motivation to multi-band ASR

We concentrate on the illustration of important psychoacoustic findings which motivated the use of multiple frequency subbands in automatic speech processing, synthesis and coding (Bourlard et al., 1996b; Hermansky et al., 1996; Goldberg and Riek, 2000).

One of the most important investigations which can be seen as laying the ground for multi-band automatic processing, is Fletcher's study on human speech recognition, carried out between 1920 and 1950. His motivation was to measure the quality of the perception of speech sounds in telephony to improve telephone speech intelligibility. In 1994, Allen (Allen, 1994) summarized this work recalling its findings and making them more accessible to the speech research community. For any work on multi-band processing, Fletcher's results need to be well understood and are, thus — based on (Allen, 1994) — discussed in detail in the following.

2.3.1 On Fletcher's independent 'articulation' bands

First experiments in (Fletcher, 1953) were carried out on normal conversational speech over a (modified) telephone channel. Fletcher already found that human speech recognition (HSR) is strongly dependent on the effects of semantic context, confounding the measurement of phone errors and therefore complicating intelligibility testing and increasing variability of the results. He thus excluded context from his subsequent experiments by changing to the use of nonsense Consonant-Vowel-Consonant (CVC) syllables, such as "yif" or "moush".

Fletcher used the term *'articulation'* to denote the empirical probability of correct recognition of sounds having *no* context, such as the nonsense syllables used, and *'recognition'* for the empirical probability of correct recognition of sounds having context, such as words. For many different speaker-listener pairs, the 'articulation' was varied by (i) changing the signal-to-noise ratio (SNR) of the speech signal, and (ii) low-pass and high-pass filtering the speech. After the listeners had noted what they had heard, the error probabilities for the consonant $(1-c)$ and vowel $(1-v)$ sounds, respectively, were computed. An average CVC-phone 'articulation' s was calculated from the 'articulation' of all C's and V's ($s = \frac{2c+v}{3}$) along with an estimate of the CVC syllable 'articulation' score S, assuming independent C and V units. For perfect conditions, i.e. very high SNR and no filtering, the average 'articulation' was 98.5%.

Fletcher found that CVC syllable 'articulation' S (i.e. the probability of correct identification of a CVC syllable) is well predicted from the phone 'articulations' c and v by the relation

$$\begin{aligned} S &= c^2 v \\ &\simeq s^3 \end{aligned} \tag{2.3}$$

This shows that the three sound units are perceived as independent sounds which means that for correct identification of the syllable all three sound units must be correctly identified. However, we have to bear in mind that this is only true for nonsense CVC's (which have maximum entropy (Allen, 1994, p. 571)) as, for a meaningful word, context decreases the entropy. A similar finding was more recently achieved by Bronkhorst et al. (1993) who also found that under conditions of low noise, phones are perceived independently.

Experiments with low- and high-pass filtering The newly discovered fundamental importance of the phone 'articulation' to HSR led to the question of how humans decode phones. For this, Fletcher studied phone 'articulation' s (i.e. context independent phone recognition) for various frequency filters and noise conditions by low-pass L and high-pass H filtering the speech. He found that the partial 'articulations' (i.e. the 'articulations' s_L and s_H of the two bands) did not sum to the wide band 'articulation' s.

After looking for a non-linear transformation A of the partial 'articulations' which would make them additive, i.e. $A(s_L(f_c,\alpha))+A(s_H(f_c,\alpha)) = A(s(\alpha))$ with f_c being the high/low-pass cut-off frequency and α the speech gain (used to vary the SNR), Fletcher found empirically $A(s)$, which he called the *articulation index* (AI):

$$A(s) \simeq \frac{\log_{10}(1-s)}{\log_{10}(1-s_{max})} \tag{2.4}$$

where constant $s_{max} = 0.985$ is the maximum 'articulation' and $\varepsilon_{min} = 1 - s_{max} = 0.015$ is the corresponding minimum error. Since this is dealing with transformations of probabilities, the additivity condition corresponds to a (statistical) independence assumption. Solving (2.4) for s one gets wide band 'articulation' $s(A) = 1 - \varepsilon_{min}^A$, and with this $\varepsilon(A) = \varepsilon_{min}^A$, which describes the error probability of the phone as the minimal error ε_{min} to the power of the articulation index.

According to Fletcher (1953), the articulation index accurately characterizes speech intelligibility under conditions of frequency filtering and noise masking, and the AI can be interpreted as a fundamental internal variable of speech recognition.

2.3.2 Fletcher's product of errors rule

For the two band example of high- and low-pass filtered speech it can be followed from (2.4) and its additivity assumption

$$\log(1-s) = \log(1-s_L) + \log(1-s_H) \tag{2.5}$$

which becomes

$$\begin{aligned} 1-s &= (1-s_L)\cdot(1-s_H) \\ \varepsilon &= \varepsilon_L \cdot \varepsilon_H \end{aligned} \tag{2.6}$$

using the error $s = 1 - \varepsilon$ in (2.6). He found that this term was true for any value of the cut-off frequency f_c.

Equation (2.6) can be interpreted as stating that 'articulation' errors in the low-pass filtered band ε_L are independent of the 'articulation' errors in the high-pass filtered band ε_H and vice versa. It means that only if there is an error in *both* subbands, the whole phone is wrongly recognized. This is equivalent to stating that the overall recognition is correct, if *any* subband is correct. Fletcher thus thought to have shown that the phones are processed in independent frequency channels, which he called *articulation bands*, and that these independent estimates of a phone in each frequency band are merged in an "optimal" way.

Allen (1994) expressed the above to mean that we are listening to independent sets of phone features in the two bands and are processing them independently, up to the point where they are fused to produce the phone estimates. Allen states that (2.6) "*may be generalized to a multichannel articulation band model*" (Allen, 1994, p. 572) of K channels, where K is the number of independent 'articulation' bands:

$$\varepsilon = \varepsilon_1 \varepsilon_2 \ldots \varepsilon_K \tag{2.7}$$

which is called *Fletcher-Stewart multi-independent channel model of phone perception*, following Allen (1994) who writes that it was first proposed by Stewart but developed by Fletcher. Following Bourlard (1999) and for ease of notation we simply refer to it as (Fletcher's) *product of errors rule*. Note that this model does not state how to identify the band in which recognition is correct, as it is just intended as a description of human recognition performance.

The idea of independent articulation bands, which correspond to frequency subbands, which are processed separately up to a certain point where their information is joint for phone estimates led to the multi-band approach in ASR. In ASR, however, the information from each subband is commonly represented by recognition probabilities rather than error probabilities. The multi-band approach will be discussed in Chapter 5.

Following Bourlard (1999), we discuss in Section 6.4 how the "multi-independent channel model", that is the "product of errors rule", could also be realized in a speech recognizer, under the obviously wrong assumption of independent and correct recognizers.

Local SNR dependency Subsequent research has shown (Green et al., 1991) that the k^{th} band 'articulation' error ε_k, is determined by the local SNR (normalized to 30 dB) in each critical band

$$\varepsilon_k = (\varepsilon_{min})^{\frac{1}{K}\frac{\mathrm{SNR}_k}{30}} \tag{2.8}$$

This relationship shows that the AI is determined by the SNR in each band and not by the band energy, i.e. the local contribution of each band basically depends on the local SNR_k. The AI can then directly be written as:

$$A(s) \quad = \quad \frac{1}{K} \sum_{k=1}^{K} \frac{\mathrm{SNR}_k}{30} \tag{2.9}$$

In later works, the equally important frequency bands were substituted by 1/3- or 1/1-octave bands which made the use of band-specific weighting factors necessary to render distributions of each octave band to speech intelligibility equally important. In automatic multi-band processing, a similar approach to weight the different subbands according to their inherent SNR can be employed to render recognition more reliable. This is discussed in Section 7.4.1.

Fletcher's studies motivated other scientists to investigate human intelligibility performance under different tasks.

Miller and Nicely (1955), for example, carried out listening experiments on 16 English consonants, spoken over voice communication systems with low-pass and high-pass filtering as well as random noise. Their results also showed that human speech recognition on narrow frequency bands was high. Miller and Nicely were especially interested in confusion between sounds. They found that low-pass filtering and noise masking led to consistent patterns of confusions, leaving phonemes audible, while errors due to high-pass filtering seem to occur randomly as consonants lose their intelligibility. Grouping the 16 consonant sounds into a set of 5 articulatory features (such as voicing, nasality, etc.) showed that the perception of each one of these features appears to be relatively independent of the perception of the others. They argue that breaking down the measure of transmission into an estimation of each of the 5 features separately corresponds to considering 5 different communication channels. If all channels really were independent, the sum of the transmitted information in each channel would equal the information transmitted by the whole channel. As articulatory features are not independent and the sum of all transmissions results in a higher value than transmission of the whole channel, it was concluded that this is an indicator of *redundancy* in the speech signal.

The argument of inherent redundancy in the speech signal further motivated multi-band processing, as it suggests that there is sufficient information present in each of the subbands to warrant a certain level of independent processing.

2.4 Recent psychoacoustic findings

New results from research on HSR illustrate the insufficiency of Fletcher's AI formula described in Section 2.3.1, when applied to more realistic test conditions. In these experiments, which had been carried out in these works on HSR, not only low- and high-pass filtering but also band-pass filtering of speech was investigated. This led to important results showing that humans integrate frequency information from non-contiguous bands which, as is seen below, often results in higher robustness than the use of adjacent bands of the same size. We will see in Section 5.6 how these findings can be employed to improve the usually used multi-band approach as applied in automatic speech recognition.

2.4.1 Discussion of the AI and subband independence assumption

Speech transmission index An extension of the articulation index, which was introduced in Section 2.3.1, is the so-called *speech transmission index* (STI) which takes into account distortions in the time domain (Houtgast and Steeneken, 1985) and non-linear distortions (Steeneken and Houtgast, 1980). The STI was proposed by Steeneken and Houtgast for the measurement of speech transmission quality. For consistency with Section 2.3.1, we nevertheless only consider degradation in the frequency domain, such as filtering. The band-specific parameter is again the SNR which is modeled by the *transmission index* (TI_k). Using seven octave bands, Steeneken and Houtgast define the STI as

$$\mathrm{STI} = \sum_{k=1}^{K} \alpha_k \cdot \mathrm{TI}_k \tag{2.10}$$

with $K=7$ and α_k being the importance-weighting of each octave band ($\sum_k \alpha_k = 1$). Just as for the AI, the STI results from a (weighted) summation over frequency bands, and thus implicitly assumes independence between frequency bands. However, in a later investigation, Houtgast and Verhave (1991) found that the energy contents in neighboring bands *can* be correlated for the case of continuous speech. Steeneken and Houtgast (1999) argue that (instantaneous) speech levels in adjacent bands show a high degree of co-variation so that the information from these bands could be correlated and even redundant. As the original experiments leading to the development of the STI only included contiguous bands, no effect of correlation had been observed. However, practical experience with the STI, which also included non-contiguous bands showed that intelligibility was underestimated by the STI for this kind of band-pass filtered speech. Missing contributions of the gap in a non-contiguous band did not result in as low intelligibility as estimated by the AI. Redundancy between bands was assumed to account for this (Grant and Braida, 1991; Steeneken and Houtgast, 1999).

To investigate the discrepancy, Steeneken and Houtgast (1999) carried out more experiments in which the octave bands which were included in the pass-band and their SNRs were varied. The database comprised nonsense CVC-words only. The STI is calculated with the help of the transmission index TI_k which is defined as follows:

$$\mathrm{TI}_k = \frac{\mathrm{SNR}_k + 15}{30} \tag{2.11}$$

with $0 \leq \mathrm{TI}_k \leq 1$. For the three selected SNR-values of 15, 7.5 and 0 dB the TI_k thus had the values 1.0, 0.75 and 0.5, respectively. Comparing the results between the sets of bands with gaps and the sets of contiguous bands showed that the former resulted in relatively low STI whereas the latter resulted in a relatively high STI value in relation to the CVC-word score. This was attributed to overestimation of the total information content for bands without gaps. It was therefore proposed that a reduced contribution from adjacent octave bands through a redundancy factor could correct the discrepancy, leading to a revised version of the STI :

$$\mathrm{STI}_r = \alpha_1 \mathrm{TI}_1 - \beta_1 \sqrt{(\mathrm{TI}_1 \mathrm{TI}_2)} + \alpha_2 \mathrm{TI}_2 - \beta_2 \sqrt{(\mathrm{TI}_2 \mathrm{TI}_3)} + \ldots + \alpha_K \mathrm{TI}_K \tag{2.12}$$

where K, α_k and TI_k are defined as in (2.10) and $\sum_{k=1}^{K} \alpha_k - \sum_{k=1}^{K-1} \beta_k = 1$. (Taking the root of the TI_k factors is not essential but makes the terms in the expression more uniform). The effect of the redundancy correction depends on the values of β_k and the simultaneous contribution to the information content by the two adjacent frequency bands TI_k and TI_{k+1}, which is given by their product. Equation (2.12) takes account of the redundancy between the six neighboring octave bands only, although it could be extended in a simple manner to also consider redundancy introduced by non-adjacent bands, with a corresponding increase in the number of parameters.

The new STI_r[3] model was applied to earlier experiments using band-pass filtering in noise producing adjacent octave bands only. Here, only a small improvement could be gained with the new model, indicating that the original STI model was well suited for these conditions of pass-bands without gaps. In the experiments on non-adjacent bands, the new model, which takes into account the correlation between neighboring bands, resulted in a more accurate prediction of intelligibility scores than the former, additive STI model. The frequency-weighting factors α_k and the redundancy-correction factors β_k, which were set up over various sub-sets of the data, resulted each time in similar values.

Research by Grant and Braida Grant and Braida (1991) also addressed the question regarding the additivity assumption of the information in frequency subbands and, with that, the independence assumption of the bands. According to their analysis, the deviation from

[3] The subscript r indicates the *revised* STI model.

the additivity assumption could be due to an overlap of the slopes of the filters which produce the non-adjacent bands. On the other hand, they argue that closely-spaced frequency bands could have been affected by self-masking. Finally, they also mention possible high correlation between neighboring frequency bands.

The aforementioned investigations indicate that *high correlation* and *redundancy* exist in the speech signal. The assumption of independent frequency bands is therefore no longer sustainable. For the processing of frequency subbands in ASR, we thus have to ensure that we appropriately account for the correlation and redundancy in the speech signal. A solution is provided in Section 5.6 through a new model to multi-band processing where correlation information between all (contiguous and non-contiguous) bands is considered.

2.4.2 Importance of non-contiguous frequency bands

Experiments on speech limited by frequency other than only low- and high-pass filtering were carried out in (Lippmann, 1996) and (Silipo et al., 1999). As we will see in this subsection, good perception of such stop-band filtered speech sustain above findings on the importance of non-contiguous frequency bands.

Investigating mid-frequencies Complementary to Fletcher's work, who only studied high-pass and low-pass filtering, Lippmann (1996) investigated the removal of mid-frequency speech energy, simulating human hearing loss. The purpose of the study was to explore the importance of high-frequency speech energy for consonant perception when mid-frequencies (from 800 Hz to 4 kHz first, then raised to 8 kHz) are missing. Results show that speech energy at high frequencies is most important when mid-frequency speech energy is not available. The additional use of high-frequency speech energy with low-pass filtered speech increased recognition accuracy by almost 30 %. In similar experiments (French and Steinberg, 1947; Kryter, 1962) where only the low-pass cut-off frequency was extended from 4 to 8 kHz, an overall gain of only 10 % had been obtained. Moreover, Lippmann's results show that listeners can integrate acoustic cues from widely disparate frequency bands.

Intelligibility of combined channels Silipo et al. (1999) investigated whether detailed auditory analysis of the short-term acoustic spectrum is required to understand spoken language. For this, the spectrum was split into 4 1/3-octave channels, which they called "slits" (each well separated by one octave from its neighbor(s)), and the intelligibility of each channel by itself and in combination with up to 3 others was measured. It was found that human word recognition remained high when 2 or 3 channels were presented simultaneously (60-83%) although intelligibility of each channel by itself was less than 9%. The two center slits resulted in the highest intelligibility, while the more distant slits were unable to profit from each other when combined. Still, channel proximity did not always result in higher intelligibility. The authors also point out that the intelligibility of their slits was much higher than that predicted by the AI, which again suggests that revision of the AI is warranted.

The above experimental results from HSR suggest that

- the information found in different (even dispersed) frequency regions is not uncorrelated,
- and that humans make good use of this correlation information and redundancy in the speech signal.

It was especially emphasized by Lippmann, and Silipo et al. that a combination of high- and low-frequencies (including a gap in frequency) often resulted in better recognition performance than an increase of the low-frequencies by the use of a higher cut-off frequency (thus without gap). These findings suggest that the multi-band approach which has been employed so far and which does neither consider any combinations of subbands nor non-contiguous subbands does not fully account for the way humans process frequency information.

We therefore propose in Section 5.6 a new approach to multi-band processing where not only the individual subbands but also all combinations of subbands (including combinations with and without gap) are employed in order to avoid excluding the joint information carried by any combination of subbands, and also to more correctly model human speech processing.

2.5 Summary

Findings on the way the human auditory system functions are sometimes used in automatic speech processing, such as in some feature extractors and in the multi-band approach, to render the automatic speech recognizer more robust to noise. For a better understanding of these human-based processing steps, we presented some basic knowledge about human speech processing and its modeling.

We illustrated Fant's well-known source-filter model for human speech production, on which several successful feature extraction techniques are based.

Perceptual experiments with humans have shown that humans appear to process speech in separate, "critical bands", and that frequency is not perceived linearly. This led to the introduction of the Bark and Mel scales. Following, the power law of hearing and the "equal loudness contours", which describe the phenomenon of increased human sensitivity to a certain frequency range, were discussed. Moreover, human perception was described to process heterogeneous information hierarchically and in parallel streams. We will see in following chapters how these characteristics can be implemented in automatic speech processing units to render them more noise robust.

Fletcher's concept of independent subband processing in human speech perception was illustrated, which is one of the main original motivations for multi-band processing in ASR. Fletcher's product of errors rule has, however, not yet been implemented and tested which will be done in the framework of this thesis.

We then discussed the speech transmission index (STI), which is an extension to the articulation index (AI), and measures the quality of transmitted speech. The STI has recently been revised to better model more realistic application conditions than only high- and low-pass filtered speech, such as stop-band filtered speech. The new-found importance of non-contiguous frequency bands with gaps is also sustained by other research. These new results confirm the need to revise the usual approach to multi-band processing in such a way as to avoid the assumption that subbands can be processed independently. A new approach will hence be introduced in Chapter 5, the "full combination" approach to subband processing. It is based on a more consistent theoretical analysis of subband expert combination and on the recent findings on human auditory processing.

Chapter 3

Useful background to statistical pattern classification and automatic speech recognition

The usual approach to ASR is to extract, at each time step, one feature vector from the entire speech spectrum and to pass it through the acoustic model which calculates probability estimates for each speech class. These probabilities are then used in decoding to find the most likely sequence of words.

In the approaches pursued in this thesis, several feature vectors are extracted in parallel, either from frequency subbands for multi-band processing, or from the entire speech spectrum for multi-stream processing. They are then treated as separate information streams and passed on to different acoustic models. The stream probability estimates are recombined before decoding.

In this chapter, we present the main features and underlying hypotheses of Hidden Markov Model based ASR for the usual approach of one-stream (fullband) processing, but we will bare in mind that the same applies to each of the separate streams in a multi-band or a multi-stream system. As in this thesis, probability combination is carried out after each time frame, decoding of the combined probability estimates can also be conducted as in a standard speech recognizer.

3.1 Structure of an automatic speech recognizer

A speech recognizer can be described as a chain of individual modules with one module producing the input to the next. At each processing step the speech signal is transformed into a new representation possibly with the help of additional, external knowledge sources. The goal of most recognizers is to produce at its output an orthographic transcription of the recorded utterance. In some cases, instead of an orthographic transcription, a machine-internal representation of the recognized utterance is better suited for further processing such as translation or speech synthesis. In this thesis, we are interested in the correct orthographic transcription at the word level of the recorded speech sequence.

Based on **Figure 3.1**, we give an overview of the most important parts employed in any automatic speech recognizer.

Discretization In computer-based ASR, after the speech signal has been recorded, the real-valued, continuous waveform needs to be transformed into an appropriate format for digital processing, which first implies the use of an analog/digital (A/D) converter. Here, the speech signal is sampled at equidistant points in time, and its amplitude is quantized. Following Shannon's sampling theorem (Shannon and Weaver, 1949), the speech signal has to be sampled at a minimum of twice the maximal band-width to guarantee the possibility for reconstruction of the continuous waveform from its sample values. In the example of telephone speech which has a band-width of around 200 or 300 to 3200 or 3400 Hz (Rabiner and Juang, 1993, p. 308), (Schukat-Talamzzini, 1995, p. 47)[1], the sampling frequency is thus usually chosen at 8 kHz. The originally continuous speech signal is now discretized in both frequency and time and thus representable in a digital computer.

Short-term analysis Due to inertia of the human articulators, the speech signal does not change too rapidly over time and can therefore be assumed short-term stationary in short segments of 5 to 30 ms (Schukat-Talamzzini, 1995). These segments, which are obtained through application of a windowing function, which only cuts out the interesting part of the signal, are used to extract, at intervals of about 10 to 15 ms in time (so-called frames), characteristic features from the signal. For this, the windowed time signal is usually converted to the spectral or cepstral[2] domain to enhance the robustness of the acoustic features. In general, the features should be designed in such a way as to allow for good discrimination between the speech units by encoding the *content* of the speech signal rather than speaker intrinsic characteristics as needed in speaker verification or identification. In some cases, acoustic feature extraction reduces the amount of information to be stored and processed by the computer.

Feature extraction and pre-processing Different feature extraction techniques have been developed over time. Later on, knowledge gained from human auditory processing was incorporated in order to render performance of automatic speech recognizers closer to human-like performance. This is often achieved by the use of non-linearities similar to those found in the human auditory system, for example cube root compression to simulate the power law of hearing as described in Section 2.2. In Section 4.2, we discuss several of the most commonly used feature extraction techniques which are especially appropriate for recognition in noise due to (different stages of) pre-processing, which are based on human auditory processing, employed during feature extraction. Extraction of acoustic feature vectors is intended to provide a first means to handle interfering noise from various sources and to derive acoustic representations of the speech signal which are well suited to differentiate between the different speech sounds as well as suppress irrelevant sources of variation.

[1] Different authors mention slightly different values in this respect.

[2] To obtain the cepstrum, the inverse Fourier transform is taken of the logarithmic spectral magnitude.

Density estimation Most of the speech recognizers are nowadays based on the theory of *Hidden Markov Models* (HMMs). Each speech unit is hereby modeled by one HMM. Such speech units can be whole words or phonemes, depending on the size of the vocabulary which is to be modeled. In the latter case, words are constructed, during recognition, from sequences of phonemes, and sentences from sequences of words. An HMM assumes that each speech unit can be modeled as a sequence of (static) acoustic vector segments, where each segment is represented by the parameters of some invariable statistical function, which was fixed beforehand. Generally, this density is supposed to be Gaussian and its parameters are estimated on a representative sample space (the *training* sample space). If this training set is segmented in terms of phonemes or states, training of the parameters is easily achievable. Usually, such a segmentation is not given and powerful training algorithms are needed which either segment the data iteratively or train the parameters without explicit segmentation. The trained model parameters are then used during recognition to estimate the likelihood that an acoustic vector has been produced by a certain HMM state.

Decoding in the testing phase The densities for each frame, together with the *dictionary* and language model or *grammar* are passed onto the decoder. The dictionary defines for each word in the database its constituent units, and the language model describes the connections of the words which are linguistically possible within a language. It is the decoder which then finds the best path (i.e. the best state sequence) through the search space of speech units, deciding for the most likely path of all possibilities. The most commonly used algorithm for decoding is the Viterbi algorithm (Rabiner and Juang, 1993).

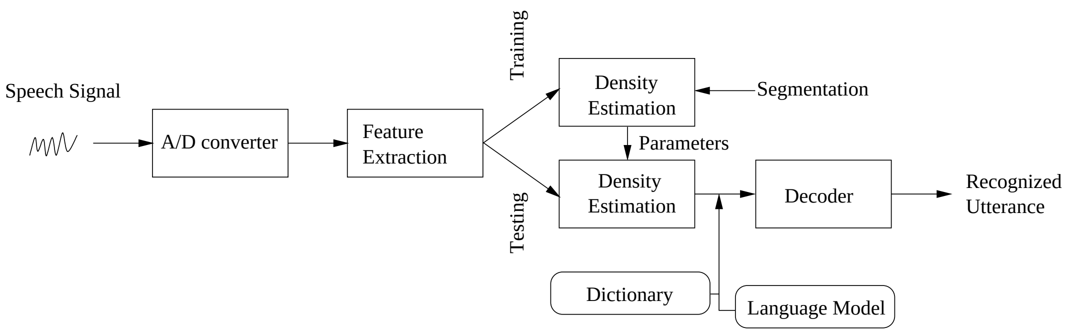

Figure 3.1: Illustration of a standard speech recognizer with training and testing phase.

3.2 Background to information theory

In this section, we give a short introduction to terms stemming from information theory which will be needed during the development of this thesis. The formulae are based on (Jones, 1979; Cover and Thomas, 1991; Applebaum, 1996).

Entropy and conditional entropy The *entropy* is a measure of the average uncertainty of a random variable, and depends only on the probabilities of the components x_i of the vector random variable x (Papoulis, 1991). For the discrete case, the entropy is defined as

$$H(x) = -\sum_i P(x_i) \log_2 P(x_i) \tag{3.1}$$

where $P(x_i)$ is the probability of x_i.

The smallest value of the entropy is 0 and occurs when $P(x_i)$ has a sharp peak ($P(x_i) = 1$) for one value of x_i and zero for the rest. The largest entropy value arises when all x_i have the same probability. This can also be interpreted in such a way that, in the first case, no information is conveyed as only exactly one value can occur, i.e. there is no uncertainty. In the second case, though, all values of the random variable are equally probable, hence leading to the largest value of uncertainty.

The *conditional entropy* is the entropy of a random variable given another random variable.

$$H(y|x) = -\sum_{i,j} P(x_i, y_j) \log_2 P(y_j|x_i) \tag{3.2}$$

with $P(x_i, y_j)$ the joint probability function of x_i and y_j, and $P(y_j|x_i) > 0$. It is a measure of the uncertainty that is still felt about y after it is known that x has occurred but without knowing which value it has taken.

Mutual information Now $H(y|x)$ is a measure of the information content of y which is not contained in x; thus, the information content of y which is contained in x is $H(y) - H(y|x)$. This is called the *mutual information* (MI) of x and y, and is denoted by $I(x, y)$, so that

$$I(x, y) = H(y) - H(y|x) \tag{3.3}$$

$$= \sum_{i,j} P(x_i, y_j) \log_2 \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \tag{3.4}$$

with $P(x_i), P(y_j)$ the marginal probabilities. It is the reduction in uncertainty of one random variable due to the knowledge of another random variable and, hence, can also be described as a measure of the dependence between the two random variables.

MI is a special case of the more general quantity of *relative entropy* $D(P_1||P_2)$, also refered to as Kullback-Leibler divergence, which measures the asymmetric "distance" between two distributions $P_1(x)$ and $P_2(x)$

$$D(P_1||P_2) = \sum_i P_1(x_i) \log_2 \frac{P_1(x_i)}{P_2(x_i)} \tag{3.5}$$

With this, the mutual information amounts to a relative entropy between the joint probability $P(x, y)$ and the product distribution $P(x) \cdot P(y)$.

Above formulae are only given for the discrete case. For the continuous case, it simply suffices to substitute the sum operator by an integral in each formula respectively.

3.3 Statistical pattern classification

Statistical classification in the framework of automatic speech recognition could be seen, in a very simplified way, as the automatic categorization of the acoustic feature vectors by assigning exactly one class label to each feature vector. These classes can either be concrete speech units or clusters in the feature space. The classification is then carried out with the help of models which represent the different speech units (or clusters). In practice, we do not want to distinguish single speech units only, but we aim at recognizing entire speech utterances. As the task would be too complex to create a model for each possible speech utterance, other more refined techniques are needed, which will be described in the next section.

The classifier, that is the mathematical function for the assignment, is not available *a priori* but has to be constructed with the help of the statistical information given by the feature vectors of the training sample space.

In the following, we outline the statistical classifier whose decision rule is known to fulfill the classification task in an *optimal*[3] way, as far as the classification error rate is concerned.

3.3.1 Optimal Bayes' classifier

In ASR, the training space consists of a set of (high-dimensional) feature vector examples which have been extracted from the training speech signals, and denoted $X = \{x_1, \ldots, x_T\}$ (with T the number of time frames obtained from feature extraction). The probability that a pattern, i.e. a feature vector, belongs to a class ω_k ($k = 1, \ldots, K$) is denoted $P(\omega_k)$ (with $0 \leq P(\omega_k) \leq 1$), the prior class probability[4]. If a class probability is conditioned on the feature vector x, this conditional probability is denoted $P(\omega_k|x)$, the posterior probability. The optimal Bayes' classifier now states that a feature vector x can be optimally assigned to class k providing the minimum probability of error, if class membership is decided according to the following rule

$$P(\omega_k|x) > P(\omega_j|x), \qquad \forall\, j = 1, \ldots, K,\ j \neq k \tag{3.6}$$

that is, the pattern has to be assigned to the class which has the highest posterior probability. This optimum strategy is often called *Bayes' decision rule* or *MAP* (Maximum A Posteriori) *criterion*.

Generally, the posterior probabilities cannot be calculated directly, and can only be estimated from the training data. Commonly, this estimation is simplified by making some assumptions about the distribution of the data, such as describing them by some parametrized model. Defining the form of this parametrized model, usually separately for each class, we are left with the estimation of the parameters of this model (for each class). (The trained parameters Θ will then be used during classification to estimate the required probabilities). Most commonly, a probability density function (pdf) $p(x|\omega_k)$ is used which is connected to posterior probabilities via *Bayes' rule*, so that maximizing of the posterior probability amounts to:

$$\arg\max_k P(\omega_k|x) = \arg\max_k \frac{p(x|\omega_k)P(\omega_k)}{p(x)} = \arg\max_k p(x|\omega_k)P(\omega_k) \tag{3.7}$$

since $p(x)$ is constant for all classes. We thus see that in order to maximize the posterior we want to find the largest likelihood, which is refered to as *Maximum Likelihood* (ML) *criterion*.

[3] "optimal" in the formal sense of "Bayes optimal" and subject to certain conditions which will be introduced below.

[4] In all of the work reported here, it is assumed that every x must belong to one of the classes ω_k. This is known as "forced choice". Under some conditions, this assumption is invalid, for example, when some sounds may come from a different language, or be due to extraneous noise, such as coughs or typing sounds.

This way of approximating posterior probabilities, though, has several disadvantages. First, some assumptions are required about the form of the parametric model of $p(x|\omega_k)$. In most systems, the density functions or *likelihoods* $p(x|\omega_k)$ are estimated using the model of a Gaussian distribution, which is introduced below. Another drawback is that the prior probabilities needed to convert the likelihoods back into posterior probabilities are usually hard to estimate. Finally, the fact that the likelihoods are usually trained separately for each class results in poor discrimination between the models.

In a subsequent section we will see how the posterior probabilities can also be estimated directly which involves discriminant training. In principle, this will require the training of parameters which are influenced by all of the input vectors, which can increase complexity.

3.3.2 Density estimation by GMMs

The most common choice for modeling the pdf $p(x|\omega_k)$ are (multi-dimensional) Gaussian functions. As one Gaussian function is usually not sufficient to appropriately model the distribution of the acoustic feature vectors of a given class a weighted sum of Gaussian functions is considered instead, which is refered to as a *Gaussian Mixture Model* (GMM). In order to restrict the number of parameters in this model, it is often assumed that the different components of a d-dimensional acoustic vector are independent. The covariance matrices of the Gaussian distributions are then diagonal.

A simple univariate Gaussian model is given by

$$p(x|\omega_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp^{-\frac{1}{2}\frac{(x-\mu_k)^2}{\sigma_k^2}} \tag{3.8}$$

where μ_k $(k = 1, \ldots, K)$ is the mean and σ_k the standard deviation associated with class ω_k. Assuming diagonal covariance matrix and writing σ_{ki}^2 for the elements on the diagonal, the pdf for a multivariate Gaussian distribution reduces to

$$p(x|\omega_k) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\sigma_{ki}^2}} \exp^{-\frac{1}{2}\frac{(x_i-\mu_{ki})^2}{\sigma_{ki}^2}} \tag{3.9}$$

where μ_{ki} denotes the i^{th} component of μ_k, and d the dimension of the acoustic vector.

While the Gaussian function is unimodal, any distribution can be approximated with a *Gaussian Mixture Model* (with a sufficient number of mixture components) which is defined by

$$p(x|\omega_k) = \sum_{l=1}^{M} P(m_l|\omega_k)\, p(x|m_l, \omega_k) \tag{3.10}$$

where $P(m_l|\omega_k)$ is the mixture weight of mixture component m_l $(l = 1, \ldots, M)$ and class ω_k, and $p(x|m_l, \omega_k)$ the component Gaussian distribution. The weights $P(m_l|\omega_k)$ are restricted to be larger or equal to zero, and to sum to to one ($\sum_l P(m_l|\omega_k) = 1$).

Given we have M Gaussian distributions per class ω_k $(k = 1, \ldots, K)$ the number of parameters of the whole set of models amounts to $(2dM + M)K$.

EM algorithm for GMM parameter estimation There is no analytical solution to estimate the model parameters, i.e. the means, variances and mixture weights, from the training data. An optimization scheme which can be used to find these parameters for each Gaussian model of the GMM according to the ML objective is the Expectation Maximization (EM) algorithm (Dempster et al., 1977). In this approach, the parameter estimation problem is structured to incorporate hidden variables, which represent information that is not directly observable but that is assumed to be part of the model that generated the data. In the case of GMMs, such hidden variables could for example be the index of the Gaussian which generated the data.

As we will see below, maximizing the likelihood corresponds to maximizing the expected value of the logarithm of the joint density between the known and these hidden components (Dempster et al., 1977; Bilmes, 1997). The prominent idea of the EM algorithm is thus to maximize this expected value by updating the parameters which are used in the probability estimation. After random initialization of the parameters, the calculation of the expected value and the re-estimation of the parameters to maximize this expectation are iterated to converge. As the EM algorithm is used in Chapter 7 to estimate the weights of similar mixture models, we will discuss the algorithm here in little more detail.

Let us first show the correspondence between data likelihood maximization and maximization of the expected value (Bilmes, 1997; Boite et al., 2000; Gold and Morgan, 2000). For this, the expectation of the joint likelihood of the observed and hidden variables, usually refered to as auxiliary function, is expressed as a function of old parameters $\hat{\Theta}$ and new parameters Θ. For random (hidden) variable m_l, observed random variable x and parameters Θ, let

$$\begin{aligned} A(\Theta, \hat{\Theta}) &= \sum_{l=1}^{M} P(m_l|x, \hat{\Theta}) \log p(m_l, x|\Theta) && (3.11) \\ &= \sum_{l=1}^{M} P(m_l|x, \hat{\Theta}) \log(P(m_l|x, \Theta)p(x|\Theta)) && (3.12) \\ &= \sum_{l=1}^{M} P(m_l|x, \hat{\Theta}) \log P(m_l|x, \Theta) + \log p(x|\Theta) \sum_{l=1}^{M} P(m_l|x, \hat{\Theta}) && (3.13) \\ &= \sum_{l=1}^{M} P(m_l|x, \hat{\Theta}) \log P(m_l|x, \Theta) + \log p(x|\Theta) && (3.14) \end{aligned}$$

Choosing the new parameters to be equal to the old parameters and subtracting this new expression $A(\hat{\Theta}, \hat{\Theta})$ from (3.14) one gets

$$\log p(x|\Theta) - \log p(x|\hat{\Theta}) = A(\Theta, \hat{\Theta}) - A(\hat{\Theta}, \hat{\Theta}) + \sum_{l=1}^{M} P(m_l|x, \hat{\Theta}) \log \frac{P(m_l|x, \hat{\Theta})}{P(m_l|x, \Theta)} \quad (3.15)$$

The last term is a relative entropy which can be shown to be non-negative (Blahut, 1990). Thus, if a change in the parameters increases A, the (logarithm of) the data likelihood $\log p(x|\Theta)$ also increases.

To illustrate the EM algorithm we assume here that a set of (independent and identically distributed) vectors $\{x_1, \ldots, x_t, \ldots, x_T\}$ is modeled by a mixture of M densities. An unknown probability density $p(x|\Theta)$ can always be decomposed as

$$p(x|\Theta) = \sum_{l=1}^{M} P(m_l|\Theta)p(x|m_l, \Theta) \quad (3.16)$$

assuming M disjoint categories. As typically done for ML training, we try to approximate the true densities by maximizing the likelihood $p(x|\Theta)$ or its logarithm. Considering m_l as the hidden random variable and knowing from (3.15) that an increase in the expectation of $\log p(x, m_l|\Theta)$ will also increase the data likelihood $p(x|\Theta)$ we can write the expected value over the T samples and M mixture components as

$$A(\Theta, \hat{\Theta}) = E[\log p(x, m_l|\Theta)] = \sum_{l=1}^{M} \sum_{t=1}^{T} P(m_l|x_t, \hat{\Theta}) \log[P(m_l|\Theta) p(x_t|m_l, \Theta)] \tag{3.17}$$

The old parameters $\hat{\Theta}$ describe the parameters which were used to generate the distribution with which the expected value will be evaluated, and the new parameters Θ are to be optimized. We can decompose (3.17) as

$$A(\Theta, \hat{\Theta}) = \sum_{l=1}^{M} \sum_{t=1}^{T} P(m_l|x_t, \hat{\Theta}) \log P(m_l|\Theta) + \sum_{l=1}^{M} \sum_{t=1}^{T} P(m_l|x_t, \hat{\Theta}) \log p(x_t|m_l, \Theta) \tag{3.18}$$

Assuming the subset of parameters in Θ of the two terms is disjoint, the terms can be optimized separately. We can continue by assuming that $p(x_t|m_l, \Theta) = p(x_t|\omega_k)$ is a simple Gaussian (with diagonal covariance matrix and assuming x_t is scalar) such as (3.8), and that $P(m_l)$ is the weight given to Gaussian m_l in the model. With this we can write (3.18) as

$$A(\Theta, \hat{\Theta}) = \sum_{l=1}^{M} \sum_{t=1}^{T} P(m_l|x_t, \hat{\Theta}) \log P(m_l|\Theta) + \sum_{l=1}^{M} \sum_{t=1}^{T} P(m_l|x_t, \hat{\Theta}) \left[-\log \sigma_l - \frac{(x_t - \mu_l)^2}{2\sigma_l^2} + c \right] \tag{3.19}$$

with c a constant which will disappear in the following differentiations. To optimize this expression for the means, we set the partial derivatives to zero ($\frac{\partial A}{\partial \mu_j} = 0$), getting

$$\sum_{t=1}^{T} P(m_j|x_t, \hat{\Theta}) \left(\frac{x_t}{\sigma_j^2} - \frac{\mu_j}{\sigma_j^2}\right) = 0 \tag{3.20}$$

and thus

$$\mu_j = \frac{\sum_{t=1}^{T} P(m_j|x_t, \hat{\Theta}) x_t}{\sum_{t=1}^{T} P(m_j|x_t, \hat{\Theta})} \tag{3.21}$$

Similarly, the optimum value for the variances can be derived

$$\sigma_j^2 = \frac{\sum_{t=1}^{T} P(m_j|x_t, \hat{\Theta})(x_t - \mu_j)^2}{\sum_{t=1}^{T} P(m_j|x_t, \hat{\Theta})} \tag{3.22}$$

For the calculation of the mixture weights update formula, (3.19) must be supplemented with a Lagrange term $\lambda(\sum_{l=1}^{M} P(m_l|\Theta) - 1)$, as the mixture weights are probabilities which must sum to one. Taking the partial derivative of the augmented A function and setting it equal to zero, we get

$$\frac{1}{P(m_j|\Theta)} \sum_{t=1}^{T} P(m_j|x_t, \hat{\Theta}) + \lambda = 0 \tag{3.23}$$

as the terms involving the means and variances can be disregarded. Summing (3.23) over all components of m_j yields $\lambda = -T$, so that the weights can be expressed as

$$P(m_j|\Theta) = \frac{1}{T} \sum_{t=1}^{T} P(m_j|x_t, \hat{\Theta}) \tag{3.24}$$

Expression $P(m_j|x_t, \hat{\Theta})$ which is needed in each of the update formulae can be calculated by Bayes' rule, as

$$P(m_j|x_t, \hat{\Theta}) = \frac{p(x_t|m_j, \hat{\Theta})P(m_j|\hat{\Theta})}{\sum_{l=1}^{M} p(x_t|m_l, \hat{\Theta})P(m_l|\hat{\Theta})} \tag{3.25}$$

which can be evaluated from the terms we have access to, that is (3.8) and (3.24), with the parameters from the previous optimization step.

Summarizing the EM After a functional approximation, such as a GMM, has been chosen for the densities associated with each class, the EM algorithm in the present context can be summarized as

- Initialize the parameters.
- Given these parameters, compute the estimates of the posterior probabilities for the hidden variables, i.e. $P(m_j|x_t, \hat{\Theta})$.
- With these posterior estimates, find the parameters that maximize the expected value of the joint density for the data and the hidden variable. After iteration, these values converge to give a local maximum likelihood for the observed data.

We will see in Chapter 7 how this algorithm can easily be used to adapt the weights in a multiple stream system where each stream employs a set of GMMs for stream likelihood estimation.

3.3.3 Discriminant probability estimation by ANN

In order to avoid some of the assumptions which are necessary to estimate the likelihoods $p(x|\omega_k)$, the posterior probabilities $P(\omega_k|x)$ in (3.7) can also be estimated directly. In (Bourlard and Morgan, 1994), *Artificial Neural Networks* (ANNs) have been proposed to approximate the unknown distribution of the data. The main difference between GMMs and ANNs is that ANNs can actually be trained to directly estimate the posteriors.

One of the most widely used kind of ANN employed for statistical classification in the framework of automatic speech recognition are *Multi-Layer Perceptrons* (MLP) (Ripley, 1996).

MLPs are feed-forward networks of an input layer, zero or more hidden layers, and an output layer, each consisting of one or more units (also called neurons). A general structure of an MLP can be seen in **Figure 3.2**. Each layer of an MLP is usually only connected to the previous layer, and the connections between the units of one layer and that of the previous layer are defined by the *weights*. The weights are estimated during training of the MLP. The output of each unit is defined as the weighted sum of its inputs from the previous layer passed through a differentiable non-linear transfer function:

$$y_j^l = f(\sum_i w_{ij}^{l-1,l} y_i^{l-1}) \tag{3.26}$$

where $w_{ij}^{l-1,l}$ is the weight from unit i in layer $(l-1)$ to unit j in layer l, y_i^{l-1} is the output of unit i in layer $(l-1)$ and f is typically a sigmoid transfer function $f(y_i) = \frac{1}{1+\exp(-y_i)}$.

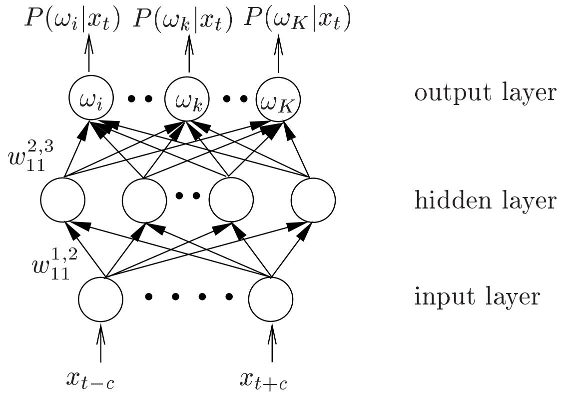

Figure 3.2: Illustration of a Multi-Layer Perceptron (MLP) with one input layer (with $(2c)+1$ time frames as input vector), one hidden layer and an output layer giving posterior probability estimates for each class ω_k ($k = 1, \ldots, K$). Each layer l is fully connected to its preceding layer $l-1$ with the weights $w_{ij}^{l-1,l}$ between unit i (in layer $l-1$) and unit j (in layer l).

In the case of automatic speech recognition, an MLP is trained to associate an acoustic input vector (which can consist of several frames of speech) with a desired output vector (i.e. the "target" output). The output vector indicates the correct speech unit class corresponding to the input and thus consists of zeros all over except for the correct class, which is indicated with a one (supervised training). Training is usually accomplished via the error back propagation algorithm which is described below (Bishop, 1995; Ripley, 1996).

Gradient descent in multi-layer networks In order to evaluate the appropriateness of the network outputs, an error function can be defined which measures the distance of the current output to the target output. Such an error function is a smooth function of the weight parameters, and can be minimized in a variety of ways. In case of the "sum of squared errors" function, for example, and a linear network, the solution for the weight values can be found in a closed form by differentiating of the error function with respect to the weights. If a non-linear activation function, such as a sigmoid is used (as in our case), and if, thus, the network is no longer linear, or if a different error function is considered, a closed form solution is no longer possible.

However, if the activation function is differentiable, the derivatives of the error function (with respect to the weights) can be evaluated. These derivatives can then be used in a gradient-based optimization algorithm to find the minimum of the error function, and with that good values for the weight parameters. One such algorithm is e.g. *gradient descent*, which functions as follows.

Given a differentiable error function E, the weights are first initialized at random. Following, the weight vector is updated by moving a small distance in the weight-space into the direction in which the error decreases most rapidly, i.e. the derivative has the largest negative value. This process is iterated, generating a sequence of weight vectors $\{w_{ij}(1), .., w_{ij}(n), .., w_{ij}(N)\}$

(connecting unit i to unit j) according to

$$w_{ij}(n+1) = w_{ij}(n) - \eta \frac{\partial E}{\partial w_{ij}(n)} \tag{3.27}$$

with η being a small positive number, called the learning rate. The *Error Back-Propagation* (EBP) algorithm (Rumelhart et al., 1986) first computes the output vectors for all training input vectors and calculates the error function E. It then allows to calculate the gradient of the error E versus every weight in the network by recursively back-propagating the error at the output layer and updating the weights according to (3.27). Under suitable conditions, the sequence of weight vectors will converge to a solution locally minimizing E. The choice of η can be critical, since if it is too small, convergence is very slow, whereas if it is too large, the correction process will overshoot and lead to divergence.

Statistical inference in MLPs For a proof that the outputs of MLPs (or ANNs in general) in classification can be interpreted as estimates of posterior probabilities of output classes conditioned on the input the reader is refered to (Bourlard and Wellekens, 1989; Richard and Lippmann, 1991; Bourlard and Morgan, 1994, 1997). These proofs are based on the following four conditions:

- The neural nets are sufficiently large, i.e. contain enough parameters, so that it can be trained to a good approximation of the mapping function between the input and the output class.
- The error criterion for gradient training is either the mean-squared difference between outputs and targets, or else the relative entropy between the outputs and the targets.
- The neural nets are trained to a global error minimum (i.e. avoiding local optima).
- The system is trained in the classification mode; that is, for K classes the target is one for the correct class and zero for all the others.

To ensure that the outputs of the MLP approximate posterior probabilities, the *softmax* function $f(y_i) = \frac{e^{y_i}}{\sum_{k=1}^{K} e^{y_k}}$ is often used (instead of the usual sigmoid) to normalize the output. To get the target outputs which are needed for network training, the training data needs to be labeled for each time frame. Such a segmentation can be obtained through iterative Viterbi alignment of the training data, starting at a linear initialization (cf. Section 3.4.1).

3.4 Statistical sequence recognition by Hidden Markov Models (HMMs)

In the last section we saw how in static classification a feature vector can be assigned to one of the different classes (in our case speech units) by acoustic models of a certain distributional form, the parameters of which are trained on the training data. Unfortunately, we are not interested in recognizing a sequence of speech units but rather whole words and sentences. As it is not feasible to construct an acoustic model for each possible sentence or even word, the word models are constructed by concatenation of the speech unit models, and the sentence models by concatenation of the word models.

3.4.1 General approach

As we saw above, in order to guarantee optimal classification minimizing the error probability, we are looking for the model which has the highest a posteriori probability. The same is true for continuous speech recognition where the sequence of feature vectors describes much longer utterances. Given an acoustic sequence $X = \{x_1, \ldots, x_t, \ldots, x_T\}$ with $x_t \in \mathbb{R}^d$, we are thus looking for the best model W_j which has the highest posterior probability:

$$P(W_j|X,\Theta) = \frac{p(X|W_j,\Theta)P(W_j|\Theta)}{p(X|\Theta)} \tag{3.28}$$

where the class W_j is the j^{th} $(0 \leq j \leq J)$ statistical model for a sequence. As it is usually difficult to train $P(W_j|X,\Theta)$ directly, two assumptions are usually made to ease the task.

- It is commonly assumed that the components in Θ which are used to estimate the prior probability $P(W_j)$ of a model are independent of the components in Θ which are used to estimate the acoustic model $p(X|W_j)$. Thus, each of them can be estimated separately.
- Moreover, during recognition, $p(X|\Theta)$ is constant for all choices of j as it is independent of the models.

We can then write the optimal decision rule as

$$j_{\text{best}} \quad = \quad \arg\max_j p(X|W_j,\Theta_A)P(W_j|\Theta_L) \tag{3.29}$$

where Θ_A is the set of parameters of the acoustic model, and Θ_L the set of parameters for the language model, parameterizing the statistical distribution of all word sequences. The parameters Θ_L of the language model can be estimated on a large text database. We concentrate on the calculation of the acoustic model $p(X|W_j,\Theta_A)$ (thus we drop index A in the following), which is usually realized in the framework of *Hidden Markov Models*, which are illustrated in the following.

Hidden Markov Models

As pointed out above, Hidden Markov modeling assumes that the sequence of feature vectors is a piecewise stationary process, that is, an utterance is modeled as a succession of discrete stationary states, with instantaneous transitions between these states.

An HMM is defined as a stochastic finite state automaton over a set of L states $W = \{q_1, \ldots, q_l \ldots, q_L\}$. In ASR we assign each state to one of the K possible classes (such as phonemes) of a set $\Omega = \{\omega_1, \ldots, \omega_K\}$[5]. Each state is moreover ascribed a stochastic output process to describe the probability of occurrence of some feature vector (see **Figure 3.3**).

Therefore, there are two concurrent processes in an HMM: the first models the temporal variability of speech through the sequence of HMM states, and the second models the locally stationary character of the speech signal through a set of state output processes. Since the sequence of states is not directly observable, the Markov models are called *hidden*.

[5] where $q_l \in \Omega$, and $p(x|q_l) \hat{=} p(x|\omega(q_l))$ when $\omega(q_l)$ is the class associated with q_l.

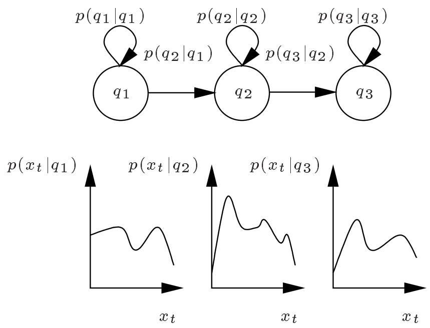

Figure 3.3: Illustration of a three-states left-to-right Hidden Markov Model (HMM). A (standard) HMM is a stochastic finite state automaton, consisting of a set of states and transitions between the states. Parameters specifying the HMM are, for each state q_l, an emission probability (density) $p(x_t|q_l)$, an initial state probability $P(q_i)$, and for each state transition, a probability $P(q_l|q_k)$ describing the possibility for transition from one state q_k (at time $(t-1)$) to the next state q_l (at time t).

Assuming that the first stochastic process is described by a first-order, time-invariant Markov model instantaneous transitions from one state q_k to the next state q_l are parametrized by the *state transition probabilities* $P(q_l|q_k)$. The models used in automatic speech recognition are usually left-to-right models allowing for transitions only from left to right, and loops which exit and enter the same state. The first state of the sequence is selected at random, according to the *initial state probabilities*.

The second stochastic process produces a feature vector $x_t \in \mathbb{R}^d$ (at time t) according to the feature vector distribution of that state. As the feature vectors usually consist of continuous observations, these vectors or *emissions* are best modeled by continuous emission pdfs $p(x_t|q_l)$, such as Gaussian or Gaussian mixture models which were described above. In the case when MLPs are used to model the distribution of the data as described in the previous section, scaled likelihoods are used instead which are obtained by dividing the posterior probabilities at the output of the MLP by the class prior probabilities. This will be discussed in more detail in Section 3.5.

Word and sentence Hidden Markov Models With the help of the dictionary, which describes the possible pronunciations of each word, the words of a language are constructed by concatenation of elementary speech unit HMMs; similarly, the sentences are constructed as sequences of words using syntactic and semantic constraints as defined by the grammar.

We now have to bare in mind that a model W consists of a concatenation of elementary HMMs and is made up from a sub-set of the L states q_l $(l = 1, \ldots, L)$. The set of parameters present in a word or sentence HMM W are denoted as Θ and is a subset of the parameters of the elementary speech unit HMMs.

The three problems When using HMMs for ASR, three problems have to be solved (Rabiner and Juang, 1993; Bourlard and Morgan, 1994):

- Parameterization and probability estimation

  Given a model W and a word or sentence realization and its short-term acoustic feature vector sequence $X = \{x_1, \dots, x_T\}$, what is the probability that the given model W actually created this acoustic sequence, i.e. how can $p(X|W, \Theta)$ be computed?

- Parameter estimation (Training)

  How are the model parameters Θ to be adapted so that each model W_j most accurately predicts the most typical realization of the acoustic observation it is associated with? In the case of ML criterion, the goal of the training is thus to find the best set of parameters in order to maximize $\prod_{j=1}^{J} p(X_j|W_j, \Theta_j)$ where W_j is the concatenation of the speech sub-units corresponding to sentence X_j, and Θ_j are its associated parameters, thus

  $$\Theta^* = \arg\max_{\Theta} \prod_{j=1}^{J} p(X_j|W_j, \Theta_j) \tag{3.30}$$

  where Θ is the whole set of parameters. Unfortunately, maximization of (3.30) in the parameter space Θ does not have a direct analytical solution. However, in Section 3.4.2 an iterative procedure is presented to maximize (3.30), or its *Viterbi approximation*

  $$\overline{\Theta}^* = \arg\max_{\Theta} \prod_{j=1}^{J} \overline{p}(X_j, Q_j|W_j, \Theta_j) \tag{3.31}$$

  where the parameters of the models are optimized iteratively to find the best parameters *and* the best state sequence Q_j.

- Decoding

  Given a set of Markov models with their trained parameters and a sequence of observations X, how should the best sequence W^* of elementary Markov models W be found to maximize the probability that W^* generated the observations, i.e.

  $$W^* = \arg\max_{W} p(X|W, \Theta) \tag{3.32}$$

  For continuous speech recognition, this problem is in general computationally intractable. However, using the Viterbi approximation, we can solve the simpler problem:

  $$\overline{Q}^* = \arg\max_{Q} \overline{p}(X, Q|W, \Theta) \tag{3.33}$$

  which gives the best sequence of states $\overline{Q}^*$ which is then transformed into the corresponding sequence of words W^*. It is computed by using (3.43) which is described in the next section.

ML parameter estimation can be carried out in the framework of HMMs using *Markov assumptions* so that the density for a complete sequence is broken up into a combination of densities corresponding to subsequences, which will be described in the following.

3.4.2 Estimation of $p(X|W)$

Following (Bourlard and Morgan, 1994), we assume that the acoustic parameters Θ are fixed and assuming them implicit in each equation, the problem consists in calculating $p(X|W)$ given a sequence of observations $X = \{x_1, \ldots, x_T\}$ of length T as well as a model W consisting of L states. We denote path Q all sequences of T states which are allowed by model W, and q_l^t represents the statistical event $\{q^t = q_l\}$ that state l has been visited at time t.

The intuitive way of estimating $p(X|W)$ consists in summing the probabilities of all possible paths admitted by the model:

$$p(X|W) = \sum_{Q \in W} p(Q, X|W) \tag{3.34}$$

With L the number of states in model W, it has to be summed over L^T state sequences, so that the number of operations which are needed for the estimation of $p(X|W)$ is approximately $O(2T \cdot L^T)$ (Rabiner and Juang, 1993), that is, it increases exponentially with the length of the acoustic vector sequence. As for longer utterances this would amount to an unsolvable task, more efficient methods have been proposed such as a recursive calculation of the forward-probabilities $P(q_l^t, X_1^t|W)$[6].

Forward recursion The likelihood $p(X|W)$ can be decomposed into the sum of joint densities with the possible final states for the length T sequence

$$p(X|W) = \sum_{l=1}^{L} p(q_l^T, X_1^T|W) \tag{3.35}$$

where X_1^t describes the acoustic vector sequence ranging from the first frame to time frame t, and q_l^t the state l at time t. Thus, to get the complete likelihood, we need to find the joint probability of the final state and all the data leading up to it. Factorizing $p(q_l^t, X_1^t|W)$ into two components, we get

$$p(q_l^t, X_1^t|W) = \sum_{k=1}^{L} p(q_k^{t-1}, X_1^{t-1}|W) p(q_l^t, x_t|q_k^{t-1}, X_1^{t-1}, W) \tag{3.36}$$

which is called the *forward recursion* and describes the probability that the model W has generated the partial sequence X_1^t and is in state l at time t. Defining

$$\alpha_t(l|W) = p(q_l^t, X_1^t|W) \tag{3.37}$$

expression (3.36) can be written as

$$\alpha_t(l|W) = \sum_{k=1}^{L} \alpha_{t-1}(k|W) P(q_l^t, x_t|q_k^{t-1}, X_1^{t-1}, W) \tag{3.38}$$

When the recursion reaches the final frame, the complete likelihood is obtained by applying (3.35).

Estimation of local contributions Let us now consider how the second term on the right-hand side of the forward recursion can be calculated. We can decompose it according to

$$P(q_l^t, x_t|q_k^{t-1}, X_1^{t-1}, W) = P(q_l^t|q_k^{t-1}, X_1^{t-1}, W) p(x_t|q_l^t, q_k^{t-1}, X_1^{t-1}, W) \tag{3.39}$$

[6] where $X_1^t = \{x_1, \ldots, x_t\}$.

As such terms are difficult to estimate, the *Markov assumption* is made that the probability of being in state q_l^t (at time t) depends only on the state at time $t-1$, and is conditionally independent of the past, so that we can write

$$P(q_l^t|q_k^{t-1}, X_1^{t-1}, W) \simeq P(q_l^t|q_k^{t-1}, W) \tag{3.40}$$

Moreover, observations are assumed to be independent of past observations (i.e. acoustic vectors are uncorrelated) and states, which amounts to

$$p(x_t|q_l^t, q_k^{t-1}, X_1^{t-1}, W) \simeq p(x_t|q_l^t, W) \tag{3.41}$$

With these simplifications, the forward recursion (3.38) becomes

$$\alpha_t(l|W) = \sum_{k=1}^{L} \alpha_{t-1}(k|W) P(q_l^t|q_k^{t-1}, W) p(x_t|q_l^t, W) \tag{3.42}$$

Thus, the total likelihood for a data sequence being produced by a particular model, can be computed using a recursion which only depends on local emission $p(x_t|q_l^t, W)$, transition $P(q_l^t|q_k^{t-1}, W)$, and initial state probabilities. To compute the emission probabilities of (3.42), each state q_l of W has to be associated with a pdf describing $p(x_t|q_l^t, W)$. As x_t is usually continuous and high-dimensional, this is often done through GMMs such as (3.10).

Within the assumptions described here, the forward recursion yields the complete likelihood and is functionally equivalent to, though more efficient than, the direct summation of the likelihoods of all possible paths, as the number of operations required to compute $p(X|W)$ is reduced to the order $O(L^2T)$. Still, the procedure can be difficult to implement as it includes both multiplications and additions of likelihoods and probabilities.

Viterbi approximation Besides numerical reasons it is often useful to find the single best state sequence for an observation sequence, such as for data alignment. Thus, in the Viterbi approximation only the most probable path is considered, instead of considering all possible state sequences Q which can have produced X. The Viterbi forward recursion is obtained by replacing the summation in (3.42) by a maximum operator, which yields an approximation to the complete likelihood $p(X|W)$,

$$\overline{p}(q_l^t, X_1^t|W) = \max_{k=1}^{L} [\, \overline{p}(q_k^{t-1}, X_1^{t-1}|W) P(q_l^t|q_k^{t-1}, W) \,]\, p(x_t|q_l^t, W) \tag{3.43}$$

where $\overline{p}(q_l^t, X_1^t|W)$ is the Viterbi approximation of the probability of the joint event that state q_l is visited at time t and the sequence X_1^t is observed, given the model W. Again, when recursion (3.43) reaches the final frame, approximation to the complete likelihood is obtained by applying (3.35), where the sum is also replaced by the maximum operator.

3.4.3 Parameter estimation

The forward recursion presented in the previous section provides a means to determine complete sequence likelihoods through a local product of state emission and transition probabilities with a cumulative value computed from allowed predecessor states. As these emission densities and transition probabilities are not known *a priori*, they have to be estimated from the training data. We will see how this can be achieved with the EM algorithm introduced in Section 3.3.2 which now takes the expected value over the space of all possible state sequences corresponding to the models for the training data, in order to increase the likelihood of a complete sequence of observations (given the models).

For this it is necessary to estimate the posterior probability for all possible states (within the topology of the HMM) for each acoustic vector x_t at time t, i.e. all possible paths through the model have to contribute to the estimation of the parameters. We see in the following how this can be achieved through the use of the above-introduced forward recursion $\alpha_t(l|W)$ and a corresponding *backward recursion* $\beta_t(l|W)$ (Estimation (E) step of the EM). The probabilities which are calculated with these recursions are then used in the M-step (Maximization step) of the EM to estimate new parameters which maximize the likelihood. This algorithm is often refered to as *Forward-Backward* or *Baum-Welch* training (Baum and Petrie, 1966).

As the exact state sequence for training of the models is usually not known, the states can be taken as the hidden variables in the EM when applied to HMMs to maximize the likelihood $\sum_{j=1}^{J} \log p(X_j|W_j, \Theta)$ (Boite et al., 2000; Gold and Morgan, 2000).

As we saw above, maximizing the likelihood is equivalent to maximizing the expectation of the joint probability of the observed and hidden variables (considering just one particular sentence X_1^t and its model W)

$$A(\Theta, \hat{\Theta}) = \sum_{Q \in W} P(Q|X_1^T, \hat{\Theta}, W) \log(p(X_1^T|Q, \Theta, W)P(Q|\Theta, W)) \tag{3.44}$$

assuming conditional independence as usual, we can approximate the two last terms in (3.44) as $p(X_1^T|Q, \Theta, W) \simeq \prod_{t=1}^{T} p(x_t|q_l^t, \Theta, W)$ and $P(Q|\Theta, W) \simeq \prod_{t=1}^{T} P(q_l^t|q_k^{t-1}, \Theta, W)$, so that we can write

$$\begin{aligned} A(\Theta, \hat{\Theta}) &= \sum_{t=1}^{T}\sum_{l=1}^{L} P(q_l^t|X_1^T, \hat{\Theta}, W) \log p(x_t|q_l^t, \Theta, W) \\ &+ \sum_{t=1}^{T}\sum_{l=1}^{L}\sum_{k=1}^{L} P(q_l^t, q_k^{t-1}|X_1^T, \hat{\Theta}, W) \log P(q_l^t|q_k^{t-1}\Theta, W) \end{aligned} \tag{3.45}$$

E-step Defining

$$\beta_t(l|W) = p(X_{t+1}^T|q_l^t, X_1^t, W) \tag{3.46}$$

we can write the backward recursion as

$$\beta_t(l|W) = \sum_{k=1}^{L} \beta_{t+1}(k|W)P(q_k^{t+1}|q_l^t, \hat{\Theta}, W)p(x_{t+1}|q_k^{t+1}, \hat{\Theta}, W) \tag{3.47}$$

which has been chosen such that $P(q_l^t, X_1^T|\hat{\Theta}, W) = \alpha_t(l|W)\beta_t(l|W)$, which describes a complete data sequence which has passed through a particular state l at a particular time t.

With the help of the forward and the backward recursions the posterior probabilities which are needed to update the parameters can be derived

$$P(q_l^t|X_1^T, \hat{\Theta}, W) = \frac{P(q_l^t, X_1^T|\hat{\Theta}, W)}{p(X_1^T|\hat{\Theta}, W)} = \frac{\alpha_t(l|W)\beta_t(l|W)}{\sum_{k=1}^{L} \alpha_t(k|W)\beta_t(k|W)}$$

$$P(q_l^t, q_k^{t-1}|X_1^T, \hat{\Theta}, W) = \frac{\alpha_{t-1}(k|W)P(q_l^t|q_k^{t-1}, \hat{\Theta}, W)p(x_t|q_l^t, \hat{\Theta}, W)\beta_t(l|W)}{\sum_{k=1}^{L}\sum_{l=1}^{L}\alpha_{t-1}(k|W)P(q_l^t|q_k^{t-1}, \hat{\Theta}, W)p(x_t|q_l^t, \hat{\Theta}, W)\beta_t(l|W)}$$

M-step Maximizing the expected value (3.45) by setting its partial derivatives to zero, as had been illustrated above when the EM was applied to GMM parameter estimation, the update formulae for the transition probabilities and the state emission probabilities are found. In case of Gaussian mixture emission probabilities, this involves updating of the mean vector, covariance matrix for each class, as well as the weights.

After a probability estimator has been chosen for the densities associated with each state, the EM algorithm for HMM parameter estimation can be summarized as

- Initialize the parameters.
- Given these parameters, apply the forward and backward recursions to estimate the posterior probabilities for the hidden variables.
- With these posterior estimates, find the parameters that maximize the expected value of the joint density for the data and the hidden variable.
- Iterate the last two steps as long as A continues to increase.

Viterbi training As in Viterbi estimation, in Viterbi training the parameters are optimized in such a way as to optimize the likelihood of the best path only, that is, of the most probable state sequence in the model. This corresponds to assuming that the posterior probabilities of the states employed in the EM are either zero or one. The sequence of states which maximizes the likelihood thus corresponds to a state segmentation, which is updated in each EM iteration. The best sequence and with that the segmentation can be back-tracked from the end of each utterance following the best path which had been found. With this and a given parameter estimator for each state, training is now supervised, and the parameters can be optimized over all observation vectors associated with each state. After calculation of new emission and transition probabilities, these can be used to obtain a new segmentation.

Local probability estimation As we saw in Section 3.3.2, the approach which is usually pursued to estimate the local emission probabilities $p(x_t|q_k)$ is by modeling them with probability density functions, such as GMMs. We refer to these systems as HMM-GMMs. Estimating the parameters of these models using the EM algorithm as described above, does not guarantee that the parameters will also reduce the likelihoods of the incorrect models. Discriminant training which increases the likelihood of the correct model and, at the same time, decreases the likelihood of the incorrect models can be achieved through the use of neural networks, which was described for static classification in Section 3.3.3. In the following section, we will point out how this approach can be realized also in HMM-based sequence-recognition systems.

3.5 Hybrid HMM/ANN systems

The combination of HMMs with ANNs is refered to as HMM/ANN hybrid. In this approach, the neural network is used to estimate the local probabilities of the HMMs, whereas the HMMs continue to model the sequential character of the signal. This allows to avoid or at least reduce some of the limiting assumptions of HMMs and, at the same time, to incorporate some of the advantages of ANNs, such as their discriminability and their ability to include acoustic context. More advantages will be discussed below.

Given that ANNs are trained to estimate posterior probabilities (assuming that each output is trained to correspond to one state) and not likelihoods as in the case of GMMs, it is necessary to convert the posterior probabilities to HMM emission probabilities. This is done by applying Bayes' rule to the output probabilities:

$$\frac{p(x_t|q_k)}{p(x_t)} = \frac{P(q_k|x_t)}{P(q_k)} \tag{3.48}$$

That is, the estimates of the posterior probabilities are divided by estimates of the class prior probabilities. The scaled or normalized likelihood of the left-hand side can be used as an emission probability for the HMM, as during recognition the scaling factor $p(x_t)$ is a constant for all classes and does not influence the classification[7].

Given the state prior probabilities, HMM training and recognition can be conducted just as in the case of HMMs employing pdf's to estimate the emission probabilities, i.e. Baum-Welch training and Viterbi decoding. The only difference being that each posterior probability needs to be divided by the respective prior probability, yielding scaled likelihoods. In the case of Baum-Welch training, the scaled likelihoods are used in the forward-backward equations to estimate posterior probabilities for each state and frame. The network is then retrained, using these probabilities as targets. In some approaches, the HMM is given fixed transition probabilities and only one state is used per sub-word unit. This way, HMM training is actually no longer needed as the emission probabilities are obtained from the ANN, which is trained on a labeled database, and with that all parameters which are needed for decoding are available.

Advantages HMM/ANN hybrid systems have several advantages as compared to HMM-GMMs in which the emission probabilities are estimated by GMMs (Renals and Hochburg, 1995; Bourlard and Morgan, 1997).

- First of all, the features do not have to be assumed uncorrelated. In the case of ANNs we can even process several acoustic vectors at a time (as input to the ANN) which allows the network to learn local correlation of the acoustic vectors. This is a simple mechanism to introduce acoustic information from long temporal contexts into the recognition process. Moreover, diverse features, such as mixtures of continuous and discrete measures, can easily be combined.

[7] It has been found during experimental evaluation by our institute and others, that the theoretically necessary division by prior probabilities can sometimes be disregarded or assumed to use equal priors to enhance recognition performance. This seems to be the case especially for speech databases where class prior probabilities do not vary much.

- Second, ANNs can be trained discriminantly (when trained with discriminant criteria) which long has not been the case for GMMs. Discriminant training attempts to model the class boundaries, i.e. learn the distinction between the classes, rather than to build as accurate a model as possible for each class. The combination of ANNs and HMMs thus leads to HMMs with *local* discriminative capacity. However, recent theoretical work shows that global discriminant training of hybrid systems can also be performed (Bourlard et al., 1995).

- Another important advantage is that performance of HMM/ANN hybrid systems on noisy test data is usually found to be higher than that of an HMM-GMM.

- Posterior probabilities as estimated by ANNs have the advantage that they are independent of the size of the input space, which is not the case for likelihoods where the magnitude depends on the size of the feature space. Especially in the case of multi-band processing when differently sized subbands are to be treated this can lead to a problem in the case of Gaussian modeling. Thus, in the framework of multi-band and multi-stream processing, the posterior probabilities at the output of the ANN make it easier to merge multiple recognizers, each of them working on different input data.

Most of the speech recognizers employed in this thesis utilize HMM/ANN hybrids, where the neural network is an MLP (referred to as HMM/MLP hybrid). For some developments where we needed access to real likelihoods (and not the scaled likelihoods as output by an MLP) we resort to HMM-GMM systems.

3.6 Summary

In this chapter, we presented important technical background knowledge in order to prepare the ground for the theoretical and experimental developments in this thesis, which are based around HMM-GMM and HMM/MLP hybrid systems. We first discussed the general structure of a state-of-the-art automatic speech recognizer and gave some background knowledge to information theory and on statistical pattern classification. The concept of the optimal Bayes' classifier was introduced and its approximation through probability functions whose parameters are estimated on the training data was discussed. These probability functions can be densities, as in the case of GMMs, for which the EM algorithm is used for ML training of the parameters. The posterior probabilities of the Bayes' classifier can also be estimated directly from ANNs, for which the Error Back Propagation algorithm for MAP training of the parameters was described.

For continuous speech recognition, statistical sequence recognition is employed where sequences are modeled through concatenation of (elementary) Hidden Markov Models. Based on the ML criterion we discussed estimation of the likelihood that the data X was produced by model W (i.e. $p(X|W)$) through application of the forward recursion, and pointed out the assumptions of conditional independence which are needed to estimate the local contributions in the recursion. The same forward recursion together with the backward recursion can be used in EM training of the parameters of the models.

Local emission probabilities in the HMM are modeled by pdf's (typically GMMs). In the

HMM ANN hybrid, they are modeled by scaled likelihoods which are obtained from the posterior probabilities from an ANN (typically an MLP) after division by their respective prior probabilities. In this thesis, the first system (usually refered to as HMM) will be refered to as HMM-GMM, while the latter (which is the recognizer which is mainly employed in the experimental work in this thesis) will be refered to as HMM/MLP hybrid system.

CHAPTER 4

Robustness in ASR

Human listeners have a strong ability to cope with speech under a vast variety of conditions. The capability to unconsciously filter out interfering speech and non-speech signals which are not part of the speech signal of interest has been manifested in speech perception experiments. In psychoacoustics and neurophysiology, well-known mechanisms of human perception have been established which describe these experimental findings as we saw in a previous chapter. This knowledge can sometimes be used as to guide the construction of an ASR system and was often shown to render automatic processing more reliable (see below). As well as this, pure engineering based approaches have been developed which contribute their part in rendering automatic systems more robust. Despite these efforts, automatic speech recognition is still far below human standards.

In the following, we will see in more detail, the different causes of disturbance of the incoming speech signal which can effect automatic speech recognition systems. We will then give an overview of the different engineering approaches which are applied to this problem, discussing more thoroughly the methods which are of interest to this thesis.

4.1 Causes of adversity

We start this section by giving a short introduction on the different kinds of perturbations which are encountered between speech production and its reception. We point out which of these disturbances we try to account for in the framework of this thesis, and discuss these in more detail in the following sections.

4.1.1 Introduction

Between speech production and its reception, there are several stages at which the speech signal can be disturbed (cf. **Figure 4.1**). Following Junqua and Haton (1996), they can be classified into two broad categories:

1. Speech and speaker variability

   This category comprises style variations (e.g. formal, casual, spontaneous, read, dictated), speaking rate (e.g. normal, slow, fast), stress (e.g. emotional factors, in noise, under cognitive load), context (e.g. interview, free conversation) and voice quality (e.g. breathy, whispery, tense). Inter-speaker variability (such as length and shape of the vocal tract, physiology of the vocal cords, etc.) and intra-speaker variability (such as the physiological and psychological state of the speaker, etc.) are implicitly included in each of the above five sub-categories. They are summarized on the left-hand side of **Figure 4.2**.

   These speech and speaker intrinsic factors are not investigated in this thesis. We concentrate on the second category:

2. Noisy speech and channel distortions

   These noise cases are independent of the speech and the speaker but depend on the transmission environment, as well as possible recording and reproduction devices. These causes are summarized on the right-hand side of **Figure 4.2**. Noises induced in this way can be linear in the power spectral domain, which is refered to as *additive noise*, or linear in the logarithmic spectral or cepstral domain, which is refered to as *convolutive noise*, or non-linear in both domains (Junqua and Haton, 1996).

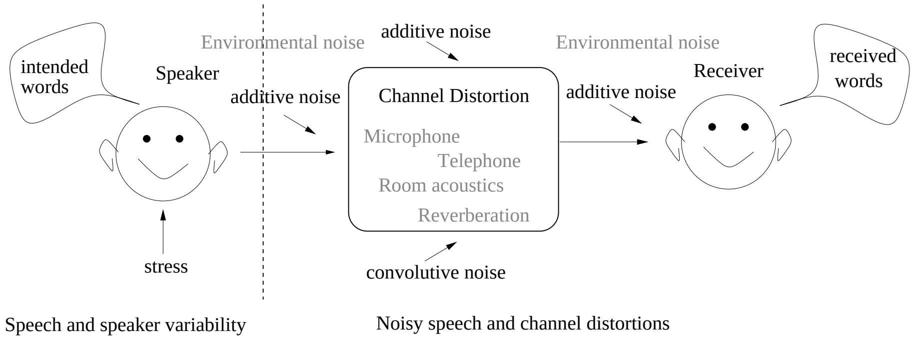

Figure 4.1: Illustration of the different causes of adversity which can occur between speech production and reception.

Environmental noises, such as office environment noise or in-car and factory noises, are usually additive (Junqua and Haton, 1996, p.160), though this is not always the case. Channel distortions can be divided into telephone distortion, microphone-induced distortions, room acoustics and reverberation. They can be additive, convolutive or both.

Mathematical modeling Let $s(t)$ describe the (clean) speech signal at time t as produced by the speaker under certain conditions that might influence the speaker such as stress and noise[1]. Signal $s(t)$ is first recorded with the help of a microphone, which has the impulse response $h_{micro}(t)$ and also picks up *background (ambient) noise* $n_1(t)$. The resulting signal is then transfered by a (short- or long-distance) transmission channel, which behaves like a

[1]The phenomenon that a speaker alters his/her way of speaking in a noisy environment is referred to as "Lombard effect".

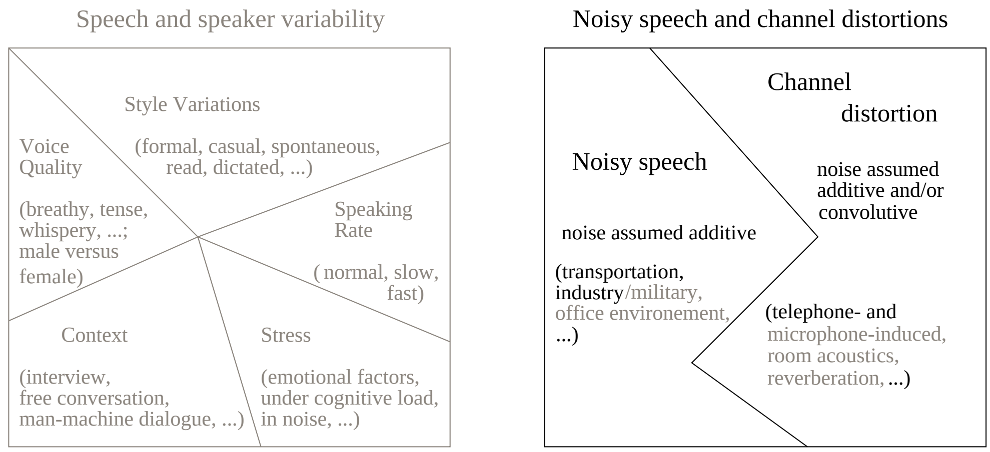

Figure 4.2: Illustration of the two broad categories of causes of adversity on the speech signal. Left part taken from Figure 4.1 in (Junqua and Haton, 1996, p. 128). The grey color indicates noise cases which are not investigated in this thesis.

linear *convolution* filter $h_{chan}(t)$ and, moreover, can add noise $n_2(t)$. Last but not least, there is additive noise present at the receiver $n_3(t)$. The observed corrupted-speech signal $y(t)$, which arrives at the recognizer, is thus related to the noise-free and distortion-free speech signal $s(t)$ by

$$y(t) = \left[\left\{ \left(\left[s(t) | \begin{matrix} Stress \\ Lombard \end{matrix} \right]_{n_1(t)} + n_1(t) \right) \otimes h_{micr}(t) + n_2(t) \right\} \otimes h_{chan}(t) \right] + n_3(t) \quad (4.1)$$

where $\otimes$ denotes the convolution operator, and $n_1(t)$ is the background noise, $h_{micro}(t)$ the impulse response of the microphone, $n_2(t)$ and $h_{chan}(t)$ are, respectively, the additive noise and impulse response of the transmission channel, and $n_3(t)$ is the noise present at the receptor (Gales, 1995).

For ease of mathematical modeling, the various additive and convolutive noise cases are, respectively, taken together. We can then describe the signal $y(t)$ which arrives at the recognizer in a simpler way as:

$$y(t) = x(t) \otimes h(t) = [s(t) + n(t)] \otimes h(t) \quad (4.2)$$

where $x(t)$ is the clean speech signal corrupted by additive noise $n(t)$ only, $h(t)$ summarizes the convolutive noise cases. Normally, the data is not modeled in the time domain, so that (4.2) needs to be transfered to the respective domain in which modeling is performed, such as the spectral or cepstral domain. In the case of the spectral domain, (4.2) becomes

$$Y(f) = X(f) \cdot H(f) = [S(f) + N(f)] \cdot H(f) \quad (4.3)$$

under the assumptions that $H(f)$ is constant over time and independent of the incoming signal level.

Other noise classifications Noises can also be distinguished according to their time and/or frequency distribution. Some noises are *stationary*, that is, the model which describes the noise does not change in time, or *non-stationary*, that is, the noise model changes in frequency over time. Other characteristics for noises are periodicity (e.g. in the case of engine noise), or impulsiveness (e.g. in the case of machine gun and some factory noises). Finally, noises can also be described as narrow-band or wide-band noise, depending on whether they only cover a certain, limited frequency region of the signal, or whether the whole signal is affected by the noise.

4.1.2 Additive noise

Environmental noise is usually assumed additive and uncorrelated with the speech signal. Assuming absence of convolutive noise, the recorded speech signal $x(t)$ can then be described as a sum of the intended speech signal and an acoustic contamination from noise-like signals: $x(t) = s(t) + n(t)$. Additive noise can either be stationary or non-stationary.

Examples of additive noise occurrences Ambient noise already occurs in office environments due to type-writers, printers, ventilator noise from computers, or people speaking in the background. The so-called *Cocktail Party Effect* is a well-known phenomenon. It refers to the ability of a listener in a "cocktail party" to be able to listen to just one of the speakers as if the others were not present, even though they may all be speaking at the same volume. On the contrary to humans, such noise occurrences severely disturb automatic speech recognizers. Moreover, in this noise case, also transient or impulse noise can be encountered, which is often of high-intensity such as door slamming, phone ringing or passing cars.

A major use of speech recognition applications is for example in an automobile. Here, the variety of ambient noises is even more striking. First, there are noise cases from outside the passengers' area such as low-frequency engine noise, tyre and windshield wiper noise, high-frequency noise from the air flow (increasing with speed), and non-stationary noise from passing vehicles (traffic noise), possibly rain falling onto the car roof or other weather-influenced conditions. Second, we encounter interferences from inside the car such as the radio playing (speech and music), passengers talking and noise from the indicator warning signal. The SNR in a passenger car can drop to -5 dB when the car is going at a speed of 90 km/h with the window closed (Lecomte et al., 1989). Degan and Prati (1988) showed that for a car moving at 100 km/h[2] the noise power in the region between 1 and 6 kHz increased considerably as compared to a stationary car with the engine turning at 4000 r.p.m. This means that the frequency region which is supposed to be the most important for human speech recognition (cf. Section 2.2) is also the region which gets the most perturbed in the car. Several publications have demonstrated the difficult task of accurate speech recognition of an in-car application (Mokbel and Chollet, 1991; Schless and Class, 1997).

The (real-environmental) additive noise cases which we chose for investigating the noise robustness of our speech recognizers are a factory noise case and an in-car recorded noise case. Moreover, artificial additive (stationary and non-stationary) band-limited noise was created. They are described in Section 8.2.

[2] with the gear in neutral position and the engine off

4.1.3 Channel distortion

If the signal is recorded, the type of microphone and its (possibly changed) position can drastically influence the speech signal (Acero and Stern, 1990). In (static and mobile) telephone applications, there is, moreover, the distortion introduced by the telephony transmission channel. Due to a large variety of telephone gadgets and transmission line characteristics, such attenuation distortions are hard to predict. The limited bandwidth of the transmission channel of 200/300-3200/3400 Hz (Schukat-Talamzzini, 1995, p. 47) additionally restricts the quality of the speech presentation. In the case of mobile telephone applications, the radio channel has a continuous variable transfer function which provokes an even more severe degradation of the speech signal (Degan and Prati, 1988, p. 43). Moreover, in cellular phones, the switching from one cell to the other can produce a switching noise (Moreno and Stern, 1994). For these reasons, utilizing a speech recognizer over the telephone line which had been trained on data not recorded over a telephone line or on a different transmission channel leads to a severe degradation in recognition accuracy if the change in application environment is not accounted for such as through the use of robust features and other techniques.

Room reverberation is another source of distortion noise. Walls and other objects in the room where a speech recognizer is employed produce, at different degrees, reflections of the speech signal which alter its spectrum.

In the experiments carried out in this thesis, we use speech data which was recorded over the telephone line for both training and testing in order to concentrate on the influence from the additive noise cases.

4.2 Robust feature processing

The notion of "(speech) features" was introduced above as a machine-internal, characteristic representation of the speech signal. The use of reliable features is a key issue in the design of an automatic speech recognition system. In order to emulate the robustness of the human auditory system, many of the most effective acoustic features, which only came into existence over the last decade, are based on human auditory characteristics. They incorporate mechanisms based on psychoacoustic and neurophysiologic evidences such as critical band filtering, equal loudness pre-emphasis, and non-linear energy compression which have been illustrated in Section 2.2. The most widely used and most promising of these features will be presented in the following.

4.2.1 State-of-the-art acoustic features

Speech features which are directly calculated from the speech signal are hardly anymore used in ASR, with the exception of the short-time energy, which is an effective measure to distinguish between voiced and unvoiced sounds (Schukat-Talamzzini, 1995). Instead, spectral and cepstral analyses are carried out which usually lead to more robust representations of the speech signal than the time features. Moreover, these analysis techniques are often combined with processing strategies based on human auditory characteristics.

Along this line, critical-band filtering aims at modeling acoustic filtering in the human cochlea. This implies narrow frequency bands for low frequencies and larger frequency bands for higher frequencies. As we already saw, different approaches have been proposed to model such critical-band representation, such as the Mel and Bark scale. The former is amongst others used in mel-frequency cepstral coefficients, the latter in perceptual linear prediction processing.

MFCC features The processing for the more robust and still inexpensive *mel-frequency cepstral coefficients* (MFCCs) (Davis and Mermelstein, 1980) can be seen on the right side of **Figure 4.3**. After spectral analysis, the (power) spectrum is integrated within overlapping critical band filters which follow the Mel scale. The spectrum is then pre-emphasized to approximate the unequal sensitivity of human hearing at different frequencies. To approximate the power law of hearing, the spectral amplitudes are compressed by a logarithmic function. The cepstral coefficients are then calculated through an inverse (discrete) Fourier transform (IDFT). For spectral smoothing to suppress the effects of further nonlinguistic sources of variance, the higher Fourier components in the compressed spectrum are ignored ("cepstral truncation").

Perceptual linear prediction *Perceptual linear prediction* (PLP) analysis (see left side of **Figure 4.3**) is based on linear prediction (LP) analysis but additionally includes auditory properties through the computation of a compressed critical band spectrum (Hermansky, 1990). To obtain an estimate of the auditory spectrum, the (FFT) spectrum is convolved with the critical-band function. The integration step is this time done with trapezoidally shaped filters which are applied at roughly Bark intervals (whereas triangular windows are used in MFCC processing). The resulting Bark-scaled spectrum is multiplied by a fixed equal-loudness curve (pre-emphasis). The amplitude of the output is compressed by a cube-root function which is used instead of the logarithm to approximate the power law of hearing. After inverse (discrete) Fourier transform (yielding autocorrelation coefficients), the autoregressive model is calculated to smooth out details from the auditory spectrum. The autoregressive coefficients are then usually transformed in to orthogonal parameters, such as cepstral coefficients. Nevertheless, PLP analysis was found to be vulnerable to linear spectral distortions (Hermansky et al., 1992). To alleviate this problem, PLP processing is often combined with a *RelATive SpecTrAl* (RASTA) filtering method, which is described in Section 4.2.4.

Both the mel-cepstral analysis and PLP provide a feature representation which corresponds to a smoothed short-term spectrum that has been compressed and equalized much as done in human hearing.

4.2.2 Spectral and cepstral mean normalization

Convolutive noise can more easily be removed in the logarithmic spectral domain or in the cepstral domain where the noise is additive (Deller et al., 1987).

In *spectral mean normalization*, the average of the input speech spectrum is measured over the whole utterance (of T frames) to approximate the channel transform function $|\hat{H}(f)|^2$, and subtracted (in the logarithmic domain) from the noisy speech spectrum $Y(f)$, as shown in (4.5). Although some speech characteristics might also be suppressed due to this long-term estimate,

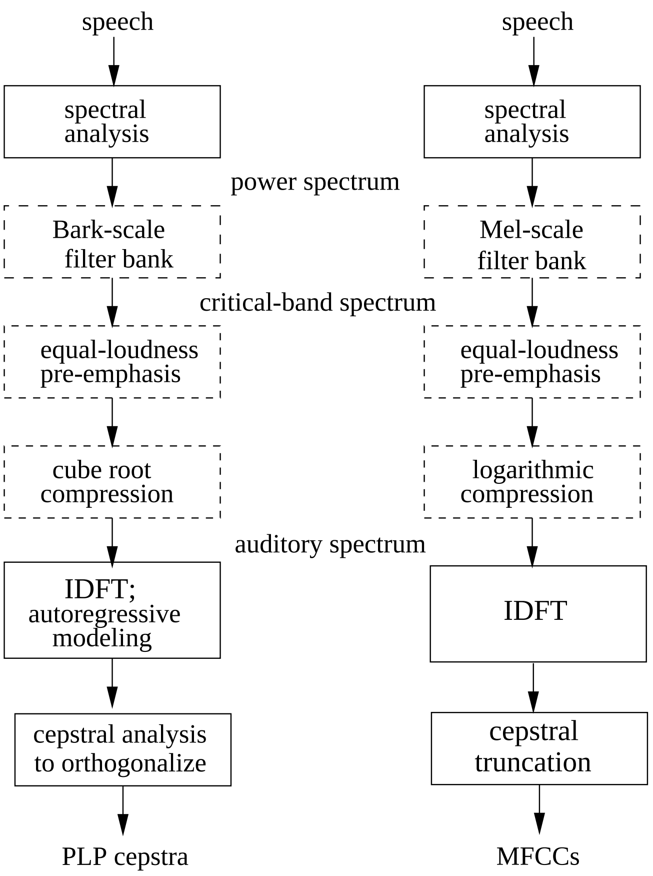

Figure 4.3: Illustration of the different processing steps for PLP (left) and MFCC (right) analysis. The dashed boxes indicate the processing modules which account for the perceptual analysis and, thus, constitute the main differences to standard LP (in the case of PLP analysis) or cepstral modeling (in the case of MFCC analysis). (Gold and Morgan, 2000).

most of the (fast-changing) phonetic speech information will be preserved[3].

$$\log \hat{S}(f) = \log Y(f) - \log |\hat{H}(f)|^2 \tag{4.4}$$

$$\simeq \log Y(f) - \frac{1}{T}\sum_{T} \log Y(f) \tag{4.5}$$

with $\hat{S}(f)$ an estimate of the original speech spectrum, $Y(f)$ the spectrum after channel distortion, $|\hat{H}(f)|^2$ the channel transfer function, and T the number of frames in an utterance (Hanson et al., 1996). Additive noise is assumed to be negligible.

The same is possible in the cepstral domain, where the mean of cepstral vectors are subtracted from the cepstral coefficients of an entire utterance on a sentence-by-sentence basis (*cepstral mean normalization*).

In real-time implementations, it is difficult to obtain such long-term mean values. Instead, a (weighted) average can be calculated over the preceding signal (of $K \ll T$ frames) instead of the whole utterance to speed up the process (Hanson et al., 1996)

$$\log \hat{S}(f) \simeq \log Y(f) - \sum_{K} \alpha_k \log Y_k(f) \tag{4.6}$$

4.2.3 Spectral subtraction

Spectral Subtraction is based on the assumption that noise and speech are uncorrelated and additive in the time domain. Moreover, convolutive distortions are supposed to be negligible. In this case, the noisy power spectrum is the sum of the noise power spectrum and the speech power spectrum. Assuming that the noise characteristics change more slowly than those of the speech, an estimate of the noise power spectrum $|\hat{N}(f)|^2$, which is usually obtained in non-speech intervals, can be subtracted from the power spectrum of the corrupted speech signal $|Y(f)|^2$ to approximate the clean spectrum $|\hat{S}(f)|^2$:

$$|\hat{S}(f)|^2 = |Y(f)|^2 - |\hat{N}(f)|^2 \tag{4.7}$$

For this method to be successful, a good estimate of the noise is needed and thus, a reliable speech versus noise detector to identify pure noise portions. Due to spectral similarities between many unvoiced speech sounds and certain background noises this is a difficult task, especially at low SNR. For calculation of the noise estimates during non-speech periods, the background noise is assumed to not vary significantly throughout the speech sections.

More recently, noise estimation techniques which no longer rely on speech pause detection have been proposed (Hirsch and Ehrlich, 1995), which also allow for quicker adaptation to slowly varying noise levels or spectra.

Subtraction of the noise power can result in negative values if the noise estimate exceeds the real noise magnitude. This is usually accounted for by setting all negative values to a non-negative threshold for example by half-wave rectification. (If the recovered speech spectrum is then re-synthesized, these remaining isolated patches of energy produce residual noise called

[3] Subtraction of logarithmic spectra corresponds to a division of power spectra which explains the name spectral "normalization".

"musical noise" which can be minimized by using non-linear subtraction techniques (Lockwood et al., 1991)).

Spectral subtraction cannot be used in the logarithmic spectral or cepstral domain where the noise is signal-dependent and can no longer be removed by subtraction.

4.2.4 Filtering of spectral or cepstral coefficients

Experiments on both human speech perception and machine recognition have shown the importance of *spectral transitions* from one phoneme to the next, i.e. the time variations or *dynamics* of a speech spectrum, rather than the phoneme-internal constant parts (Furui, 1986a). As compared to distortion noise, which is usually characterized by very low modulations, the speech-specific variations happen over a shorter time period. Thus, assuming that slow-changing variations in the spectrum are due to channel effects and anyhow carry little phonetic information, spectral dynamics can be emphasized by *suppressing the slower varying parts* in the spectrum.

RASTA filtering

One approach to suppress slow-varying changes in the spectrum is by RASTA (RelAtive SpecTrAl) filtering (Hermansky and Morgan, 1994; Hermansky et al., 1992) the short-term spectrum. The low *modulation* frequencies can be suppressed by a simple high-pass or band-pass filter of the parameter feature vectors. The filtering can be carried out in different spectral domains, such as the linear spectral domain or the logarithmic spectral domain, depending on the kind of noise that should be suppressed:

- linear spectrum to suppress additive noises
- logarithmic spectrum to suppress convolutive noise, such as linear distortion introduced by the transmission channel.

RASTA filtering is carried out after decomposition of the spectrum into critical bands, such as during PLP analysis or mel-cepstral analysis (and before auditory processing), as illustrated in **Figure 4.4**. In the case of filtering in the compressed domain, the logarithm (or any other non-linear compression function) is taken of the critical-band spectrum. Then the filter, which has the transfer function (4.8) is applied to the (logarithmic) spectral component of each frequency channel:

$$H(z) = 0.1z^4 \frac{2 + z^{-1} + z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \tag{4.8}$$

After the filtering, the conventional (PLP) analysis is resumed at the step after critical-band integration. In the case of non-linearly compressed spectral components, a following de-compressing non-linearity is carried out, which for prior logarithmic compression is the exponential function.

The low cut-off frequency which approximates a sharp spectral zero at the zero frequency of each channel, defines the fastest spectral change (i.e. the modulation frequency) which will still

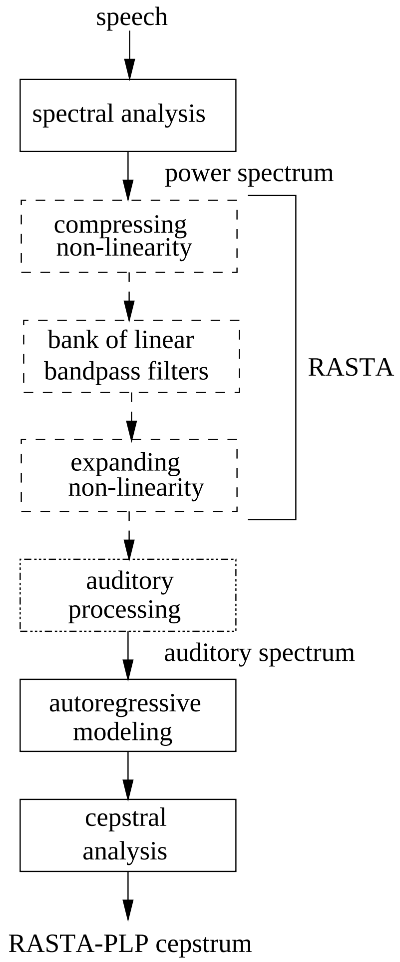

Figure 4.4: Illustration of the different processing steps for RASTA-PLP analysis. The dashed boxes are the RASTA processing modules, the dashed-dotted box indicates the perceptual analysis of PLP. Commonly, the compressing non-linearity is a logarithm and the expanding non-linearity an exponential. (Gold and Morgan, 2000).

be suppressed. It is usually located at around 0.26 Hz, so that the constant and very slowly changing components of the respective frequency channel are filtered out (high-pass part). The high cut-off frequency of the filter determines the highest modulation which will be preserved by the filter; thus, also higher modulation frequencies than those usually found in speech can be suppressed (low-pass part), such as frame-to-frame changes due to analysis artifacts. Thus, the passband is established in the domain of critical band modulations to a range that appears to be required for speech intelligibility (Gold and Morgan, 2000).

Simultaneous filtering of additive and convolutive noise To deal with both additive and convolutive noise at the same time, an extension of RASTA filtering, called J-RASTA filtering, can be used (Morgan and Hermansky, 1992; Koehler et al., 1994). For this, the logarithmic non-linearity in the RASTA filtering procedure is replaced by a function which is nearly linear for low values of signal power relative to noise (i.e. reducing additive noise effects) and logarithmic for high signal power values (i.e. reducing convolutive noise effects). Thus, the filter is applied on a function which is roughly identity in the case of additive noise, and logarithmic for the case of convolutive noise. The exact form of the function is controlled by the value of the parameter J, which is adapted according to the additive noise level (Koehler et al., 1994).

Time difference features

Another possible set of spectral dynamic features are the differenced coefficients of the static features which measure the change in coefficients over time (Furui, 1986b).

First order derivatives can for example simply be calculated by taking the difference of feature vectors of a certain distance m:

$$\Delta x_{i,t} = x_{i,t+m} - x_{i,t-m} \qquad \forall\, i\,,t \tag{4.9}$$

for short-term feature vector coefficient $x_{i,t}$ at time t. The derivative window $m\,\Delta T$ defines the amount of time over which the derivatives are calculated and is usually different for the first and higher order derivatives. The choice of the size of the window is crucial and not easy to conduct. If the window is too short, differences pick up irrelevant details in the speech signal. If the window is too long, on the other hand, no information is gathered at all.

Calculation errors can be introduced by the fact that the spectral features, on which the derivatives are based, are calculated over very short analysis windows. Therefore, linear least square regression, which considers all speech frames in the derivative window, is used to smooth the derivatives:

$$\delta x_{i,t} = \frac{\sum\limits_{j=-m}^{m} j\, x_{i,t+j}}{\sum\limits_{j=-m}^{m} j^2} \qquad \forall\, i\,,t \tag{4.10}$$

that is, the slope of the straight line is calculated which minimizes a least squares criterion of the sum of the distances between the straight line and the $2m+1$ considered points.

Second order derivatives are obtained by simply applying the difference equation of (4.9) a second time:

$$\Delta\Delta x_{i,t} = \Delta x_{i,t+m} - \Delta x_{i,t-m} \qquad \forall\, i\,, t \tag{4.11}$$

with m depending on the span of time information that is to be considered: medium-term or long-term information. The same is valid for second or higher order derivatives calculated via regression (iterative application of (4.10)).

The time derivative features are usually concatenated to the static feature vector and passed on to the classifier conjointly.

4.2.5 Linear discriminant analysis

In statistical pattern recognition, features which are orthogonal are often advantageous for modeling purposes, such as in GMMs. In the two approaches illustrated below, the feature vectors are directly and completely orthogonalized with the help of a given data set. A set of orthonormal basis functions which span the feature space are computed from statistics estimated on the training data, and then applied to the feature vectors. The basis functions can, if desired, be ranked by order of importance according to special criteria.

Linear Discriminant Analysis (LDA) tries to find a linear transformation of one parameter space into another by minimizing intra-class variance and maximizing inter-class variance (Duda and Hart, 1973). The basis functions are determined as the eigenvectors of the "between"-to-"within" covariance ratio. LDA assumes that the data in each class can be modeled by a single Gaussian distribution that shares its covariance matrix with all the other classes.

Principle Component Analysis (PCA) (also termed *Karhunen-Loeve* (KL) transform (Bishop, 1995)) can be applied after LDA. Its basis vectors are obtained from minimizing the "mean squared error" between the reconstructed and original set of vectors. PCA can select the direction of greatest variance of the inter-class covariance matrix which allows for dimensionality reduction of the final feature vector rendering the representation more compact.

Non-linear Discriminant Analysis (NLDA) is carried out in the so-called *Tandem approach* (Hermansky et al., 2000; Ellis et al., 2001) where the speech features are post-processed by an ANN (usually an MLP) before they are passed on to an HMM-GMM classifier. In order to keep the network outputs approximately Gaussian distributed, the usually used softmax non-linearity at the network output is removed (Sharma et al., 2000). The network outputs are then diagonalized through KL transform for subsequent modeling using diagonal covariance matrices.

4.2.6 Frequency difference features

A novel approach for the generation of robust spectral features is so-called *Frequency Filtering* (FF) introduced by Nadeu and Juang (1994); Nadeu et al. (1997). In this method, instead of taking derivatives over time, constant or slow-varying parts in the spectrum are removed by taking the derivative *over frequency* (Nadeu et al., 2001; Gemello et al., 2000). This is based on computationally simple first or second order FIR filters such as for example

$$H_1(z) = 1 - z^{-1} \tag{4.12}$$

$$H_2(z) = z - z^{-1} \tag{4.13}$$

which corresponds to simple differences over frequency $\Delta x'_1(f) = x(f) - x(f-1)$ and $\Delta x'_2(f) = x(f+1) - x(f-1)$ performing a combination of decorrelation and liftering[4].

The filters are usually applied either to logarithmically[5] or cube-root compressed filter bank energies (Nadeu et al., 2001; Macho et al., 1999), or to LP parameters, and substitute the conventionally following Discrete Cosine Transform (DCT) or KL transform which are usually used to decorrelate feature coefficients. The frequency-filtered parameters are thus not converted into cepstral representation but stay in the spectral domain. In order to account for loss of the first and last spectral values due to the filtering, a zero is appended at the beginning and end of each spectral vector. For this reason, the energy coefficient is usually not included in the feature vector as the endpoints of the feature vector consisting of the second and last but one logarithmic energies already include energy information. Due to their spectral representations these features seem to be especially suited for the design of classifiers which rely on non-orthogonalized but robust features such as "missing data" processing, which will be discussed in Section 4.4.

Tests with frequency filtered logarithmic energies showed competitive performance in clean and improved noise robustness as compared to MFCC features.

The FF coefficients can, just as any other coefficients, also be augmented by their time derivatives which renders them even more noise robust (Macho et al., 1999). This approach is refered to as *time and frequency filtering* (tiffing).

4.2.7 Speech enhancement

Speech enhancement techniques try to recover either the waveform or the parameter vectors of the clean speech from the noise-corrupted input signal. This way, the mismatch between test data and the original training data is reduced.

Active noise cancellation An example of speech enhancement for additive noise is microphone array processing, in which multiple microphones are employed. The first microphone records the desired speech signal $s(t)$ together with the ambient noise $n_1(t)$, whereas the second microphone measures a secondary noise input $n_2(t)$ which is supposed to be correlated with noise $n_1(t)$ but not with speech input $s(t)$. Noise input $n_2(t)$ is then subtracted from the corrupted-speech input $s(t) + n_1(t)$ of the first microphone. It has been shown (Widrow et al., 1975) that noise $n_1(t)$ in the primary input can be well eliminated from the corrupted speech, when the noise $n_2(t)$ is free from speech but coherent with noise $n_1(t)$. Sufficient coherence of the two noise signals can only be achieved when the microphones have a maximal distance of less than 5 cm from each other (Degan and Prati, 1988), which at the same time, makes it difficult not to capture any speech with the noise-only microphone.

[4] "liftering" means weighting in the cepstral domain.

[5] Note that if frequency filtering is applied to the linear spectral energies, it better attenuates additive white noise components (Nadeu et al., 2001, p.13).

Noise-cancelling microphone A way to overcome this problem is to use a special noise-cancelling microphone which consists of a pair of microphones picking up the corrupted speech signal, on the one hand, and the pure noise signal, on the other hand. Such a microphone can for example be installed in a helmet used in an air-fighter, where one microphone picks up the speech signal inside the helmet and the other microphone the noise signal from outside the helmet. Use of noise-cancelling microphones in an automobile also result in good speech enhancement. Again, satisfactory noise suppression is only achieved if the noise-cancelling microphone is correctly positioned, which means located in the right angle and very close to the speaker's mouth. Too large a distance of 10 cm and rotation of more than 30 degrees resulted in a decrease of 15 dB speech power for automobile and aircraft interiors (Degan and Prati, 1988, p. 53). Unfortunately, in real-world applications where users demand ease and comfort in application, such microphones would not find high acceptability.

4.3 Robust modeling

Several approaches for handling noise robustness in ASR systems have been introduced above, most of which are based on the assumption that noise is present in the speech signal and either should be removed or its influence should be diminished. Depending on the noise, this has to be done in the spectral, cepstral or in both domains, making it difficult for one and the same approach to account for all different noise cases. Approaches using model compensation or robust training, on the other hand, allow the presence of noise in the recognition process.

4.3.1 Model compensation

In model compensation, HMMs are defined for both the speech signal and the noise signal simultaneously. In *Signal Decomposition* (Varga and Moore, 1990) and *Parallel Model Combination* (PMC) (Gales and Young, 1992), concurrent additive signals, such as the speech and the noise signals, are recognized *simultaneously*. This is achieved through the use of different sets of parallel HMMs where each set models one part of the signal. The clean speech models which are supposed to contain sufficient information about the statistics of the clean training data, are combined with noise models, created with the help of available noise samples, which are supposed to approximate the background noise. The combination is carried out by an appropriate mismatch function to create the *corrupted-speech models*. Recognition is carried out by a three-dimensional Viterbi decoding, extended to a search through the combined state-space of the model sets. The decoder attempts to simultaneously calculate the optimal state sequence in both the speech and the noise models.

In *predictive* PMC, the corrupted-speech models are generated before any speech data from the new environment has been observed. This is achieved by using models of various noise sources, such as additive noise, stress or channel noise models, which are appropriately combined to create the new corrupted-speech models (Gales, 1998).

On the down-side of both models, Signal Decomposition and PMC, it has to be mentioned that the noise models need to be correctly trained for the respective noise and updated for each changed noise environment. Moreover, the extra computational cost involved in three-dimensional Viterbi decoding can be prohibitive.

4.3.2 Robust training

Another possibility to account for various noise conditions at the same time is by training of the recognizer for the various noise cases which are expected to be encountered during the application phase (Hirsch and Pearce, 2000), or by narrow-band training in noise. These approaches are discussed in the following.

Re-training Analogous to the enrollment process for new speakers in a speaker-dependent system a “noise-independent” system can be approximated by re-training the system on new testing conditions. Unfortunately, the collection of new data, which is inevitable for this approach, is expensive and time-consuming. Moreover, the demand of training time of the recognizer itself, which can take hours or even days for large tasks, as well as storage requirements must be considered. The fact that such a time- and resource-expensive re-training would be needed for every new and unpredictable test environment quickly shows the limitations of this approach.

Multi-style training Multi-style training pools training data from different acoustic environments, such as various noise types and levels (Hirsch and Pearce, 2000), different speaking styles (Lippmann et al., 1987) and/or microphones. It was shown that this approach significantly increases recognition performance on the respective noise cases. Performance in a certain noise condition, which was included in the multi-style training data, however, is reduced in comparison to the same system trained only on that noise (Acero, 1990). This increase in error rate for matched conditions is due to a loss of acoustic discrimination through noise contamination. The vast amount of training material, which covers all possible testing environments needed to render a system powerful on all noise cases, is not available and it is questionable whether it could ever be collected.

Narrow-band training A recent approach proposed by Dupont (2000) seems to alleviate this problem of training data sparsity. It is based on two equally important parts, one is training in noise, the other is decomposition of the frequency domain into subbands. In (Cerisara, 1999a; Dupont, 2000), it had been observed that when considering narrow frequency bands, noise occurrences often exhibit little difference other than a difference in noise power. For this reason, a subband speech recognizer which is trained on such a narrow frequency band with any kind of noise will behave relatively robustly to any other kind of noise.

In (Dupont, 2000), a system of seven frequency subbands (each covering roughly four critical bands) was found to work best for this approach. After critical band analysis, each subband feature vector was normalized and passed onto a system (in this case an ANN classifier), the goal of which it is to estimate a feature vector relatively insensible to noise. The output of each such system is a new, more discriminative subband vector. The subband feature vectors are then concatenated and passed onto a standard speech recognizer. For estimation of the robust parameters, the subband ANN classifiers are trained on a noise-contaminated database. In order to cover the entire range of possible noise levels which can be encountered, the training data is assembled from both clean data and data contaminated with white noise at different SNR levels (0, 5, ... , 20 dB). Experimental results averaged over six different noise cases with each at a range of SNR levels (5, 10, 15, 20 dB) resulted in a 30 % relative improvement in word error rate of the seven subband system with contaminated training as compared to reference (fullband and multi-band) systems using spectral subtraction or J-RASTA-PLP features for noise robustness (Dupont, 2000, p. 199).

In (Cerisara et al., 1999a, p. 105), for comparison, only the recombination module (a Single-Layer Perceptron) of a multi-band system was trained in white noise, whereas the subband recognizers were trained in clean. The resulting multi-band system showed higher noise robustness to most of the noise cases tested (i.e. white noise, high-frequency (pink) noise, low-frequency (pink) noise, hair-dryer noise and car noise) than the same system whose recombination unit had been trained on clean data. Tests in clean data, on the other hand, and also in canteen noise, resulted in higher performance by the system whose recombination unit was trained in clean speech than the one whose recombination unit was trained in white noise.

4.4 Missing Data (MD) approach

Missing Data approach Contrary to robust training or training of multiple streams which usually ignores what exactly renders the recognizer more robust, in the *"missing data"* (MD) approach the data to be recognized is analyzed in more detail. In their most general form, MD methods (Cooke et al., 1994b, 1997, 2001; Morris et al., 1998) try to segregate the different sound sources in the input signal, and then to recognize the evidence which has been assigned to each of these speech sources. This implies the necessity to handle so called *occluded speech* when some part of the spectral data frame is assigned to one source and another part to another source.

Identifying reliable data The first problem in MD processing is therefore *how to identify the reliable spectro-temporal regions* for the source to be recognized. This is usually achieved with local SNR estimation. The noise spectrum is estimated for each data frame either just once from the first few supposedly non-speech frames, or adaptively throughout the utterance. Spectral data values for which the local SNR is less than 0 are then treated as if they were not available, or "missing", while other values are treated as 100% clean. The decision that each spectral value is either all speech or all noise is an important part of the basis for the MD approach and is referred to as the *maximum assumption*. This is justified by the fact that for logarithmically (or otherwise) compressed spectral power values a and b, if $a > b$ then $\log(a+b) = \log[(1+\frac{b}{a})a] = \log(a)+\log(1+\frac{b}{a}) < \log(a)+\log(2)$, so if $a \gg 2$, $\log(a+b) \simeq \log(a)$.

Hard and soft "missing data masks" Reliability estimation is performed in a separate pass before recognition starts, and this information is stored as a *"missing data mask"* for each spectral data value in each time frame. The decision as to which values are missing can be either "hard", in which case the MD mask $P(\text{not missing})$ values are discrete ($P(\text{not missing}) = 1$

if SNR > 0, else $= 0$) or "soft" (Barker et al., 2000; Morris et al., 2001a), in which case P(not missing) mask values are continuous, and are estimated by applying a suitable sigmoid squashing function to the estimated SNR (Barker et al., 2000).

Recognition with MD The second problem which arises with occluded speech is *how to carry out recognition on partial data*? It will be seen below that, in the framework of GMMs, HMM state emission probabilities are easily adapted to handle partial data while everything else stays the same. When some part of the clean spectral data X is missing or uncertain, the posterior probability $P(Q|X)$ required for MAP decoding (cf. Section 3.4.1) cannot be calculated directly. In this case the estimate for $P(Q|X)$ which maximizes the probability of correct classification is given by its expected value, conditioned on any knowledge (κ) that we still have about X (Morris et al., 1998, 2001a)

$$\hat{P}(Q|X) = E\,[\,P(Q|X)|X^{tr}, X^{obs}, \kappa\,] \tag{4.14}$$

where X^{tr} is the clean training data, X^{obs} is the observed noisy test data, and κ is any other knowledge on which the pdf for X may depend, such as the "bounds constraint" whereby lower and upper bounds for X are given by 0 and X^{obs}. Decoding with clean speech finds Q^* to maximize $P(Q|X) = \frac{p(X|Q)P(Q)}{p(X)}$, where $p(X)$ can be ignored because its value is not affected by the choice of Q. In this case

$$Q^* = \arg\max_Q P(Q) \prod_t p(x_t|q_k^t) \tag{4.15}$$

In the case where $p(x|q)$ is a GMM, and each Gaussian component $p(x|m_l, q)$ has diagonal covariance, and if i denotes a subscript over coefficients of x, then

$$p(x|q) = \sum_l P(m_l|q) \prod_i p(x_i|m_l, q) \tag{4.16}$$

With missing data $P(Q)$ is unaffected; for the most general case considered in the MD ASR literature, which uses the soft MD decision together with the bounds constraint given above, (4.16) becomes

$$\hat{p}(x|q) = \sum_l P(m_l|q) \prod_i (a_i + b_i) \tag{4.17}$$

where

$$a_i = \Phi_i \, p(x_i|m_l, q) \tag{4.18}$$

with Φ_i denoting the estimated probability that x_i is clean, and where

$$b_i = \frac{(1 - \Phi_i)}{x_i^{obs}} \int_0^{x_i^{obs}} p(x_i|m_l, q)\, dx_i \tag{4.19}$$

assuming additive noise and filterbank energy features, so that the unreliable features are bounded below by 0 and above by the value of the feature in the noisy speech mixture x_i^{obs}. Without additional knowledge, it is assumed that they are distributed uniformly in $[0, x_i^{obs}]$.

In the case where the soft MD decision is used but the bounds constraint is not used, b_i above becomes zero (the integral becomes 1, but the $\frac{1}{x_i^{obs}}$ factor becomes $\frac{1}{\infty}$). In the case where the hard MD decision is used the same equations apply but $\Phi_i = 1$ for x_i present and 0 for x_i missing.

Marginalization Let (x_r, x_u) denote the partition of data vector x into its present or "reliable" and missing or "unreliable" components respectively. When a hard MD decision is used with no bounds constraint, the only difference with the clean data case, as far as MAP decoding is concerned, is that $p(x|m_l, q)$ becomes

$$p(x_r|m_l, q) = \int_{x_u} p(x|m_l, q)\, dx_u \tag{4.20}$$

i.e. each mixture component $p(x|m_l, q)$ is replaced by its marginal $p(x_r|m_l, q)$ by integrating over all of its unreliable components. We will see in Section 6.1.2 how marginalization can also be applied to (likelihood-based) multi-band processing, in order to permit evaluation of the likelihood for any subband combination from the single likelihood model which has been trained for fullband data.

4.5 Multi-band processing

Multi-band processing is one of the major parts of this thesis and Chapter 5 is entirely devoted to this subject. Hence, only a short introduction will be given here.

In multi-band processing, the speech signal in the spectral domain is split into several frequency subbands, in order to separate possibly occurring noise from the clean frequencies. Each subband (or combination of subbands) is then processed separately. The techniques for feature extraction can be the same or different for each subband (Christensen et al., 2000). The subband feature vectors can then

- either be concatenated to form a fullband feature vector which is then processed just as in the standard fullband approach (this is termed *feature combination*), or
- be processed independently by a recognizer for each subband to yield subband (speech unit) probability estimates, which are combined before decoding. (The subband recognizers are based on the same implementation as a respective fullband recognizer). This approach is termed *probability combination*.

In the case of probability combination, the output probability estimates from the subband recognizers can be combined at different levels, such as the state, phoneme, syllable, word or sentence level. In our case, frame level (phoneme) probabilities are estimated by each subrecognizer and combination is carried out on these estimates throughout this thesis.

The strength of multi-band systems lies in the fact that possibly occurring noise in one subband does not get mixed with neighboring subbands, as is usually the case in fullband processing. In fullband processing, feature extraction is carried out only once for the whole frequency domain, which results in a feature vector in which noise in any one subband is usually spread over all features. Multi-band processing permits us to process each frequency subband separately, so that the noise is not spread, and to down-weight or discard noisy subbands – if they can be detected.

The multi-band approach is the subject of Chapter 5. Different combination approaches will be discussed in Chapter 6 and weighting strategies in Chapter 7.

4.6 Multi-stream processing

As we saw in the discussion on psychoacoustic studies on human speech perception in Section 2.4, the speech signal possesses high information redundancy. It is hypothesized and shown by recent publications in this area (Janin et al., 1999; Christensen et al., 2000; Hagen et al., 2000), that the redundancy and diversity of the speech signal can be well exploited by the use of *various fullband streams* representing the auditory (and/or sometimes visual, such as lip movement) input signal (Dupont, 1997; Bourlard, 1999).

Such diversity can be achieved at different processing stages of an automatic speech recognizer the structure of which was illustrated in **Figure 3.1**. In the feature extraction module, diverse strategies can be employed as far as the kind and manner (i.e. window size) and the pre- and post-processing strategies of the feature extractor are concerned. In the design of the classifier, there are also numerous ways of obtaining diverse functional characteristics through: (i) different random initialization or different structure of the models, (ii) different training strategies and/or data (Mak, 1997), (iii) use of different classifiers the outputs of which must be comparable.

Depending on the stage at which the diversity is achieved the combination of the fullband streams can be achieved by simple concatenation of the (fullband) feature vectors, or demands for more sophisticated combination strategies just as it is the case in multi-band processing. *Multi-stream processing can therefore be seen as a generalization of the multi-band approach.* Multi-stream processing is another major part of this thesis and Chapter 9 is devoted to this subject. The same combination and weighting strategies as in multi-band processing can be employed.

The scope of possibilities for multi-stream processing is large. Recently, it was found at our (Hagen et al., 2000) and various other speech research laboratories (Janin et al., 1999; Christensen et al., 2000) that employing several fullband recognizers trained on *different underlying features*, performs better than a monolithic model trained on any one feature stream alone (also in the case when one of the streams was significantly worse than the others). Significant performance improvement and noise robustness can thus be achieved by the combination of complementary fullband streams using different features (Christensen et al., 2000; Janin et al., 1999; Hagen et al., 2000). In this thesis, we concentrate on this approach to multi-stream processing.

An important question which arises also in multi-stream processing of diverse feature streams is up to which level the feature streams should be processed separately. Just as in multi-band processing, the different feature streams could either be modeled jointly to capture correlation or be independently processed up to a higher level for probability combination. Our approach to multi-stream processing will demonstrate a new direction where this question no longer arises but where consistent modeling of both strategies is employed. This approach is the "full combination" approach originally proposed for multi-band processing which is discussed for multi-stream processing in Chapter 9.

4.7 Summary

In this chapter, we described a range of different causes of adversity with which an automatic speech recognizer is confronted. We then came to the discussion of several of the most widely employed and most promising strategies to handle noise-corruption in the framework of ASR. Robust feature processing techniques, such as the MFCC, PLP and J-RASTA-PLP features, and feature post-processing techniques as spectral and cepstral mean normalization, time differentiation as well as LDA were presented. Other approaches to achieve higher noise robustness include the calculation of frequency difference features or the use of spectral subtraction. We discussed speech enhancement such as through noise-cancelling microphones and the approach of model compensation. Robust training and modeling techniques were also described.

MD and multi-band processing were shown to share a similar approach to noise robustness, as they both attempt to separate clean from noisy frequencies, and process them independently. In the MD approach it is tried to separate reliable and unreliable feature coefficients for each frame respectively, and base recognition on the reliable ones only. For this, a noise detection algorithm is needed. In multi-band processing, on the other hand, estimation of the noise is not required as several different frequency subbands are processed in parallel and independently. Possible noise in one band can thus not spread into neighboring bands, and some part of the spectrum will provide reliable information at each time frame. Finally, in the multi-stream approach several fullband streams are being considered as complementary information streams instead of frequency subbands. The latter two approaches are investigated in more detail in subsequent chapters.

In the next chapter, we introduce the paradigm of multi-band processing, before coming to the combination strategies as employed in both multi-band and multi-stream processing in Chapter 6.

Chapter 5

Multi-band speech recognition

In this chapter, we introduce the general multi-band idea which the thesis will build upon. Multi-band speech recognition was inspired by early Fletcher's findings on human auditory perception which suggested that the linguistic message gets decoded independently in different frequency subbands.

In conventional (fullband) speech recognition, as was briefly outlined in Section 3.1, acoustic feature vectors are extracted from the whole frequency band of input speech. Many feature extractors further transform these spectral features linearly or non-linearly to decorrelate the feature coefficients and enhance discriminability between feature classes. This can be done, for example, by applying an inverse Fourier transform (IFT) or DCT to the log spectral values. In the case when some of the feature coefficients are corrupted by noise the noise is spread to the clean feature coefficients, due to these transformations. The probability estimation module (which is normally trained on clean speech) is then confronted with mismatched data, resulting in unreliable probability estimates.

In multi-band processing, the frequency band is split into a given number of subbands which are processed separately to a certain point. In order to divide the speech signal into frequency subbands, the speech waveform is usually first converted from the time to the frequency domain representation. This is typically done by a Fourier analysis followed by (Mel or Bark scale frequency) warping which reflects the human auditory function. In the (warped) spectrum different frequency subbands can be easily distinguished[1]. Feature extraction and possibly decorrelation of coefficients is conducted for each band separately, so that the noise does not spread to neighboring subbands.

[1] In the same way, division into subbands could also be performed by applying a bank of filters directly to the time domain signal.

5.1 Formal view of the multi-band approach

This section discusses the general idea of multiple stream decoding which allows for asynchronous development of the streams and thus different segmentations in each stream, which might more accurately account for the respective sequence of stationary segments.

When combining multiple input streams the different streams are usually supposed to act in a synchronous way, that is, the different parameter groups are assumed to be synchronous. From this follows that the segmentations of the different streams have to be identical, meaning that the HMM state transitions have to occur at same instances in time for all streams. This can be a rather restricting assumption, if the different streams are stationary at different moments in time. Moreover, synchronous processing also implies that the topologies of the HMMs are the same for each stream. In order to circumvent these restrictions, an approach to process multiple, independent input streams was proposed, which allows for asynchronous development of the streams (Bourlard et al., 1996b; Dupont, 2000).

Let us assume that there are I streams of information X_I which are to be recognized. Moreover, assume that model W corresponds to a possible transcription of a given utterance and that this hypothesized model is the concatenation of J sub-models W_j ($j = 1, \ldots, J$) representing the sub-word units. The choice of sub-word unit depends on the level at which the recombination is to be performed, as for example the syllable or phoneme level. To process each stream independently of each other up to the defined sub-word unit level, each sub-word model W_j is composed of I parallel models W_j^i ($i = 1, \ldots, I$). For a given speech unit (j fixed) the different HMMs do not interact with each other, and may have different topologies. The topologies can be adapted specifically to each information stream. The resulting model is illustrated in **Figure 5.1**. The parallel models W_j^i are forced to recombine their respective segmental scores at some temporal "anchor" points (denoted $\otimes$ in **Figure 5.1**).

The recombination states $\otimes$ are not regular HMM states since they are responsible for recombining (according to the rules discussed below) probabilities or likelihoods accumulated over a same temporal segment for all the streams. Since this needs to be done at all possible segmentation points, a special case of HMM decomposition (Varga and Moore, 1990), refered to as *HMM recombination*, has to be used for decoding (Bourlard and Dupont, 1996).

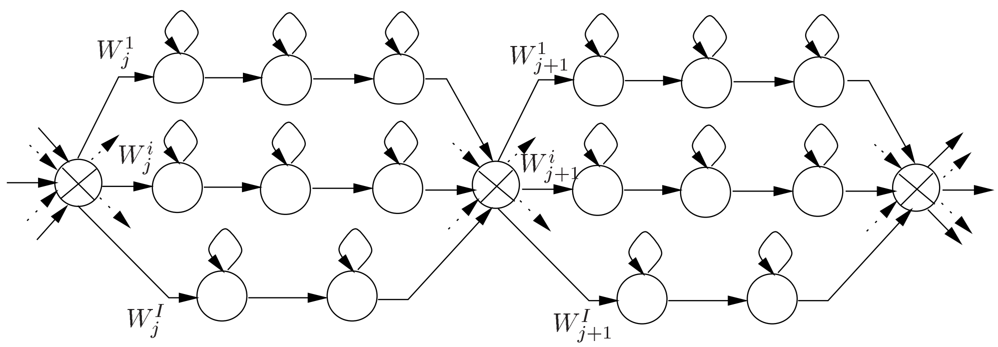

Figure 5.1: General form of an I-stream recognizer with "anchor" points $\otimes$ between speech units (from (Bourlard et al., 1996b, p. 3)).

As described in Section 3.4 for MAP decoding of the most probable word sequence, we follow Equation (3.29). Applying Bayes' rule and separating the acoustic model and the language model, as discussed before, we are left with the need to maximize the likelihood $p(X|W)$. We thus want to find model W maximizing

$$p(X|W) = \prod_{j=1}^{J} p(X_j|W_j) \tag{5.1}$$

where X_j represents the stream sub-sequence associated with the sub-word unit model W_j. Assuming that there are different experts[2] E_i for each input stream X_i $(i = 1, \ldots, I)$ and that these experts are mutually exclusive and exhaustive, it follows:

$$\sum_{i=1}^{I} P(E_i) = 1 \tag{5.2}$$

where $P(E_i)$ is the probability that expert E_i is the best among all experts. It can then be written

$$p(X|W) = \prod_{j=1}^{J} \sum_{i=1}^{I} p(X_j, E_i|W_j) \tag{5.3}$$

$$= \prod_{j=1}^{J} \sum_{i=1}^{I} p(X_j^i|W_j^i) P(E_i|W_j) \tag{5.4}$$

where X_j^i represents the i^{th} stream of the sub-sequence X_j, W_j^i the sub-word unit model for the i^{th} stream, and $P(E_i|W_j)$ the reliability of expert E_i given the considered sub-word unit. In (5.4), it can be seen that, given any hypothesized segmentation, the hypothesis score may be evaluated using multiple experts and some measure of their reliability.

In the specific case in which the streams are assumed to be statistically independent, an estimate of the expert reliability is not needed, as the full likelihood can be decomposed into a product of stream likelihoods for each segment model:

$$\log p(X|W) = \sum_{j=1}^{J} \sum_{i=1}^{I} \log p(X_j^i|W_j^i) \tag{5.5}$$

More generally, any non-linear system could also be used to recombine the probabilities or likelihoods:

$$p(X|W) = \prod_{j=1}^{J} f(\{p(X_j^i|W_j^i), \forall i\}) \tag{5.6}$$

Different combination strategies are possible for f, such as non-linear combination by MLP (Dupont, 2000, p. 100), and other non-linear combination strategies described in Sections 6.4 and 6.5.

During recognition, the best sentence model W maximizing $p(X|W)$ needs to be found. Different solutions have been investigated such as recombination at the sub-word unit level,

[2] In this thesis, an *expert* be defined as a trained classifier which has a fixed set of feature components at its input and outputs class probability estimates.

where the W_j's are the sub-word unit models composed of parallel sub-models, one for each stream, as illustrated in **Figure 5.1**. This requires a significant adaptation of the recognizer, as pointed out above. This can be realized by the *HMM recombination algorithm* (Bourlard et al., 1996b).

State-level recombination Recombination can also be done at the HMM-state level. This can be realized in many ways, including untrained linear, or trained linear or non-linear ways, as the examples given in Section 5.4.1 will show. This amounts to performing a standard Viterbi decoding in which local probabilities are obtained from a linear or non-linear combination of the local subband probabilities. Although this approach does not allow for any asynchrony, we will see in Section 5.4.2 that the synchrony constraint usually leads to at least as good results as when asynchrony between the subbands is allowed. All the work presented in this thesis employs the synchrony constraint by recombining at the frame level. The combination strategies which are investigated in this thesis are presented in Chapter 6.

5.2 The multi-band paradigm

The usual approach to multi-band processing is based on the independence assumption between subbands, so the bands should be chosen in such a way as to minimize overlap between the subband frequencies. In each subband, a vector of features characteristic to this subband is extracted, which is referred to as a feature sub-vector. This is shown in **Figure 5.2**.

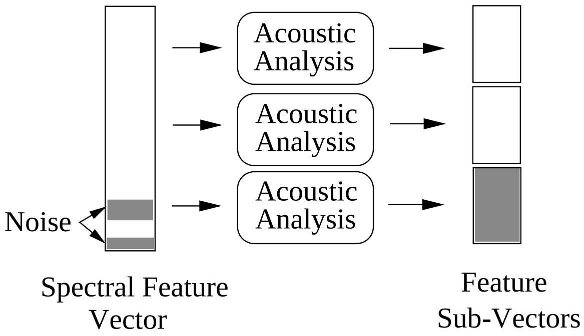

Figure 5.2: An illustration of multi-band processing, where the input speech signal in the spectral domain is split into several frequency subbands which are then processed separately. Noise in some feature components does thus not spread to other components in a different subband.

After orthogonalization or further processing, the feature sub-vectors can be treated in two ways: They are either concatenated and used to replace the original features (*feature combination*), or else each of them is processed by a separate subband recognizer (sub-recognizer) trained on the respective subband and outputting subband probability estimates. In this case, a statistical formalism is needed to treat and recombine the respective probability estimates. This approach is refered to as *probability combination*. We, thus, distinguish between two main approaches to multi-band processing:

1. Feature Combination: In feature combination, the feature sub-vectors are recombined *before* classification.
2. Probability Combination: In probability combination, the subband feature vectors are passed to their respective subband classifier, the outputs of which are recombined *after* classification.

These two approaches are examined in the following sub-sections.

Feature combination

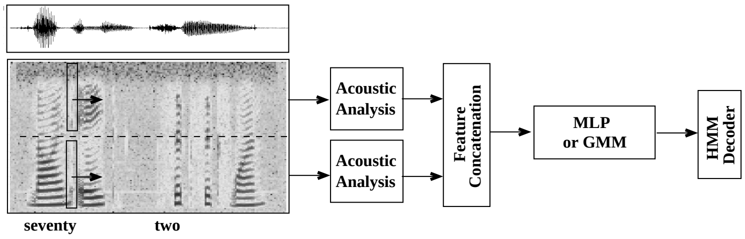

Figure 5.3: Illustration of feature combination for two subbands with following MLP or GMM probability estimation and HMM decoder. The MLP outputs scaled likelihoods or posterior probabilities whereas the GMM outputs likelihoods.

Subband feature combination through concatenation of already pre-processed feature sub-vectors, as illustrated in **Figure 5.3**, was first proposed in (Okawa et al., 1999). This approach offers the advantages that, first, the separately processed feature sub-vectors do not spread noise from one corrupted sub-set into others, as decorrelation and further transformations are carried out on each sub-vector separately. Second, the recombination scheme is rather simple as it only needs concatenation of the feature sub-vectors. We will see in the following sub-section that in the case of probability combination more sophisticated and difficult recombination strategies are necessary. While feature combination is simple, it does not allow the different bands to be weighted separately according to their reliability. This constitutes a weakness in feature combination because, as we see in the experiments (Chapter 8 and 10), an appropriate weighting strategy can lead to performance improvements under certain conditions.

Concatenation of the feature sub-vectors results in one feature vector which can then be modeled as in standard fullband processing. This implies that possible correlation between feature sub-vectors can be captured by the acoustic model (Okawa et al., 1998), which usually renders the model more powerful and reliable. Unfortunately, orthogonalization of the feature coefficients is carried out in the feature sub-vectors only so that the sub-vectors themselves are not orthogonal to each other.

Probability combination

One of the major advantages of multi-band processing — namely the fact that noise from one subband is not spread into the others — is also guaranteed in probability combination, as

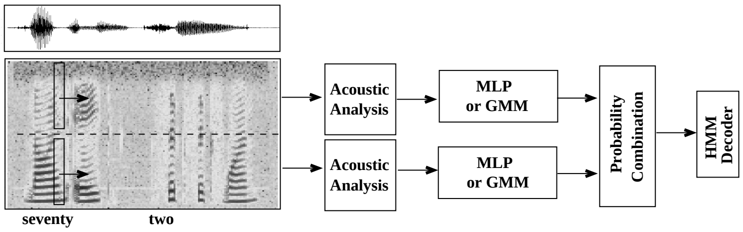

Figure 5.4: Illustration of standard probability combination with MLP or GMM probability estimation for two subbands. The MLP outputs scaled likelihoods or posterior probabilities whereas the GMM outputs likelihoods.

pre-processing for both feature and probability combination is identical. In probability recombination, as illustrated in **Figure 5.4**, each sub-frequency region is treated as a distinct source of information. During recognition, each subband recognizer outputs probability estimates which need to be combined at some level of time segmentation, such as the phone, syllable or sentence level, before the decoding process. Several studies have found that no significant improvement was achieved when recombination was carried out at levels higher than the state level (Bourlard et al., 1996a), even though these approaches make it possible to consider asynchrony between the subbands (Mirghafori and Morgan, 1998b; Cerisara, 1999b; Mirghafori and Morgan, 1999; Cerisara et al., 2000). For this reason, multi-band work in this thesis employs combination at the frame level, which allows for *standard* (one-dimensional) *Viterbi decoding*.

Importance of combination strategy The combination process is an important part in multi-band processing. The right choice in how to combine the probability estimates from the different subband recognizers essentially influences the performance of the combined system. Depending on the nature of the (subband) classifiers, whether they are likelihood-based, such as in the case of HMM-GMMs, or posterior-based, such as in the case of HMM/ANN hybrid classifiers, the statistical formalism changes. A range of well-known strategies, such as the sum and product rule, and the recombining MLP, are discussed in Section 5.4.1. New combination strategies which have been investigated in this thesis are presented in Chapter 6.

Reliability weighting In most combination strategies, a weighting function can, or has to, be employed to measure the relative reliability of each stream, and assign higher weights to the more reliable recognizers. Reliability weighting constitutes a powerful tool in any multiple recognizer system (Jacobs et al., 1991; Jordan and Jacobs, 1994; Hashem, 1997) especially in the case of on-line adaptive weighting, which can account for changing environmental conditions (Hagen et al., 2001). Well-known weighting functions within the multi-band approach are for example based on SNR-estimates (Bourlard and Dupont, 1996) or the Mutual Information criterion (Okawa et al., 1999), which estimate the reliability of each subband relative to the others. Reliability weighting forms an important part of this thesis and is debated in Chapter 7. Experimental results within multi-band processing are presented in Chapter 8.

Training Standard training procedures for HMM-GMMs such as the Viterbi training or Baum-Welch still hold for the subband models. The same is true if HMM/MLP hybrid systems are used for which the respective training algorithms can as well be utilized in this context. Joint training of all stream models and the weights have been proposed (Cerisara et al., 1999b; Kirchhoff and Bilmes, 2000), the discussion of which though lies not within the scope of this thesis.

Choice of subbands Discussions on issues such as the choice of frequency subbands (their number and location as well as possible degree of overlap (Bourlard and Dupont, 1996; Hermansky et al., 1996)), and the choice of subband feature extraction and recognizer techniques, are abundant in other works. They are thus, not discussed in this framework but are briefly described in Section 5.4.2.

5.3 Engineering motivation for multi-band processing

In Section 2.3 we discussed the psychoacoustic motivations for multi-band processing. In this section, other than psychoacoustic motivation to multi-band processing are presented, which comprise improved linear prediction and noise robustness as well as phonetic motivation, and the issues of improved modeling and convergence of smaller subband experts.

Superiority of linear prediction in subbands Subband processing can be widely found in data compression applications such as image or audio compression due to the coding gains observed during subband processing as compared to fullband processing. Such gains arise from the superiority of finite-length linear prediction in subbands to that in a fullband (Rao and Pearlman, 1996). The prediction error variance of the fullband was found to always exceed the total prediction error variance of the combined subbands (Rao and Pearlman, 1996). Moreover, analyzing subband spectra with respect to their spectral flatness measure[3] shows that subband differential pulse code modulation (DPCM)[4] provides coding gain over fullband DPCM (Rao and Pearlman, 1996).

Noise robustness Environmental noises which can be encountered, rarely exhibit the same characteristics in each frequency band. Some noises, such as siren and car noise, are known to mainly affect certain frequencies. One of the prevailing motivation for multi-band processing is the expectation for *increased noise robustness towards (band-limited) noise*. In conventional feature extraction for fullband systems the corrupted speech features are mixed with the clean features from the non-corrupted parts during orthogonalization or other feature post-processing. As we saw, in subband processing this is not the case. Even if other than orthogonalized (cepstral) features were employed, the (fullband) corrupted feature coefficients would affect probability estimation which would no longer result in a reliable output. In the case when the noise does not cover the whole spectrum, subband processing leads to some feature sub-vectors

[3] When the source spectrum is white, the spectral flatness measure (sfm) is 1, and when the source spectrum is maximally correlated, the sfm is 0. The inverse of the sfm is a measure of the predictability of the source. Thus, a smaller value of the sfm means lesser entropy and more predictability.

[4] In DPCM the difference between a data sample and the linear prediction of this sample from past samples is encoded.

being unaffected by the noise; thus, probability estimation in these subbands can be judged as being reliable because the input data corresponds more closely to the (clean) data used for training.

Phonetic motivation In (Ghitza, 1994), it was proposed that different frequency regions have different dynamic characteristics. Through independent processing of each frequency region, the type of feature extraction and the size of the analysis window used in each band can be well adapted to the dynamic characteristics of each frequency subband. In the same way, spectral information which characterizes a certain phoneme is often limited to a specific sub-region of the spectrum. For this reason, certain phonemes could be better modeled if only a restricted frequency region was considered.

Improved modeling In the context of ANN systems, combining the outputs of several trained ANNs is similar to using a single ANN with sub-networks working in parallel. The combination weights are the connectionist weights of the output layer. The main difference between the two setups is that, when the trained sub-ANNs are combined, all connectionist weights are fixed and only the combination weights need to be estimated. On the other hand, when a large ANN is trained, it includes all the sub-networks, and thus a vast number of additional parameters would need to be trained simultaneously. Although increase in computational power and decrease of costs would make it possible to train large ANNs, this can quickly result in over-fitting of the data when the number of the model parameters gets large as compared to the size of the database used for training. For this reason, several smaller models may sometimes be preferable to one large model. In general we can state that the smaller the dimension of the input features, the smaller the size of the models can be. Reduction of the number of necessary parameters can lead to *improved performance* for a given data set, even though information is being discarded, as the mapping in the lower dimensional space can be better specified by the fixed quantity of data (curse of dimensionality) (Bishop, 1995). This often also implies *better convergence* of the training algorithm, as the learning task is easier. High computational power might therefore be better utilized for *parallel processing* of several smaller models than for the processing of one large model at a time.

5.4 Overview of previous research

We now come to the state of the art research in the framework of multi-band processing. We point out the probability combination approaches which are mainly used in these approaches before describing the different investigations in more detail.

5.4.1 Probability combination approaches in previous research

Motivated by Fletcher (1953)'s assumption of independent subband processing in humans, the original multi-band approach and its recombination schemes only consider the individual frequency subbands as streams and for these also only one expert per stream. Later on we will see how also the *combinations* of subbands should be considered. In order to distinguish between these two approaches, we will refer to the first one together with its probability combination strategies as the *"standard" approach to multi-band processing*. In this section we present several standard approaches to the combination of single-subband experts, the standard sum rules, the standard product rules, and non-linear recombination by MLP. As already pointed out, in case of posterior-based systems, after combination the posterior probabilities are divided by the respective class prior probabilities to convert them to scaled likelihoods which can be used in the Viterbi decoder.

Sum rules In this section, we discuss several realizations of the "original" sum rule which have been used so far for posterior-based systems. No mathematically correct derivation can be given as the set of events b_i[5] $(i = 1, \dots, B)$ is not exhaustive, though it was often assumed to be.

The *"standard" sum rule* (STD SUM) for posteriors is written as follows:

$$P(q_k|x) = \sum_{i=1}^{B} P(q_k|b_i, x) P(b_i|x) \quad (5.7)$$

$$\simeq \sum_{i=1}^{B} P(q_k|x_i) P(b_i|x) \quad (5.8)$$

where B is the number of frequency subbands and $P(q_k|x_i)$ the probability estimate from expert i which is trained on subband data x_i. $P(b_i|x)$ is the reliability term which depends on both the expert i and the acoustic vector x.

When the assumption is made that the choice of the best classifier is independent of the input vector $P(b_i|x) = P(b_i)$, expression (5.8) results in the (static) weighted arithmetic mean rule (STD ARITHM MEAN): $P(q_k|x) \simeq \sum_{i=1}^{B} P(q_k|x_i)P(b_i)$. In (Dupont, 2000) further variations of the sum rule are discussed where e.g. a dependency on the specific state q_k is added to the weight yielding $P(q_k|x) \simeq \sum_{i=1}^{B} P(q_k|x_i)P(b_i|x, q_k)$.

These standard sum rules have several disadvantages. First, it was found in ASR that in the case of matched training and testing conditions (i.e. in our case clean test data) smaller frequency bands do not supply the recognizers with enough information to render performance satisfying (Hermansky et al., 1996; Cerisara et al., 1998; Tibrewala and Hermansky, 1997). Too much correlation information is lost so that under matched conditions a multi-band system is no longer competitive as compared to a standard fullband system. Second, in the case of totally mismatched (i.e. corrupted) speech, it might be advisable to disregard all frequency bands and only rely on prior information (for a given speech frame). These extreme but very realistic cases are not considered in standard subband decomposition schemes where the recombination unit forces probability estimates from at least one subband expert. We will see in Section 5.6 how a set of mutually exclusive and exhaustive events can be defined which also takes above cases into consideration.

The respective combination strategies for likelihoods can be derived from the posterior-based formulae via Bayes' rule and are, thus, not explicitly stated here (but are included in the summarizing tables in Appendix D).

[5] "b_i" denotes the event that "subband i must contain the best selection of data".

Product rules for likelihoods Non-linear recombination by product rule is one of the most widely used combination strategies for likelihood (and posterior) estimates. The "standard" product rule for likelihoods (STD PRODUCT) writes as

$$p(x|q_k) \simeq \prod_{i=1}^{B} p(x_i|q_k) \tag{5.9}$$

where $p(x_i|q_k)$ is the stream likelihood of the i^{th} stream of x. Introducing exponential weights, (5.9) results in the standard geometric mean (STD GEOM MEAN) rule:

$$p(x|q_k) \simeq \Theta_k \prod_{i=1}^{B} p^{w_i}(x_i|q_k) \tag{5.10}$$

with $\Theta_k = \frac{1}{\prod_i \theta_{ik}}$ a normalization constant, where $\theta_{ik} = \int_{x_i} p^{w_i}(x_i|q_k)dx_i$.

Product rules for posteriors Turning to the posterior-based approach, the decomposition of the full posterior probability into the respective stream probabilities is derived, assuming conditional independence of the acoustic vector components (cf. derivation of (6.24) in Section 6.2), as

$$P(q_k|x) = \Theta_k\,\Theta\,\frac{\prod_{i=1}^{B} P^{w_i}(q_k|x_i)}{P^{(\sum_i w_i)-1}(q_k)} \tag{5.11}$$

with Θ_k and Θ as in (6.23) and (6.24). After normalizing the right-hand side in order to assure that the posterior probabilities sum to one, this is the geometric mean (STD GEOM MEAN) for posteriors. The second term in (5.11) is independent of q_k and therefore disappears after normalization (such as in (6.23)).[6] In the case when all weights are equal to one, we refer to this approach as product rule (STD PRODUCT) for posteriors.

This rule is essentially different from the often used "product rule" which assumes independence of the posterior probabilities of one class given the data from different streams, which amounts to assuming equal class priors. This rule is for example employed in (Kirchhoff et al., 2000; Kirchhoff and Bilmes, 2000):

$$P(q_k|x) = \Theta \prod_{i=1}^{B} P(q_k|x_i) \tag{5.12}$$

In this thesis, this combination rule is refered to as the "independence assumption" (STD INDEP ASMPT) rule. As we see in the experiments, the product rule (5.11) as derived from the likelihood-based case usually leads to more robust results than the independence assumption for posteriors (5.12).

When recognizer outputs are combined by multiplication, the recognizers which possess low entropy, dominate in the combination more than they dominate when combined by addition. This could be one of the reasons why the product rule, though theoretically based on an incorrect independence assumption, results in good performance of automatic speech recognizers.

[6] For this reason, it is not necessary to assume full independence.

Non-linear recombination by MLP The most promising non-linear recombination scheme is one that can approximate any possible combination function. This can be achieved by the use of an MLP as the combining unit. The inputs to the MLP are, in this case, the probabilities at the output of each of the different classifiers, which are to be combined. The net's outputs are estimates of the posterior probabilities for the considered classes. The training of the MLP parameters Θ corresponds to classical MLP parameter training. We can note this combination scheme by writing

$$P(q_k|x) \quad = \quad f(\Theta, \{P(q_{k'}|x_i), \forall i, k'\}) \tag{5.13}$$

with f the non-linear mapping function realized by the MLP, and Θ the set of MLP parameters.

This approach has several advantages. First, the input to the recombining MLP does not necessarily need to consist of estimates of posterior probabilities or likelihoods of each class but can be any kind of parameters. The MLP classifies the parameters at its input independently of their interpretation. For this reason, one can introduce, for example, linear (Fukunaga, 1990; Haeb-Umbach and Ney, 1992) or non-linear discriminant analysis (LDA or NLDA) (Fontaine et al., 1997) before the recombining MLP in order to reduce the dimension of the input vector. LDA computes discriminant features with the help of a linear transformation of the input vectors into output vectors of smaller dimension such that class separability is maximal. For NLDA it suffices to remove from the classifier MLPs the last layer[7], as each hidden layer of an MLP automatically performs an NLDA (and the last hidden layer of the MLP needs to be smaller than the input layer). Moreover, the recombining MLP can treat, just like every MLP, several temporal frames at once, thus considering context information which could, also for the recombining MLP, help the classification task.

In the case when training and testing conditions are alike, recombination by MLP gives among the best results (Mirghafori, 1999; Hermansky et al., 1996; Bourlard and Dupont, 1996), which is due to its capability of approximating any non-linear function. This is achieved by its large number of parameters which, though, first need to be trained. This combination rule thus involves extensive training and is difficult to quickly adapt to new conditions. As soon as we are confronted with a certain mismatch between training and testing conditions other combination schemes are therefore often more appropriate. One way to adjust a recombining MLP to new testing conditions is by re-training it for the new environment which is time consuming and demands sufficient training data. Besides this, "multi-style" training could be used as described in Section 4.3.2, which however does not result in as good performance as when training for a specific noise. Moreover, in a recombining MLP it is not convenient to down-weight streams in case additional knowledge about the unreliable streams is available. Finally, it is also impossible to incorporate new streams without having to retrain a new recombining MLP.

5.4.2 Description of various multi-band research approaches

Based on the psychoacoustic studies outlined in Section 2.3, several different approaches to multi-band ASR have been investigated over the last years, such as the possibility for asynchrony between the bands, and different feature processing strategies applied in each band.

[7] in the case when there are more or at least two hidden layers.

Work by Nikki Mirghafori Mirghafori (1999) investigated the possibility of developing specialized phone-like classes for each subband. This was based on findings by Ghitza (1994) where phonetic feature transmission was shown to be better in a multi-band system than in a traditional fullband system, possibly due to certain bands distinguishing better among certain phonetic categories. In (Mirghafori, 1999), the subband classes leading to most of the confusions were fused in order to reduce the number of classes and enhance generalization. This resulted in improved frame level discrimination, but no significant reduction in word error rate. During this analysis, it was found that phonetic transitions do not always occur synchronously among subbands. Asynchrony of the bands was investigated on the word- and multiple-state level, which significantly increased computation. Unfortunately, no improvement in word recognition accuracy was achieved. The author argued that by disregarding synchrony between bands, important information might be lost, and that relaxation of the synchrony constraint might only be advisable for very particular phone transitions.

Mirghafori worked with a subband system of four subbands, the recognizers being HMM/-MLP hybrid systems. High recognition performance on the Numbers95 database (cf. Section 8.2) was obtained only when the multi-band system was combined with the fullband system by multiplying the (scaled) likelihoods from both systems (Mirghafori and Morgan, 1998a). This result was enhanced when a PLP-feature based subband system was combined with a RASTA-PLP based fullband system. Combination of two different fullband systems, one using PLP features, one RASTA-PLP features did not result in the same improvement.

Work by Christophe Cerisara Contrary to most of the studies on subband processing, which utilize HMM/MLP hybrid systems as (subband) classifiers, Cerisara (1999a) developed a multi-band system employing (second-order) HMM-GMMs (Mari et al., 1997). The (four) normalized subband likelihoods were combined by themselves, as well as in combination with the normalized fullband likelihood using either a weighted sum or a recombining MLP. It was found in general that the system of all 5 bands (i.e. including the fullband recognizer) outperforms a system consisting of the 4 subband classifiers only, *except* in the case of very low SNR ($<$ 0dB). When the noise level was not too high (above 0 dB SNR), best recognition was achieved when all five streams were combined with a recombining MLP.

Moreover, two algorithms for asynchronous combination of the subbands were proposed, with combination being carried out at the sentence level or after smaller speech segments, such as phonemes. The latter is based on the "two-level dynamic programming" (Rabiner and Juang, 1993) and the "level building" (Myers and Rabiner, 1981) algorithms. Finally, a global training algorithm was developed facilitating simultaneous training of both the HMM-GMM classifiers of all subbands and the recombination unit. The algorithm is based on the Minimum Classification Error (MCE) criterion. Tests in clean and noise-corrupted speech showed performance improvement of the globally trained multi-band system compared to the reference system.

Work by Stéphane Dupont Several acoustic feature extractors as well as combination schemes were investigated in the framework of multi-band and multi-modal speech recognition by Dupont (2000). Special emphasis was placed on robustness to additive noise. In the multi-modal speech recognizer, in addition to a standard speech recognizer, a module for visual analysis of the speaker's lip movement was employed. By the use of different sets of HMMs for each stream of the multi-band or multi-modal system, asynchronous development of the different streams was made possible in two different ways: (i) the HMMs in each stream cooperate with the models in the other streams in such a way as to resynchronize the paths through the HMM states in each stream at pre-defined "anchor" points, as was discussed in Section 5.1. (ii) use of multi-dimensional HMMs in which each composite state consists of a combination of constituent states. The contribution of the composite state is calculated from the contribution of each constituent state. The topology of the composite HMM allows all possible paths to be pursued as defined by the original constituent HMMs. This approach has the advantage that normal (one-dimensional) Viterbi decoding can be employed as combination is carried out at the frame level.

Releasing the synchrony constraint between streams did not result in any robustness gain, just as it was the case in (Mirghafori, 1999). Audio-visual speech recognition, on the other hand, led to a reduction in error rate of up to one third in the case of additive noise (with synchrony constraint) as compared to the acoustic recognizers alone. Finally, the approach of "narrow-band training" as discussed in Section 4.3.2, was developed in this work.

Work by Hermansky et al. In (Hermansky et al., 1996), the authors compare multi-band systems of 2 and 7 frequency subbands on an isolated digits task. Different (linear and non-linear) recombination strategies of the logarithmic likelihoods were carried out, with the non-linear scheme, which uses an MLP for recombination, consistently leading to better performance. It was shown that both subband systems were comparable to the fullband classifier in clean conditions, but were more robust in the case of additive (artificial) sinusoidal noise. The authors also investigated the influence of combinations of subbands on the recognition performance of the multi-band system. Using the 7-subband system, the performance of different setups was compared: (i) sub-sets of the seven sub-recognizers were combined by a recombining MLP (probability combination), and (ii) single sub-recognizers were trained on different combinations of the seven input features streams (feature combination). This resulted in a total of 127 possible combinations for each of the two configurations, only the most promising of which were actually set up and tested. *Manually* choosing the best sub-recognizer for each word, the performance of both systems stayed the same for *all SNR-values* (ranging from clean to 0 dB). This suggests that, at any given point in time, at least one of the 127 combinations exhibits extremely high noise robustness. A similar approach based on (*a priori*) SNR estimates was tested, in which the subbands with an SNR-value below a certain threshold were left out of the recognition task. This resulted in close to optimal performance. A *majority vote* among all available sub-recognizers still yielded about half the error rate of the conventional fullband recognizer.

Work by Hervé Glotin Our colleague Hervé Glotin (Glotin, 2000) investigated in the framework of "full combination" processing (see Sections 5.6 and 6.1 for derivation and explanation of this approach), the interfacing of multi-band models with speech reliability cues like voicing (Berthommier and Glotin, 1999a,b) and localization (Glotin et al., 1999) in order to reinforce robustness.

In (Glotin and Berthommier, 2000), for example, based on the fact that most of the speech segments are voiced, the speech harmonicity cue is exploited in order to derive a time–frequency reliability weighting scheme. This estimation method was evaluated together with direct integration of the *a priori* SNR values. The average word error rates for a panel of noises at different SNR-levels showed that these functions improve recognition in case of stationary band-limited noise. On non-stationary noise, however, they are less efficient compared to constant weighting (that is, when no information about the SNR is given), as well as on wide band noises compared to the baseline (J-RASTA-PLP) fullband recognizer. However, other experiments show that the harmonicity cue is efficient in the case of multi-stream ASR (Neti et al., 2000) (see Section 9.3).

5.5 Limitations of previous multi-band processing approaches

As could be seen in the last section, previous multi-band systems (when used by their own) do not for all noise cases lead to higher performance than a regular fullband recognizer. Under certain conditions even an increase in word error rate is encountered. In the following, we discuss some of the problems with which standard multi-band processing is confronted.

Loss of information Some of the aforementioned advantages of multi-band processing can, under certain conditions, become disadvantages. Reduction of input data, which allows for smaller sized models, also includes severe reduction in information, resulting in smaller recognition rates of each subband recognizer. This can easily be illustrated in the case of stop consonants, the main characteristic of which is that their energy is equally distributed over frequency. This characteristic is lost when the frequency domain is split into separate subbands and thus, these phonemes become rather difficult to recognize. In this respect, not only the decreased quantity of information plays a major role, but also the possible loss of spectral correlation between the subbands is responsible for lower performance of the subband recognizers.

Choice of number, position and features of the subbands Another issue in multi-band processing is the choice of the number and position of the frequency subbands, which usually has to be decided *before* the models are constructed, as well as choice of features which are extracted in each of the subbands. Several studies (Hermansky et al., 1996; Cerisara, 1999a; Christensen et al., 2000) have shown that no "generic" rule of thumb can be established for either of these choices and that the optimal system varies according to the application. The same is true for the recombination module. Its selection is especially crucial, as it is this module which is expected to dampen the errors committed by each of the subband recognizers.

Inadequacy for realistic noise Multi-band systems work well on band-limited (stationary and non-stationary) noise (Hermansky et al., 1996; Bourlard and Dupont, 1997). Unfortunately, it should be mentioned that noises are not as band-limited as would be required to fully exploit the advantage of a multi-band system. In most cases where a standard multi-band system is employed under real-environmental noise, it cannot compete with the performance of a regular fullband recognizer.

Unmotivated additional use of the fullband recognizer To overcome this problem, a certain modification to the standard multi-band approach has recently gained increased popularity. In this modified approach, the standard fullband recognizer is added as another independent recognizer to the multi-band system (Haton et al., 1999; Mirghafori, 1999). This strategy is similar to the combined fullband and subband approach investigated in (Bourlard and Dupont, 1997; Mirghafori and Morgan, 1998a) and yields higher recognition performance than either of the systems alone. Although the combination of the fullband recognizer and the subband system improves recognition performance, in one way or an other, this approach has *no sound mathematical motivation*, and even contradicts the assumption of independence necessary for these multi-band approaches.

In the framework of this thesis, we introduce possible solutions to the mentioned problems of multi-band processing, such as the problem of (i) reduced information in each subband (and loss of cross correlation information), and (ii) the demand for prior choice of position and number of subbands. The proposed solution to the first problem is "full combination" subband processing which is presented in the next section. Its detailed implementations for posterior- and likelihood-based systems will be described in the next chapter. A solution to the second problem is presented in Section 6.1.2 through marginalization applied within the likelihood-based "full combination" approach. Here, any subband likelihoods can be derived from the fullband likelihood without training other than the fullband expert.

We will see in the experiments how the proposed implementations can better account for real-environmental noises as well as artificial band-limited noise.

5.6 Full combination approach to subband processing

As discussed in Section 2.3, Fletcher (1953) had thought to show that humans process frequency bands separately, and that correct recognition in any band leads to correct recognition of the entire input. More recent findings in HSP, however, have shown that high correlation and redundancy exist in the speech signal between both adjacent and non-adjacent frequency regions. Moreover, it was shown that humans can integrate even dispersed frequency information to make significant use of such correlation. This dispersed information can sometimes result in higher robustness than the use of adjacent frequency bands. In psychoacoustic experiments it was found that combination of high- and low-frequencies (including a gap in frequency) often resulted in better recognition performance than a broadening of the low-frequency band by the use of a higher cut-off frequency (thus without gap).

In multi-band ASR it was up to now assumed that subbands could be processed independently, with each subband modeled by a distinct recognizer. In the case of noise-corrupted speech in one subband, correct recognition on the remaining clean subbands could then provide enough information to decode the entire input data. In case of clean speech and speech corrupted with wide-band noise, however, experiments in ASR have shown that a multi-band system of this type very often leads to decreased performance as compared to a fullband recognizer, due to missing cross correlation information. To model more closely what is actually going on in humans and to obtain higher performance in both clean and (wide-band) noise corrupted speech by a multi-band system, we have to find a revised model which also exploits correlation information between (adjacent and non-adjacent) subbands. This should be done by integrating also dispersed frequency information, when some frequency regions are missing, in order to exploit this correlation and redundancy in the spectrum.

Thus, at each time frame, as much clean correlated information as possible should be modeled. In the MD approach (see Section 4.4), noise corrupted frequencies in each frame are detected and excluded, while the remaining reliable data is modeled as a single stream. However, accurate noise detection is very difficult. In the "full combination" (FC) approach taken here, data is divided into subbands and recognition is performed on *every possible combination of subbands*, after which the output from these experts are integrated by one of several possible combination strategies.

The FC paradigm for multi-band ASR

For most application areas, the position of the noise is not known and can be in any subband and any number of subbands. We therefore have to find a way in which we can consider all possible subsets of the frequency domain in order to find the clean dataset.

For this, let us define the set of all possible combinations of B subbands, which include the streams consisting of no, one, two etc. (adjacent and non-adjacent) subbands up to the combination of all subbands, as C, and the set of events b_i ($i = 1, \dots, \mathcal{B} = 2^B$) as follows:

> *$\mathcal{B}$ denotes the set of events b_i that data in combination i is clean speech data, and data not in combination i is completely uninformative and can therefore be regarded as missing.*

On the assumption that each subband is either completely clean or completely uninformative, such a set of events is mutually exclusive and exhaustive, as only one combination of subbands can be the largest clean combination, and one or other must be the true clean combination, because all possible combinations have been considered. Denoting $P(b_i)$ the probability that event b_i occurs, we can write:

$$\begin{aligned} P(\cup_i b_i) &= \sum_{i=1}^{\mathcal{B}} P(b_i) && \text{(mutually exclusive)} \\ &= 1 && \text{(exhaustive)} \end{aligned} \tag{5.14}$$

Note that, in order to be able to refer to a 'subband' and 'combination of subbands' in one term, we use the term '*(data) stream*' to account for both. If some subbands are not corrupted by noise, it is likely that the best stream is the largest combination of clean subbands[8]. However, under wide-band noise conditions it can also be the case that some less noisy subset carries more useful information than the empty set.

Let us now consider how this new FC approach to subband processing can be implemented in a speech recognizer. Considering all possible combinations of subbands means that features have to be extracted not only in the nominal subbands but also in each combination of subbands

[8]This is under the assumption that the stream acoustic models are trained on clean speech only.

which, in the case of B subbands, amounts to $\mathcal{B} = 2^B$ feature streams (note that this includes the empty set). Data within each feature stream can be further processed for decorrelation and/or other transformations, as required[9]. We can then associate with each event b_i an expert i which has at its input the clean data defined by event b_i.

As in usual multi-band processing, we can thus distinguish between *probability combination* and *feature combination* also for the FC approach.

Probability combination In the *usual* multi-band approach, the set of events ($\cup_i$ band i is clean and all other bands are missing) is not exhaustive as it does not cover all possible positions of missing data. In this approach, the fullband posterior probability needed for MAP decoding is estimated through some combination function of the posterior probabilities from B subband recognizers

$$P(q_k|x) = f(P(q_k|x_1)\dots P(q_k|x_i)\dots P(q_k|x_B)) \tag{5.15}$$

Similarly, combination of B likelihood-based subband recognizers can be carried out.

In full combination processing, the probability estimates of all $\mathcal{B} = 2^B$ subband-combination recognizers are needed, where the events $\cup_i b_i \in \mathcal{B}$ are now mutually exclusive and exhaustive. In (posterior-based) probability combination, a recognizer has thus to be trained on each of the $\mathcal{B}$ feature streams, as shown in **Figure 5.5** for the case of two subbands. Realization of the FC approach in posterior- and likelihood-based systems is discussed in Chapter 6.

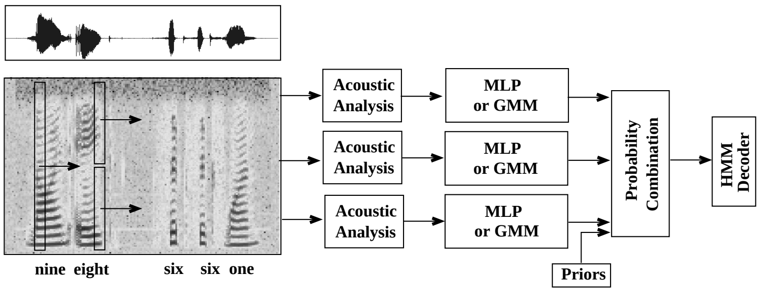

Figure 5.5: Illustration of full combination processing with MLP or GMM classifiers for two subbands. Features are extracted from all possible combinations of subbands.

A further advantage of FC processing over "standard" subband processing is that the question of how many subbands are to be chosen and the exact position of the subbands gets less important as in the FC approach all subbands are considered by themselves *and in combination* and thus correlation between all subbands is considered.

Feature combination For feature combination, all possible feature streams would be concatenated. Due to the highly redundant information in the concatenated feature vector, a principal component analysis (PCA) (cf. Section 4.2.1) should be carried out to extract only the most representative features for each class, thereby orthogonalizing the feature vector and

[9]This is an important advantage of the FC multi-band approach over the MD approach, where decorrelation etc. cannot be applied without spreading noise over all of the data.

reducing its dimensionality. The resulting feature vector is then processed as for a regular fullband recognizer.

Advantages over MD processing

As we saw in the discussion of MD processing (cf. Section 4.4), ASR under noisy conditions can often be improved by simply ignoring the parts of the spectral signal which are most affected by noise. MD processing does not rely on any assumptions about the noise type or level, but, on the other hand, this approach also involves serious drawbacks. First, mismatch is not easy to detect, and, second, the need to avoid mixing clean with noisy spectral coefficients rules out the possibility of data orthogonalization (e.g. the use of cepstral features), which results in unacceptably low performance in clean speech. In FC multi-band processing, we can exploit the advantages of MD processing, that is, ignoring the unreliable parts, while avoiding its disadvantages, because multi-band processing is not so strongly dependent on mismatch detection, and is not restricted to spectral features. This is realized through the use of *multiple* "missing data masks" which, in reality, are the definitions of the subbands and combinations of subbands. This way we do not have to detect the exact mismatch for each coefficient, but can instead integrate over all possible positions of mismatched data, by combining experts trained on each subband combination. Within each subband combination, the spectral features can then also be orthogonalized.

The marginalization approach used with MD techniques (Cooke et al., 1997; Morris et al., 1998) can also be applied to FC multi-band processing which will be shown in Section 6.1.2.

5.7 Summary

In this chapter, we described multi-band processing for ASR where the speech data in the frequency domain is split into several frequency subbands. After feature extraction and possible further acoustic processing, the feature sub-vectors from the different subbands are either concatenated (feature combination) or independently processed for acoustic modeling before probability recombination.

In the framework of HMMs, the parallel subband streams can be recombined at different levels. In case of recombination other than the frame level, special decoding algorithms are needed to allow for asynchrony of the streams. In this thesis, only combination at the frame level is employed, thus standard Viterbi decoding can be used.

We then discussed engineering motivations for the multi-band approach, as well as previous multi-band processing and its combination strategies, such as the standard sum and product rules and the recombining MLP. These approaches are often found to encounter several limitations, due primarily to loss of correlation information, and are generally inadequate in realistic, wide-band noise conditions. These limitations have led us to the development of the "full combination" approach, which not only considers the separate frequency subbands as information streams, but also all possible combinations of subbands. Psychoacoustic and engineering reason motivating the FC approach, and its advantages as compared to MD processing, were discussed.

Combination of the probability estimates from each stream can be carried out according to different rules which will be discussed in the next chapter, together with the details of FC implementation.

CHAPTER 6

Combination strategies

In the previous chapter, we described different strategies for the recombination of subband probabilities. In each of them, we have to combine likelihoods or posteriors according to a function (usually a weighted sum) which often depends on weights representing the reliability of each subband stream. Estimation of the weights will be discussed in the following chapter.

In this chapter, we present the probability combination strategies which were developed in the framework of this thesis. These comprise combination schemes based on the full combination approach introduced in the preceding chapter. Moreover, we present new combination strategies which were motivated from models of human speech perception. Each of the combination strategies is presented for the posterior-based case as well as for the likelihood-based case, where reasonable. The posterior probabilities need to be converted to (scaled) likelihoods after recombination and before the decoding stage. The likelihoods need to be normalized before combination to account for the different range they usually cover. As combination of the probability estimates from each classifier is conducted on the frame level, regular ("one-dimensional") Viterbi decoding can be carried out on the combined (scaled) likelihoods.

Despite the fact that the preceding chapter was concerned with combining subband experts, the expert combination strategies discussed in this chapter are not specific to subband expert combination, but can be applied to combinations of experts trained on any (preferably complementary) data streams. More specifically, the combination strategies presented in this chapter are also an important part of the multi-stream approach which is presented in Chapter 9.

6.1 FC sum rule

In the "full combination" approach to subband processing, which was introduced in Section 5.6, all possible combinations of streams are being considered at each frame in time.

6.1.1 FC posterior decomposition

For posterior decomposition, a separate expert is trained for each of the $\mathcal{B} = 2^B$ possible combinations, where B is the number of frequency subbands. Introducing the hidden variable b_i ($i = 1, \ldots, \mathcal{B}$) indicating which band subset is clean, as defined in Section 5.6, and with the b_i's being mutually exclusive and exhaustive, $P(q_k|x)$ can be expressed as

$$\begin{aligned} P(q_k|x) &= \sum_{i=1}^{\mathcal{B}} P(q_k, b_i|x) && (6.1) \\ &= \sum_{i=1}^{\mathcal{B}} P(q_k|b_i, x) P(b_i|x) && (6.2) \\ &= \sum_{i=1}^{\mathcal{B}} P(q_k|x_i) P(b_i|x) \quad \text{by definition of } b_i && (6.3) \end{aligned}$$

$P(q_k|b_i, x) = P(q_k|x_i)$ follows from the definition of b_i. $P(b_i|x)$ is the *reliability term* for each expert. If b_i is true, then $P(q_k|x_i)$ should be accurately estimated by expert i (which was trained on clean data). Otherwise the estimate will not be reliable. Different approaches to how this weighting term can be estimated are discussed in Chapter 7. We refer to combination rule (6.3) as the (adaptive) *FC weighted arithmetic mean* or simply as FC SUM rule for posteriors.

If we make the assumption that the reliability of a classifier is independent of the input vector then $P(b_i|x) = P(b_i)$ and we arrive at the static FC weighted arithmetic mean, which can be written as

$$P(q_k|x) \simeq \sum_{i=1}^{\mathcal{B}} P(q_k|x_i) P(b_i) \qquad (6.4)$$

If $P(b_i)$ are uniformly distributed, this is just the *simple average* of the outputs from each classifier corresponding to class q_k.

A limitation of posterior-based FC In the case of posterior-based experts (such as MLPs), it is necessary to train 2^B (MLP) experts, and the approach is thus limited to a small number of subbands. We, therefore, propose in Section 6.3 an approximation scheme which estimates the probabilities for each *combination* of bands based on the *single* band experts only.

An advantage of likelihood-based FC We see in the following section how with FC for likelihoods, under certain conditions the stream likelihoods can be derived from the fullband likelihood without training other than the fullband expert.

6.1.2 FC likelihood decomposition

We derive two different equations for FC likelihood decomposition. It will be seen in Section 7.3.4 that these equations, though only slightly different, have important and distinct theoretical advantages.

First possibility for likelihood decomposition We can convert the sum rule for posteriors (6.3) to a sum rule for likelihoods by using Bayes' rule.

$$\frac{P(q_k|x)}{P(q_k)} = \sum_{i=1}^{\mathcal{B}} \frac{P(q_k|x_i)}{P(q_k)} P(b_i|x) \tag{6.5}$$

$$\frac{p(x|q_k)}{p(x)} = \sum_{i=1}^{\mathcal{B}} \frac{p(x_i|q_k)}{p(x_i)} P(b_i|x) \tag{6.6}$$

where

$$p(x_i) = \sum_{k=1}^{K} p(x_i|q_k)P(q_k) \tag{6.7}$$

Expression (6.6) will be refered to as *likelihood-based* FC SUM *rule 1*. This rule shows the necessity in likelihood-based recombination by sum rule to *normalize* the likelihoods $p(x_i|q_k)$ by dividing by the respective density $p(x_i)$. Likelihoods are not comparable before normalization because their scale depends on the different dimensions, and distributions, of the vector x_i in each subband combination.

As decoding is independent of the full data density $p(x)$, it can be ignored on the left side of (6.6).

Second possibility for likelihood decomposition With b_i defined as above we obtain a second likelihood-based FC sum decomposition as follows

$$p(x|q_k) = \sum_{i=1}^{\mathcal{B}} p(x, b_i|q_k) \tag{6.8}$$

$$= \sum_{i=1}^{\mathcal{B}} p(x|b_i, q_k)P(b_i|q_k) \tag{6.9}$$

This, which will be refered to as *likelihood-based* FC SUM *rule 2*, has exactly the same form as (3.10) for a mixture pdf, except that the event m_i = "x is from mixture i" is now replaced by the event b_i = "largest clean combination is i". However, $p(x|b_i, q_k)$ in (6.9) cannot be evaluated directly because b_i tells us that part of x is missing. To overcome this problem we can use the following intuitive approximation

$$\frac{p(x|b_i, q_k)}{p(x)} \simeq \frac{p(x_i|q_k)}{p(x_i)} \tag{6.10}$$

which also avoids the scaling problem which would have occurred if we only had $p(x|b_i, q_k) = p(x_i|q_k)$.

It can be shown by rearranging (6.6) and (6.9) that

$$\frac{p(x|b_i, q_k)}{p(x)} = \frac{p(x_i|q_k)}{p(x_i)} \frac{P(b_i|x)}{P(b_i|q_k)} \tag{6.11}$$

Expression (6.10) is therefore true when $P(b_i|x) = P(b_i|q_k)$, which will often be approximately true. The reliability weights $P(b_i|q_k)$ in rule (6.9) are conditioned only by the class identity q_k, whereas the weights in (6.6) are conditioned only by the local data x. It will be seen in Section 7.3.4 that the approximation (6.10) is essential for the purpose of estimating the ML weights for the likelihood based FC SUM rule 2 (6.9).

Limitation of these two approaches The above two approaches can be implemented as in the posterior-based approach (Hagen et al., 1998; Morris et al., 2001b), with *one expert per event* b_i which amounts to $\mathcal{B} = 2^B$ (likelihood-based, such as GMM) experts that need to be trained. In this way, the number and definition of the subband feature streams have to be defined and fixed beforehand. This is a restricting condition as (i) it can never be known in advance which grouping of subband features (although all combinations of these are considered) will result in the most reliable classification, which moreover can vary from task to task, and (ii) considering all feature vector components as separate streams would lead to too large a number of experts to be trained. If we could find a way to only train one (GMM) expert and then induce during recognition from this expert all possible combinations of reliable subband pdfs, this problem could be overcome[1]. In the following we are going to show how this can be achieved using the marginalization approach as described for MD processing in Section 4.4.

FC likelihood decomposition using marginalization

In the FC SUM rules for likelihoods (6.6) and (6.9), we sum over all possible positions ($i = 1, \ldots, \mathcal{B}$) of reliable subbands. Under the condition that subband combination coefficients are selected from fullband coefficients without further processing (such as orthogonalization within a combination), the parameters for the marginal pdfs $p(x_i|q_k)$ can be obtained directly from the parameters for the fullband pdf by marginalization.

Following the derivation which led to expression of the marginal pdf (4.20) for the data "present" in MD processing we can derive the state likelihoods $p(x_i|q_k)$ for each stream i by integrating over the unreliable, that is, "missing" part $x'_i = x - x_i$ of the data, which is disregarded in the respective stream:

$$p(x_i|q_k) = \int\limits_{x'_i} p(x|q_k)\, dx'_i \tag{6.12}$$

For the mixture pdfs of M mixtures m_j as commonly used for likelihood modeling it holds:

$$\begin{aligned}
\int\limits_{x'_i} p(x|q_k)\, dx'_i &= \int\limits_{x'_i} \sum_{j=1}^{M} P(m_j|q_k) p(x|m_j, q_k)\, dx'_i && (6.13)\\
&= \sum_{j=1}^{M} P(m_j|q_k) \int\limits_{x'_i=-\infty}^{\infty} p(x|m_j, q_k)\, dx'_i && (6.14)\\
&= \sum_{j=1}^{M} P(m_j|q_k) p(x_i|m_j, q_k) && (6.15)\\
&= \sum_{j=1}^{M} P(m_j|q_k) \prod_{l \in s_i} p(x_{(l)}|m_j, q_k) && (6.16)
\end{aligned}$$

where s_i denotes the set of feature coefficients in subband combination i. Note in (6.14) we used $\int_{x'_i=-\infty}^{\infty} p(x|m_j, q_k)\, dx'_i = \int_{x'_i=-\infty}^{\infty} p(x_i|m_j, q_k) p(x'_i|m_j, q_k)\, dx'_i = p(x_i|m_j, q_k)$. In the case

[1] However, the equally important problem of having to evaluate and then combine all of these separate expert probability estimates would remain. See Section 6.3 for ways of reducing the computational complexity of obtaining FC probability estimates.

where each mixture component pdf $p(x|m_j, q_k)$ is modeled as a diagonal covariance Gaussian, with mean μ_{jk} and variance vector σ^2_{jk}, the mean and variance vectors for the marginal pdf $p(x_i|m_j, q_k)$, i.e. μ_{ijk} and σ^2_{ijk}, are simply obtained by striking out the rows and columns from the mean vector μ_{jk} and covariance matrix σ^2_{jk} corresponding to the missing components (Cooke et al., 1994a).

Substituting (6.16) back into (6.6) (and (6.7) for calculation of the normalization factor) we get the *full combination formulae using marginalization* (FC Sum (Marg)) for likelihood-based systems.

In the case when each stream only comprises one feature component, the above implementation of the FC approach can be interpreted in MD terminology as *a weighted sum over all possible sets of hard MD masks* using marginalization without bounds.

Preliminary experiments employing marginalization in FC multi-band ASR revealed that although this avoids the need to train more than one fullband expert, the remaining problem of having to evaluate the marginal likelihood for every combination of subbands is still very computationally expensive, and this prevented us from running further experiments.

Bounded marginalization

Marginalization above involves integration over an unbounded interval for all "missing" data components. Following the MD approach (cf. Section 4.4), we can also apply *bounded* marginalization, so that the observed values of coefficients being treated as unreliable are not completely ignored. In the case where we bound each observation above by its observed value and below by zero, the stream state likelihood for (6.6) (and (6.7)) becomes (continuing from (6.13))

$$p(x_i|q_k) \quad = \quad \sum_{j=1}^{M} P(m_j|q_k) \prod_{h \in s_i} p(x_{(h)}|m_j, q_k) \prod_{l \notin s_i} \frac{1}{x^{obs}_{(l)}} \int_{x_{(l)}=0}^{x_{(l)}=x^{obs}_{(l)}} p(x_{(l)}|m_j, q_k)\, dx_{(l)} \qquad (6.17)$$

Substituting (6.17) back into (6.6) (and (6.7) for calculation of the normalization factor), we get the *full combination formula using bounded marginalization* (FC Sum (Bnded Marg)) for likelihood-based systems.

In the FC approach each combination i can be seen, for one frame, to correspond to one hard MD mask. This shows that no noise estimator is needed in FC processing as it is the case for MD processing, but instead all possible combinations of masks are considered at each time frame. Which coefficients are considered as reliable and which ones as unreliable is thus defined by their membership to a certain combination, i.e. all coefficients within subbands of a combination are reliable whereas all coefficients of subbands not in the combination are unreliable.

6.2 FC product rule

One of the main advantages of the FC approach is the fact that no assumption of independence or conditional independence between subbands is needed. Moreover, by integrating over all possible positions of clean data (in the FC SUM rules) it is assured that for each time frame the most reliable combination of bands is included in the combination process. On the other hand, the assumption that the events b_i (cf. Section 5.6) are exhaustive is open to question[2]. Experimental results have often shown that, despite the limitations of the inaccurate independence assumption between the different recognizers working on each combination of subbands, the recombination by a product can be a more effective method of combining the outputs of multiple classifiers than the sum rule (Haberstadt and Glass, 1998; Christensen et al., 2000; Dupont, 2000; Hagen and Bourlard, 2000; Kirchhoff et al., 2000; Meinedo and Neto, 2000), as will be seen in Chapter 8.

6.2.1 FC product rules for likelihoods

Under the inaccurate assumption of independence between the different recognizers, the full likelihood can be decomposed into a product of $\mathcal{B}$ stream likelihoods for each state q_k ($k = 1, \ldots, K$), according to the FC PRODUCT rule for likelihood-based systems:

$$p(x|q_k) \quad \simeq \quad \prod_{i=1}^{\mathcal{B}} p(x_i|q_k) \tag{6.18}$$

with $p(x_i|q_k)$ the state likelihood of expert i, which was trained on part x_i of data x only.

In a product, a single expert can suppress recognition of a certain class q_k when the probability for this class is close to zero.

The FC PRODUCT in (6.18) can also be implemented as the weighted FC geometric mean (FC GEOM MEAN) of likelihoods, motivated by the fact that the reliability of the input streams can be different[3]. In the weighted geometric mean exponential weights w_i are included for each expert likelihood:

$$p(x|q_k) \quad \simeq \quad \Theta_k \prod_{i=1}^{\mathcal{B}} p^{w_i}(x_i|q_k) \tag{6.19}$$

with $\Theta_k = \frac{1}{\prod_i \theta_{ik}}$ a normalization constant, where $\theta_{ik} = \int_{x_i} p^{w_i}(x_i|q_k)\, dx_i$ so that $\int p(x|q_k)\, dx = 1$. Except for the case where all $w_i = 1$, no probability theoretical derivation of this rule has been proposed in the literature[4]. For this reason the weights are usually constrained to be ≥ 0, but may or may not be made to sum to one.

[2] The full set of events $\{b_i\}$ is only exhaustive if each subband is either 100% clean or 100% uninformative or "missing", which is not necessarily true.

[3] This non-linear combination formula is equivalent to a linear weighted sum of the logarithms of the likelihoods $\log p(x_i|q_k)$.

[4] For the case when all weights are one ($w_i = 1\ \forall\, i$), such as in (6.18), θ_{ik} equals one so that the normalization constant Θ_k also amounts to one.

6.2.2 FC product rules for posteriors

Under the assumption of conditional independence used in (6.18), we can derive for the posterior-based case the FC PRODUCT rule as follows:

$$
\begin{aligned}
P(q_k|x) &= \frac{P(q_k)}{p(x)} p(x|q_k) && (6.20)\\
&= \frac{P(q_k)}{p(x)} \prod_{i=1}^{\mathcal{B}} p(x_i|q_k) && (6.21)\\
&= \frac{P(q_k)}{p(x)} \prod_{i=1}^{\mathcal{B}} \frac{P(q_k|x_i)p(x_i)}{P(q_k)} && (6.22)\\
&= \Theta\, P^{1-\mathcal{B}}(q_k) \prod_{i=1}^{\mathcal{B}} P(q_k|x_i) && (6.23)
\end{aligned}
$$

where Θ is a normalization constant, independent of q_k, such that $\sum_k P(q_k|x) = 1$.

In the case where $p(x|q_k)$ in (6.20) is replaced by the weighted geometric mean in (6.19), we obtain the weighted FC GEOM MEAN rule for posteriors

$$
P(q_k|x) = \Theta_k \Theta\, P^{1-(\sum_i w_i)}(q_k) \prod_{i=1}^{\mathcal{B}} P^{w_i}(q_k|x_i) \qquad (6.24)
$$

with Θ_k as defined for (6.19).

6.3 Approximation of FC (AFC)

Most multi-band systems which are nowadays still employed utilize only one expert per band. As we saw in Section 5.6 the FC approach offers advantages over these one expert per band systems. However, one disadvantage of FC is that the number of experts required can be a problem when the number of subbands is much greater than three or four. We, therefore, propose an estimation strategy which approximates the FC setup, but which only uses the single-stream experts, i.e. the experts from standard multi-band processing. With this, every standard multi-band system can easily be extended to an *approximated full combination* (AFC) system. In AFC we approximate each combination probability from the probabilities from the single band experts which are part of this combination.

Under the assumption of conditional independence between subbands (l) in a combination x_i $p(x_i|q_k) \simeq \prod_{l \in x_i} p(x_{(l)}|q_k)$, we can derive the posteriors $P(q_k|x_i)$ for each subband-combination from the single-subband posteriors $P(q_k|x_{(l)})$ in this combination (i.e. $l \in x_i$) as follows

$$
\begin{aligned}
P(q_k|x_i) &= \frac{P(q_k)}{p(x_i)} p(x_i|q_k) && (6.25)\\
&\simeq \frac{P(q_k)}{p(x_i)} \prod_{l \in x_i} p(x_{(l)}|q_k) && (6.26)\\
&= \frac{P(q_k)}{p(x_i)} \prod_{l \in x_i} \frac{P(q_k|x_{(l)})p(x_{(l)})}{P(q_k)} && (6.27)\\
&= \Theta\, P^{1-|x_i|}(q_k) \prod_{l \in x_i} P(q_k|x_{(l)}) && (6.28)
\end{aligned}
$$

where Θ is a normalization constant independent of q_k, such that $\sum_{k=1}^{K} P(q_k|x) = 1$, that is

$$\Theta = \frac{1}{\sum_{k'=1}^{K} P^{1-|x_i|}(q_{k'}) \prod_{l \in x_i} P(q_{k'}|x_{(l)})} \tag{6.29}$$

where $|x_i|$ is the number of subbands in x_i.

These approximated combination posterior probabilities (6.28) can now be used in any combination strategy where separately trained posteriors are used. For the AFC SUM rule we thus substitute them in (6.3).

6.4 Product of errors rule

Fletcher's multi-independent channel model for human phone perception, summarized in the "product of errors rule", was an early description of the error characteristics of low level human speech perception. It is depicted in Section 2.3. It implies that humans possess an ideal ability to detect which articulation bands are correct and which ones are incorrect. This ability is not (yet) available for ASR due to lack of infallible automatic detection of which recognizers are correct. However, it is surprising that the product of errors rule itself has actually never been used for the purpose of stream combination.

In order to apply Fletcher's *product of errors rule* to recombine stream error probabilities we define the error probability of stream i ($i = 1, \ldots, B$) for phoneme q_k as $\varepsilon(q_k|x_i) = 1 - P(q_k|x_i)$. We can then derive the product of errors rule (PoE) from (2.7) for a multi-band system of B bands as follows:

$$P(q_k|x) = 1 - \prod_{i=1}^{B}(1 - P(q_k|x_i)) \tag{6.30}$$

6.5 Error correction in posteriors combination (ECPC)

A model for quantifying the influence of contextual information on human recognition performance was recently proposed by Bronkhorst et al. (1993) and is presented in detail in Appendix C. This model was set up to describe how humans incorporate time contextual information to correct recognition errors.

In the model, the recognition of a written or spoken word is described as the independent recognition of the constituent letters or phonemes, where each constituent can be correctly or incorrectly recognized. The probability of (correct or incorrect) recognition for all constituents in a word are then multiplied to calculate the probability of correct recognition of the word. It is argued that in HSP, the mis-recognized constituents are then corrected in a subsequent processing step which is modeled in Bronkhorst et al. (1993) by multiplication with a correcting weight. To account for all possible combinations of correctly and incorrectly recognized constituents within a word, it is summed over all possible combinations of these, i.e. over all possible recognition "events" of a word.

Although the authors state that it is not a model for the recognition process itself, we show here how the ideas behind this model can be used in combination with the FC approach.

In our framework, we want to calculate the probability of correct recognition of the *frequency information* of one time frame, and a constituent is thus the phonetic information in a *subband*, which can be correctly or incorrectly recognized.

Subband probability of correct or incorrect recognition In multi-band ASR, we can interpret the probability of correct recognition of a constituent (such as a letter or phoneme) to correspond to the probability of correct recognition of a the information in a subband. The probability of erroneous recognition (for a mis-recognized constituent) then corresponds to the probability of incorrect recognition of a subband by its expert, that is, the expert's error probability.

Following Bronkhorst et al. (1993), a recognition event (which the authors termed "percept" (cf. Appendix C)) consists of a certain combination of correctly and incorrectly recognized subbands. The probability of occurrence of such an event is calculated from the probabilities of *all* (correctly and incorrectly recognized) subbands. Thus, there are as many recognition events as there are possible combinations of correctly and incorrectly recognized subbands. One such an event is illustrated in **Figure 6.1** for an example of five subbands, two of which are corrupted by noise, and thus result in erroneous recognition.

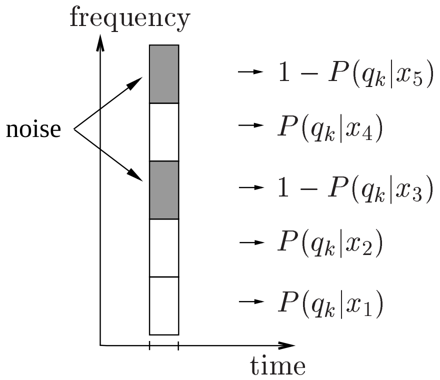

Figure 6.1: Example of a corrupted time frame, where 2 frequency subbands of 5 are corrupted by noise and were mis-classified.

In (Bronkhorst et al., 1993) it is implicitly assumed that the constituents, i.e. the subbands, are statistically independent, so that a recognition event S_j can be expressed by the product of correct subband probabilities $P(q_k|x_j)$ and subband error probabilities $1-P(q_k|x_j)$, which describe this event. For a subband system of B subbands, the set of events S_j thus amounts to

$$
\begin{aligned}
S_0 &= P(q_k|x_1)P(q_k|x_2)\cdots P(q_k|x_B), \\
S_1 &= (1-P(q_k|x_1))P(q_k|x_2)\cdots P(q_k|x_B)+\cdots+ \\
&\quad\; P(q_k|x_1)\cdots P(q_k|x_{B-1})(1-P(q_k|x_B)) \\
&\vdots \\
S_B &= (1-P(q_k|x_1))(1-P(q_k|x_2))\cdots(1-P(q_k|x_B))
\end{aligned}
\tag{6.31}
$$

which includes for each S_j ($j = 0, \ldots, B$) all permutations within the set of $(B-j)$ correctly recognized subbands.

Correcting factors for mis-classified subbands In Bronkhorst et al. (1993)'s model it is assumed that a listener has a certain chance w of correctly guessing a mis-classified letter or phoneme in a word using context. If an automatic recognizer is "asked" to make a (random) guess at an element it previously mis-classified, the probability of correct classification would remain equal to the phoneme prior probability.

Following Bronkhorst et al. (1993), we now multiply the probability of occurrence of each recognition event S_j by the chance w^j of correcting all j error probabilities in an event, and explicitly sum over *all possible events* $i = 1, \ldots, \mathcal{B}$. With this, we obtain the recognition probability of the combined system of B subbands for phoneme q_k as

$$P(q_k|x) \simeq \sum_{j=0}^{B} w^j S_j \tag{6.32}$$

$$\simeq \sum_{i=1}^{\mathcal{B}} \prod_{j \in c_i} P(q_k|x_j) \prod_{l \in c'_i} (1-P(q_k|x_l))\, w^{\varsigma'_i} \tag{6.33}$$

where c_i is the set of correctly recognized subbands, c'_i the set of erroneously recognized subbands of event i ($i=1, \ldots, \mathcal{B}$), and ς'_i the number of mis-classified subbands in set c'_i. We name (6.33) the ECPC[5] (Error Correction in Posteriors Combination) formula.

Combining ECPC and FC

Comparing the FC Sum rule (6.3) to the ECPC formula (6.33) we can see that both (i) add over all possible combinations of (reliable) subbands and then (ii) multiply by a correcting term which is different for each combination. The main difference between the two approaches is that in ECPC the error of the *unreliable* subbands is not discarded but multiplied as error probabilities to the posterior probabilities of the reliable subbands. In ECPC the probability of correct recognition of the reliable data is approximated by a simple product of the reliable subband probabilities, whereas in the FC approach a separate expert is used to estimate the probability for each reliable combination. We were, thus, interested in combining the two approaches in order to not degrade performance due to a coarse approximation of the subband combination probabilities when applying ECPC as is.

For each subband combination i ($i = 1, \ldots, \mathcal{B}$), there also exists a subband expert working on the data which is *not* part of combination i, that is, on $x'_i = x - x_i$. The error probability of this expert is thus used instead of the product of error probabilities of the single-subband experts $l \in c'_i$ in (6.33) to model the error probability of the unreliable data x'_i for each combination. With this we can write the combined FC-ECPC formula as

$$P(q_k|x) = \sum_{i=1}^{\mathcal{B}} P(q_k|x_i)(1 - P(q_k|x'_i))\, w^{\varsigma'_i} \tag{6.34}$$

[5]This can also be read as "Easy-Peasy" meaning "very easy" in English child's speech.

with ς'_i the number of subband error probabilities corresponding to combination i. For each position of clean data x_i ($i = 1, \ldots, \mathcal{B}$), the error of the corrupted part of the data x'_i is not discarded but, for each combination, multiplied as an error probability to the posterior probability of the clean data. It is then summed over all possible combinations i. The error in each event is accounted for through the multiplication by the error correcting term $w^{\varsigma'_i}$. It is initially approximated by $w = P(q_k)$ as motivated above.

6.6 Other combination strategies

Besides the before-mentioned, maybe best-known combination schemes, a large number of other combination strategies can still be found in the literature. In (Kittler et al., 1998; Kirchhoff et al., 2000) for example, the *maximum*, *minimum* and *median* rules are discussed, as well as the *Vote rule*. Moreover, we present the "Union model" as described in (Ming and Smith, 1999; Ming et al., 2000) for its slight similarity to the FC approach.

In this section, B denotes the number of experts being combined, as these rules are not specific to subband expert combination.

Maximum rule The maximum rule (MAXIMUM) approximates the posterior probability of a class by the maximum over the posterior probabilities (for this class) from the different experts:

$$P(q_k|x) = \frac{\max_{i=1}^{B} P(q_k|x_i)}{\sum_{k'=1}^{K} \max_{i=1}^{B} P(q_{k'}|x_i)} \tag{6.35}$$

The maximum rule and the sum rules can be categorized as *OR functions* as the output probability is large when any of the input probabilities is large.

Minimum rule The minimum rule (MINIMUM) approximates the posterior probability of a class by the minimum over the posterior probabilities (for this class) from the different experts:

$$P(q_k|x) = \frac{\min_{i=1}^{B} P(q_k|x_i)}{\sum_{k'=1}^{K} \min_{i=1}^{B} P(q_{k'}|x_i)} \tag{6.36}$$

The minimum rule and the product rules can be described to implement *AND functions* which result in a large output probability only if all of the input probabilities are large.

Median rule The median rule (MEDIAN) is motivated by the observation that the (arithmetic) mean rule is highly effected by wrong classification of outliers. If an outlier class is given a high posterior probability by one of the experts, the mean value will be distorted and thus result in wrong classification. As the median is a robust estimate of the mean, this can be exploited as another, more robust combination rule:

$$P(q_k|x) = \frac{\mathrm{med}_{i=1}^{B} P(q_k|x_i)}{\sum_{k'=1}^{K} \mathrm{med}_{i=1}^{B} P(q_{k'}|x_i)} \tag{6.37}$$

Vote rule In voting, it is counted how many votes each class q_k received from the experts. The class with the highest number of votes is selected. The main advantage of this rule is the fact that it can be used with almost any classifier, for example if they do not output posterior probabilities or likelihoods. In the case where we are given posterior probabilities, the vote rule (VOTE) can be written as follows:

$$P(q_k|x) = \frac{\sum_{i=1}^{B} \Delta_{ki}}{B} \tag{6.38}$$

with

$$\Delta_{ki} = \begin{cases} 1 & : \text{ if } P(q_k|x_i) = \max_{k'=1}^{K} P(q_{k'}|x_i) \\ 0 & : \text{ otherwise} \end{cases} \tag{6.39}$$

a $(k \times i)$-matrix of zeros and ones, where a one in row i indicates for which class q_k expert i had highest posterior probability.

This decision rule can also be described as a "hard-level combination" as the outputs of the experts are first binarized before they are used in the combination scheme. The other three combination strategies mentioned in this section as well as the sum and product rules are "soft-level combination" rules as the estimates of the posterior probabilities from each expert are directly used for the decision (Kittler et al., 1998).

The combination rules can be adapted for the likelihood-based case by using Bayes' rule. The resulting likelihood-based combination strategies are given in **Table D.1** in Appendix D.

Combination by entropy criterion For each frame and each posterior-based MLP expert, the entropy of the MLP is calculated according to

$$\text{Entropy} = -\sum_{k=1}^{K} P(q_k|x) \log P(q_k|x) \tag{6.40}$$

where K is the number of output classes q_k.

The entropy is a measure of the confidence an expert has in its outputs. In the case when an expert is 100% sure about its outputs (that is, it outputs '1' for only one and the correct class, and '0's for the other classes), the entropy value will be zero. In the other extreme case, when an expert cannot decide on any class (that is, all classes receive the same probability), the entropy value is highest ($-\log \frac{1}{K} > 0$). The probability vector which is passed to the decoder is thus the output vector from the expert which had the smallest entropy value.

Combination by "Union model" At this point, we would like to remark on the similarity of a recently introduced model, called the "Union Model" (Ming and Smith, 1999; Ming et al., 2000; Jancovic and Ming, 2001) to the FC approach.

The "Union model" for multi-band noise robust ASR proposes that the likelihood $p(x|q_k)$ is evaluated via combination of B likelihood-based experts trained on subbands $x_1, \ldots, x_B$ as follows

$$p(x|q_k) = p(x_1 \vee x_2 \vee \ldots \vee x_B|q_k) = 1 - \prod_{i=1}^{B} (1 - p(x_i|q_k)) \tag{6.41}$$

The derivation of this model is based on four assumptions:

1. A continuous observation vector can be regarded as a discrete random variable.
2. P(correct) = P(any subband expert correct) (Fletcher's product of errors rule).
3. Subbands $(x_1, \ldots, x_B)$ are independent when conditioned on q_k.
4. For any subband x_i and state q_k, $p(x_i|q_k, \Theta)$ is much smaller when x_i is noisy than when x_i is clean.

and one identity, that for discrete and independent events a_i $(i = 1, \ldots, B)$:

$$P(\vee_i a_i) = P(\text{any } a_i) = P(\text{not none of } a_i) = P(\neg \wedge_i \neg a_i) = 1 - \prod_{i=1}^{B}(1 - P(a_i)) \tag{6.42}$$

Expression (6.41) follows from identity (6.42) under assumptions 1, 2 and 3 above. The attraction of this model is that under assumption 4 the effect of noisy subband x_i in (6.41) will be small, because the factor $(1 - p(x_i|q_k))$ will be approximately equal to one. This way of evaluating the likelihood should therefore be automatically robust to noise. However, assumptions 1 to 4 above are open to the following respective criticisms:

- Probability densities do not follow all of the same rules as probabilities for discrete events. As a result assumption 1 is false and (6.41) would be highly inaccurate even if all of the other assumptions below were true.
- Fletcher's product of errors rule is approximately true for humans when data is divided into two subbands, but in ASR no single one-band expert can match the performance of a fullband expert when all data is clean. Exclusion of the fullband expert will therefore always lead to reduced ASR performance in clean speech.
- Subbands (or multiple data streams) $(x_1, \ldots, x_B)$ are generally not independent, even when conditioned on q_k. Ming et al. (2000) provide formulae which can avoid this assumption of independence, but they do not use them.
- $p(x_i|q_k, \Theta)$ is not always much smaller when x_i is noisy than when x_i is clean. It is not uncommon that noisy data from one phoneme resembles clean data from another noise-like phoneme.

As each $p(x_i|q_k)$ factor in (6.41) is typically much less than one, it is easy to see that this expression, when expanded, is dominated by product terms with smaller numbers of factors, irrespective of whether each factor is noisy or not. The authors of the "Union model" reduced this problem by selecting only terms in (6.41) with the same number M of $p(x_i|q_k)$ factors (or no factors). This they referred to as the "Union model of order M". Optimal selection of model order was then achieved by performing recognition with all model orders and selecting the result from the model for which the duration model gave the maximum likelihood.

The "Union model" shows some similarities with our FC multi-band model. The main differences are as follows. First, our model for combination of posterior probabilities $P(q_k|x_i)$ (6.3) was derived from the rules of probability without making any of the above assumptions (though our AFC model, as presented in Section 6.3, does use assumption 3). Second, when Bayes' rule is applied to (6.3) to obtain the rule for combination of densities (6.6), this rule still differs from the "Union model" rule in a number of important ways. One is that it introduces a weight estimate for each expert. Another is that densities $p(x_i|q_k)$ are always "scaled" by dividing by $p(x_i)$, so that the tendency for contributions from likelihood experts with fewer inputs to dominate the sum does not arise.

6.7 Summary

In this chapter, we presented various strategies for the recombination of stream probability estimates, such as posterior probabilities and likelihoods, stemming from multiple (subband or fullband) recognizers. The "full combination" (FC) rules combine the stream probabilities of a set of mutual exclusive and exhaustive experts. We saw how the posterior-based approaches need the training of the whole set of mutually exclusive and exhaustive experts, whereas in likelihood-based processing the stream likelihoods can (under certain conditions) be easily calculated from the fullband pdf. For the FC sum rules, no assumptions are needed. In the FC product rules the streams are assumed to be independent. An approximation to (posterior-based) FC processing was presented which only employs the single-stream experts, thus being easier applicable for a high number of streams.

We then presented another set of new combination strategies which were motivated from models of human perception: the "product of errors" rule and the "error correction in posteriors combination" approach. The latter was combined with the FC Sum rule so that for each combination the error probability of the expert which has the unreliable part of the data at its input is multiplied to the probability of correct recognition of the expert working on the reliable part.

Finally, a set of well-known combination strategies such as the minimum and the maximum rule were described. In Appendix D, the standard approaches (**Tables D.1**) and the new combination strategies (**Tables D.2** and **D.3**) are summarized.

The newly introduced combination strategies will be evaluated in subsequent chapters in the framework of both multi-band and multi-stream processing. They will be compared to standard combination strategies as well as a regular fullband recognizer. Some combination strategies can or have to employ reliability weighting factors which can further enhance performance. Possible ways for weight estimation will be presented in the next chapter.

CHAPTER 7

Weighting strategies

Multi-band and multi-stream systems can achieve higher noise robustness than a one-stream recognizer already through the diversity and complementarity of their constituent streams, and an appropriate combination strategy. However, all combination strategies discussed in the previous chapter involve weights (reliability of the different experts), which were set to uniform values so far. Another additional possibility to enhance performance of the combined system is through the optimization of weights in the combination process, where the probability estimates from each expert are weighted according to their respective reliability.

In this chapter, we investigate different weighting strategies, comprising both stationary (assuming stationary noise) and non-stationary (assuming non-stationary noise) weights.

We start by presenting weighting functions as proposed in the literature. We then come to the new weighting approaches developed in this thesis. These include estimation of fixed weights, which have to be trained prior to application, and adaptive weights, which are estimated during recognition.

7.1 Introduction

In order to adapt to a changing, acoustic environment, humans use 'perceptual weights' (Arai and Greenberg, 1998) to switch from less to more reliable auditory channels to maintain recognition performance. Similarly, different weighting strategies are also employed in multi-band and multi-stream ASR.

In the derivation of different combination strategies, discussed in the previous chapter, weighting factors may either appear naturally as an essential part of that model or else be introduced as heuristic weighting factors. For example, in the case for the FC SUM and AFC SUM, each expert is assigned a certain reliability which is represented mathematically by $P(b_i|q_k)$. In other approaches, such as the standard or FC product rule heuristic weighting factors are introduced as exponents to the stream probabilities, giving each stream more or less weight.

Streams which work on different sets of input features will generally not carry the same weight of evidence for clean speech, and even more so for noise. Reliability of each stream, even for clean speech (i.e. matched training and testing conditions), depends on the phonemes being hypothesized. In case of noise (mismatched condition), the reliability has to be adapted depending on the kind and position of the noise. The reliability factors or *evidence* weights thus constitute an important part in any recombination scheme where they can be employed. Depending on their derivation and the assumptions which are made in the respective recombination model, the weights can depend on

- the stream index b_i
- the local acoustic observation x
- the speech units (such as the phonemes in our case) q_k
- any combination of the above.

Weights need to be tuned for each multiple recognizer system. If we know that the recognizers will only be used in matched conditions, the weights can be trained on the training data and kept fix during recognition. They do not need to depend on the acoustic observation, and are thus usually chosen to depend on the stream and possibly on the phoneme.

In the case when it is not known what application conditions will be encountered, it is advisable to develop weighting functions which can adapt to the changing conditions.

In the following, we illustrate some of the already proposed and most promising weighting strategies which can be found in the literature. We then come, in Sections 7.3 and 7.4, to the motivation and illustration of the weighting functions which were developed in the framework of this thesis.

7.2 Weighting functions proposed in the literature

7.2.1 Fixed weights used in multi-band and multi-stream ASR

Bourlard and Dupont (1996); Hermansky et al. (1996) propose weights derived from phoneme or word recognition rates, which were obtained from the performance of the individual subband recognizers on a cross validation set. These weights (normalized to sum to one) can be interpreted as representing the relative information content for each speech unit present in each of the subbands. This approach is also motivated from the acoustic-phonetic point of view considering that the acoustic correlates for some phonemes are rather situated in the higher frequency bands (such as fricatives), whereas the distinguishing features for other phonemes are rather situated in the lower frequencies (such as front-vowels). It was thus hypothesized that a certain subband recognizer could better account for some phonemes than for others, i.e. those phonemes whose discriminatory features lie mostly in the frequency region that recognizer was trained on. Weights derived from subband frame level recognition rates are also employed in this thesis in the framework of FC processing, and will be described in Section 7.3.3.

When the dependency of the recognizer weight on speech units is ignored, an average weight (over all speech units) can be calculated for each subband. For performance evaluation, the weights are usually compared to the equal weights approach, which can often give surprisingly good results (Bourlard and Dupont, 1996; Hermansky et al., 1996; Cerisara et al., 1998).

Weight training in "recombining MLP" Non-linear recombination by MLP (cf. Section 5.4.1) of the posterior or likelihood outputs from the subband recognizers can also be interpreted as employing a particular weighting function. Such a "recombining MLP" is especially promising under matched conditions (Bourlard and Dupont, 1996; Hermansky et al., 1996; Cerisara et al., 1998). The "recombining MLP" is usually trained according to the *Least Mean Squared Error* (LMSE), the *Relative Entropy* (Dupont, 2000), or *Minimum Classification Error* (MCE) criterion using a gradient descent algorithm. The LMSE criterion is described in Section 7.3.2. For the MCE criterion (Katagiri et al., 1991; Juang and Katagiri, 1992), a differentiable cost function needs to be introduced, as the number of errors is itself not differentiable. Such a cost function is for example a simple softmax or sigmoid activation function. To quantify the classification error it is combined with a misclassification measure, such as the difference between the probability of the best class and the mean probability of the other classes, or the wrong class which had highest probability (Cerisara et al., 1999b). The cost function is then minimized during training, generally using a gradient descent algorithm which in this framework is refered to as Generalized Probability Descent (GPD) to iteratively estimate the parameter values.

MCE training of weights in linear combination Supervised, discriminative training using the MCE algorithm was also employed to estimate the weights of a (weighted) sum of subband (logarithmic) likelihoods (Beyerlein, 1998; Cerisara et al., 1999b). In (Cerisara et al., 1999b), evaluation of the MCE-based weights in recombination by sum rule did not result in any significant performance difference as compared to a "recombing MLP" trained with the same criterion. In high-noise conditions the "recombining MLP", as well as a simple average of the fullband and all subband recognizers, were more robust.

Moreover, the MCE criterion was applied in a global training scheme where the subband HMM-GMMs and the recombination module were trained jointly. Here, the MCE criterion had to be used as the outputs from the recombining sum do not constitute real likelihoods[1]. As we saw above, the MCE algorithm works on the *difference* between the 'scores' and thus does not depend on any *statistical* interpretation of the 'scores'. We see below other weighting algorithms which employ entropy and mutual information which are defined as probability distributions (Cover and Thomas, 1991) and can, thus, only be used with probability measures. Global training of the subband HMM-GMMs and the linear recombination module resulted in improved performance only in clean speech or in the case when global training was carried out on a similar noise condition as encountered during testing.

[1]The weighted sum of likelihoods does not result in a likelihood as its integral does not amount to one. For this reason, the output of the weighted sum of likelihoods is refered to as 'scores' in (Cerisara et al., 1999b).

7.2.2 Adaptive weights used in multi-band and multi-stream ASR

In the last section, we saw several methods how the weights in the different combination strategies can be optimized. However, these methods are mainly applicable when there is no mismatch between training and testing conditions. If this is not the case, the weights should be optimized in an adaptive manner.

SNR-based weights The principle idea on which these weighting strategies are based is the fact that the higher the mismatch between training and testing condition is, the worse becomes the recognition rate. In the case when the acoustic models are trained in clean speech, it can therefore be assumed that the higher the noise level is, the more the respective model should be penalized. Signal to noise ratio (SNR) measures in each frequency subband can then be used as the basis for adaptive weighting strategies (Bourlard et al., 1996c; Okawa et al., 1998; Dupont, 2000). Such an approach is described in more detail in Section 7.4.1.

Mutual Information Criterion In a likelihood-based multi-band system, Okawa et al. (1999) employ the Mutual Information (MI) between all HMM states and phoneme categories Q and the observation X_T for a certain length of frames T to estimate the recombining weights. Recombination is carried out by geometric mean so that the weights occur as probability exponents. To treat the entropy as a relative value between each subband, posterior probabilities should be used which are approximated by $P(q_k|x_t) \simeq \frac{p(x_t|q_k)}{\sum_{k'=1}^{K} p(x_t|q_{k'})}$ with q_k an HMM state and x_t the observation at one time frame t.

The conditional entropy of all HMM states Q, given the acoustic observations X_T for a certain length of frames T is defined as:

$$H(Q|X_T) = -\sum_{t=1}^{T}\sum_{k=1}^{K} p(q_k, x_t) \log P(q_k|x_t) \tag{7.1}$$

The self entropy of the HMM states Q is

$$H(Q) = -\sum_{k=1}^{K} P(q_k) \log P(q_k) \tag{7.2}$$

With (7.1) and (7.2) the MI can then be calculated according to

$$I(Q, X_T) = H(Q) - H(Q|X_T) \tag{7.3}$$

The MI allows to measure the amount of information contained in Q with respect to the observation X_T. Expression (7.3) is evaluated for each subband to estimate its respective weight. All weights are then normalized to sum up to the number of subbands. The optimal length of frames was evaluated experimentally by using between one frame and the entire sentence. Performance improvement using the MI-based weights could be achieved as compared to equal weighting in both clean and noise-corrupted speech.

7.3 Fixed weights investigated in this thesis

In this section, we present the weighting strategies developed in this thesis which employ fixed weights. The weights are estimated on the training data before being applied in the recognition task. The weights are independent of the test data, but depend either on the phoneme or on the (combination) expert, or both.

7.3.1 Equal combination weights

Combination by equal weights for classification experts has often been used in the statistics community to compare the performance of the combined system to that of the individual experts, and in many cases already improves the resultant model accuracy (Clemen, 1989). The use of equal combination weights is a straightforward approach to the combination of multiple recognizers. Its main advantage is that no data for estimation of the weights is needed and no extra time for the calculation of the weights has to be expended. This approach bases on the assumption that all component recognizers are equally good, which will not always be fulfilled. Especially in the case of subband processing, where experts are trained on different sub-frequency regions, one would expect each expert to perform differently for different speech units (due to their position in the spectrum) and noise conditions (as in the case of high- versus low-frequency noise).

On the contrary though, we will see in the evaluation of the different weighting methods in Chapters 8 and 10 that equal weights often lead to some of the best results among all weighting strategies. One possible explanation (mentioned by Bourlard and Dupont (1997)) is the fact that if one or more subbands are noisy, they will obviously yield noisy local likelihoods or posteriors for *all* classes. Entropy increases so that in the worst case, all local probabilities are the same. As a consequence, it can be expected that during the recognition process, where all these local estimates are integrated over time, the contribution of the noisy bands will appear as a constant in the global probability, for any phone sequence hypothesis, and will thus naturally cancel out when picking the hypothesis with the highest probability.

7.3.2 Least mean squared error (LMSE) criterion

The *Least Mean Squared Error* (LMSE) criterion as usually used for MLP training can also be employed to discriminantly estimate the weights in a (linear or log-linear) combination of multi-band or multi-stream recognizers.

Offline weights estimation using the LMSE criterion with an FC posterior-based system is as follows. For each phoneme q_k, $P(q_k|x)$ is estimated as a linear combination of all of the posteriors $P(q_j|x_i)$ $(j = 1, \dots, K)$ from all of the experts i $(i = 1, \dots, \mathcal{B})$

$$P(q_k|x) \simeq \sum_{i=1}^{\mathcal{B}} \sum_{j=1}^{K} w_{i,j,k} P(q_j|x_i) + w_{0,k} \tag{7.4}$$

Let $d_{x,t}$ be the target or *desired* probability for each phoneme q_k and time frame t,

$w_k = (w_{0,1,k}, w_{1,1,k}, w_{2,1,k}, \ldots, w_{\mathcal{B},K,k})^\mathsf{T}$, $y^t = (1, P(q_1|x_{1,t}), P(q_1|x_{2,t}), \ldots, P(q_K|x_{\mathcal{B},t}))$, $Y = (y_1, y_2, \ldots, y_T)$, $D_k = (d_{k,1}, d_{k,2}, \ldots, d_{K,T})^\mathsf{T}$

For each q_k, the sum of squared errors between (7.4) and the desired output is

$$E_k = \sum_{t=1}^{T} (P(q_k|x_t) - d_{k,t})^2 \tag{7.5}$$

We therefore require

$$w_k = \arg\min_w E_k(w) \tag{7.6}$$

$$= \arg\min_w \sum_{t=1}^{T} (\sum_{i=1}^{\mathcal{B}} \sum_{j=1}^{K} w_{i,j,k} P(q_j|x_{i,t}) + w_{0,k} - d_{k,t})^2 \tag{7.7}$$

Expression (7.6) can be solved by setting $\frac{\partial E_k}{\partial w_{i,j,k}} = 0$ for all $w_{i,j,k}$. This gives rise to the "normal equations"

$$\mathbf{Y}\mathbf{Y}^\mathsf{T} w_k = \mathbf{Y}\mathbf{D}_k \tag{7.8}$$

having solution

$$w_k = (\mathbf{Y}\mathbf{Y}^\mathsf{T})^{-1}\mathbf{Y}\mathbf{D}_k \tag{7.9}$$

If $\mathbf{Y}\mathbf{Y}^\mathsf{T}$ is nonsingular then (7.8) can be solved for a unique vector w_k, but a more robust general solution is given by way of the pseudo-inverse[2]

$$w_k = \mathbf{Y}^+\mathbf{D}_k \tag{7.10}$$

These weights are the same as would be obtained by training on a one layer neural network of linear units under the LMSE criterion.

Initial tests used a simplified form of (7.4) where inputs to the output for $P(q_k|x)$ were limited to posteriors $P(q_k|x_i)$ for $i = 1, \ldots, \mathcal{B}$, but for q_k only, as follows

$$P(q_k|x) = \sum_{i=1}^{\mathcal{B}} w_{i,k} P(q_k|x_i) + w_{0,k} \tag{7.11}$$

This simplified form of LMSE weighting was considered to be of interest because it more closely resembles the FC Sum rule (6.3) than (7.4). Unfortunately, due to time restrictions and the need to get on with testing the more powerful adaptive weighting techniques, the full LMSE weights given in (7.10) were not tested.

The LMSE-weights are evaluated in Chapter 8 in the framework of multi-band processing for posterior-based systems, for which the desired probabilities and the combination posterior probabilities are directly available.

7.3.3 Relative frequency weights

As introduced in Section 7.2.1, the weights can also be derived from the recognition rate of each subband expert. In the framework of FC processing, we evaluate these "relative frequency" (RF)

[2] If the pseudo-inverse is defined by $\mathbf{Y}^+ = \lim_{\epsilon \to 0}(\mathbf{Y}\mathbf{Y}^\mathsf{T} + \epsilon\mathbf{I})^{-1}\mathbf{Y}$, then it can be shown that the limit always exists, and that this limiting value minimizes E_k (Bishop, 1995).

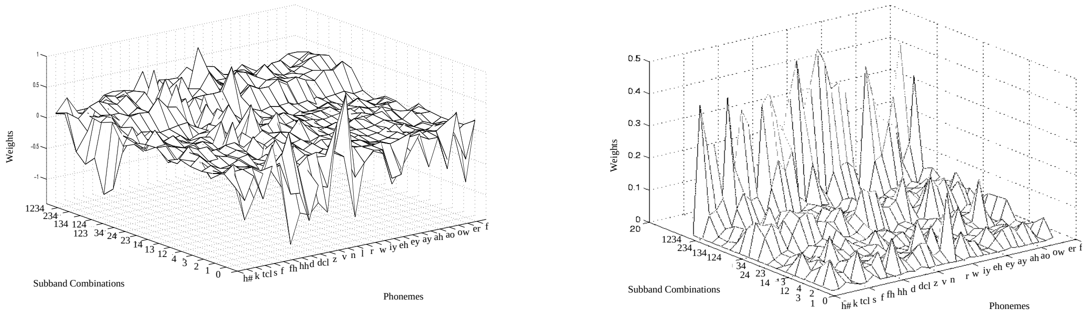

Figure 7.1: Comparison between LMSE- (left) and RF-weights (right). The multi-band system used to train the weights consists of four frequency bands and thus 16 streams altogether. The combinations are denoted by the subband-numbers which are included in the respective stream, such as "123" denoting the stream including subband 1, 2 and 3.

measures on the frame recognition rate of each subband and subband-combination expert. Employing the segmented training data which has also been used to train the experts, we calculate the ratio between the number of times an expert performs best for a given phoneme and the number of times this phoneme occurs in the database. For each stream i ($i = 1, \ldots, \mathcal{B}$) and phoneme k ($k = 1, \ldots, K$) weight $P(q_k|b_i)$ is thus approximated as

$$w_{i,k} = P(q_k|b_i) \simeq \frac{n_{i,k}}{n_k} \tag{7.12}$$

where $n_{i,k}$ is the number of frames of training data for which expert i has the largest posterior probability, over all experts, for phoneme k (and therefore has the smallest Kullback-Leibler distance from the target probability distribution), and n_k is the number of times the phoneme k occurs in the training data. Relative frequency weights thus range between zero and one $0 \leq w_{i,k} \leq 1, \ \forall i, k$.

In **Figure 7.1**, a comparison between LMSE-weights and RF-weights can be seen. As was discussed in Section 7.3.2, the linear LMSE-weights are not restricted to be positive or sum to one, while the RF-weights are all positive and sum to one. Both weighting approaches give, on average, higher weights to the larger stream combinations (employing three or all subbands, in our case). Moreover, it can be observed that some phonemes are apparently better modeled by the single-stream experts whereas others are better modeled by the two-stream experts. This confirms the assumption that each combination models a different sub-set of information and is equally important in the combination process. Both weighting strategies will be evaluated in the experiments to multi-band FC processing which are presented in Section 8.4.

7.3.4 Maximum-likelihood weights

Another criterion to optimize the weights could also be to estimate on some training or adaptation data the weight values that maximize the likelihood of the data given the model, that is (with w the weights to be optimized, and Θ the fixed model parameters)

$$\begin{aligned} w^* &= \arg\max_w p(X|w, \Theta) && (7.13) \\ &= \arg\max_w \prod_{t=1}^{T} p(x_t|w, \Theta) && (7.14) \end{aligned}$$

with $w = \{w_{i,k}\}$, $w_{i,k} = P(b_i|q_k)$, $(i = 1, \ldots, \mathcal{B}, \; k = 1, \ldots, K)$, and $X = \{x_1, \ldots, x_t, \ldots, x_T\}$, where x_t are assumed independent and identically distributed. As we saw in Section 3.3.2 we can decompose any likelihood such as $p(x_t|w, \Theta)$ according to

$$\begin{aligned} p(x_t|w,\Theta) &= \sum_{i=1}^{\mathcal{B}} \sum_{k=1}^{K} p(x_t, b_i, q_k|w, \Theta) && (7.15) \\ &= \sum_{i=1}^{\mathcal{B}} \sum_{k=1}^{K} P(b_i, q_k|\Theta) p(x_t|b_i, q_k, \Theta) && (7.16) \\ &= \sum_{i=1}^{\mathcal{B}} \sum_{k=1}^{K} P(q_k|\Theta) P(b_i|q_k, \Theta) p(x_t|b_i, q_k, \Theta) && (7.17) \end{aligned}$$

Expression (7.16) corresponds to decomposition (3.16) of the data likelihood into a weighted sum of e.g. Gaussians, where in our case $P(b_i, q_k|\Theta)$ now is the weight for expert i and phoneme q_k. We can therefore apply the same EM weight update rules that were used in Section 3.3.2 for GMM mixture weights estimation, for estimating $P(b_i, q_k)$, from which the expert combination weights $w_{i,k} = P(b_i|q_k)$ needed in FC SUM rule 2 (6.8) can be obtained as

$$w_{i,k} = P(b_i|q_k) = \frac{P(b_i, q_k)}{P(q_k)} \tag{7.18}$$

Following the mixture weight update formula (3.24), we can obtain an updated estimate for the weights $P(b_i, q_k|\Theta)$ in (7.16) as

$$P(b_i, q_k|\Theta) = \frac{1}{T} \sum_{t=1}^{T} P(b_i, q_k|x_t, \hat{\Theta}) \tag{7.19}$$

where

$$P(b_i, q_k|x_t, \hat{\Theta}) = \frac{p(x_t|b_i, q_k, \hat{\Theta})}{p(x_t|\hat{\Theta})} P(b_i, q_k|\hat{\Theta}) \tag{7.20}$$

follows from Bayes' rule, as in (3.25).

The only problem here is that $p(x_t|b_i, q_k)$ in (7.20) cannot be evaluated directly because b_i tells us that part of x_t is missing. To overcome this problem we use the approximation (6.10) to give

$$P(b_i, q_k|x_t, \hat{\Theta}) \simeq \frac{p(x_{i,t}|q_k, \hat{\Theta})}{p(x_{i,t}|\hat{\Theta})} P(b_i, q_k) \tag{7.21}$$

where

$$p(x_{i,t}|\hat{\Theta}) = \sum_{k=1}^{K} p(x_{i,t}|q_k, \hat{\Theta}) P(q_k) \tag{7.22}$$

The weights $w_{i,k} = P(b_i|q_k)$ can now be estimated from the training data using (7.18), (7.19), and (7.21) as well as the prior probability of each class $P(q_k)$, and initial weight estimates. The weights can be initialized either all equal (which was done in the experiments described in the next chapter), random, or as fixed weight values using methods from the previous sections (RF or LMSE weights). For experimental evaluation of these weights see Section 8.4.2.

Remark The likelihood-based FC SUM rule 1 (6.6) does not permit ML estimation of the combination weights $P(b_i|x)$ because it is non-linear in the likelihood components $p(x_i|q_k, \Theta)$ and does not lead to a closed form for EM estimation equations. If required, the weights $P(b_i|x)$ needed in (6.6) could be estimated by summing $P(b_i, q_k)$ in (7.19) over all q_k.

7.3.5 Quasi-optimal weights

In order to obtain some idea of the best performance a system of multiple recognizers can achieve with an *optimal* weighting strategy, we can artificially define the weights in such a way as to always choose the recognizer which has the highest output probability for the class which is known to be correct. This corresponds to a one/zero weighting scheme where the best recognizer receives all the weight and all others are excluded from the recombination process by giving them zero weight. Such a weighting strategy is of course not applicable to unknown data, but can be employed for evaluation when the correct class label for each time frame is known.

The weight for expert i for time frame t and phoneme k is calculated as

$$w_{i,k}(t) = \begin{cases} 1 & : \text{ if } P(q_l|x_{i,t}) = \max_{i=1}^{\mathcal{B}} P(q_l|x_{i,t}) \text{ for } l \text{ the correct class} \\ 0 & : \text{ otherwise} \end{cases} \tag{7.23}$$

These quasi-optimal weights allow us to obtain a target baseline performance against which we can compare the performance of new weighting strategies in Section 8.4.

7.4 Adaptive weights developed in this thesis

In the preceding section, we saw several weighting schemes where the weights are fixed in advance and no adaptation during recognition is carried out. Fixed weights can be chosen heuristically or trained on the training set of the database used to train the recognizers. As the source of mismatch between training and test conditions cannot always be anticipated, fixed weights may not be appropriate. In mismatched conditions, weights are needed which can adapt to and account for the change in application environment by gradually penalizing recognizers dependent on their specific reduction in performance.

In this section, we therefore propose adaptive weighting strategies. The first scheme works on the input signal of each stream, that is the acoustic speech data, and assumes that unreliability is due to noise corruption. The second weighting scheme works on the probability output of each recognizer and estimates weights through observation of the development of the recognizer outputs.

7.4.1 SNR-weighting

As we saw in Chapter 4 some of the most severe mismatch conditions arise when the speech recognizer, which was trained on clean speech, is to be applied in a noisy environment. In the case of unknown (additive) noise, some recognizers will degrade more than others depending on the frequency location of the noise and on the noise level at each location. Recognition of the combined multiple stream system will improve when each recognizer is weighted depending on the SNR level encountered in the data the recognizer is working on.

For each time frame, an estimate for the reliability of a stream j ($j = 1, \dots, B$) which is corrupted by additive noise can be based on the signal to noise ratio (SNR) estimated in that stream, which we denote by $\widehat{\text{SNR}}_j$. To estimate $\widehat{\text{SNR}}_j$ for a stream j we can use the estimated noise spectrum $|\widehat{N}_j(f)|^2$ and the observed spectral value $|Y_j(f)|^2$ of stream j and compute

$$\widehat{\text{SNR}}_j = 10 \log\left(\frac{|Y_j(f)|^2}{|\widehat{N}_j(f)|^2} - 1\right) \tag{7.24}$$

The noise estimate can for example be obtained during the first 100 ms of the input utterance under the assumption that no speech is yet present. Similarly, the noise estimate could also be gained (and updated) in speech pauses which is more appropriate for non-stationary noise cases, but this demands a speech/silence detector. Other estimation algorithms which do not explicitly depend on silence portions are for example described in (Hirsch, 1993; Martin, 1993; Dupont, 2000). They are based on the assumption that the noise is more stationary than speech, so that the noise can be considered stationary in speech segments of several frames in length.

In deciding which streams are clean we define two SNR thresholds: a lower threshold $\text{SNR}_{\min}$ below which a stream most certainly leads to unreliable performance of a recognizer which has this frequency stream at its input, and an upper threshold $\text{SNR}_{\max}$ above which a frequency stream most certainly leads to reliable performance of a recognizer which has this frequency stream at its input. The reliability of a stream j can then be estimated by

$$P(j\ \textit{reliable}) = \frac{\min(\max(\widehat{\text{SNR}}_j, \text{SNR}_{\min}), \text{SNR}_{\max}) - \text{SNR}_{\min}}{\text{SNR}_{\max} - \text{SNR}_{\min}} \tag{7.25}$$

As in (Morris et al., 1999), the lower threshold was fixed to $\text{SNR}_{\min} = 0\ dB$, below which $P(j \text{ reliable}) = 0$, and the upper threshold to $\text{SNR}_{\max} = 30\ dB$, above which $P(j \text{ reliable}) = 1$.

In the framework of FC processing, we need weights not only for each subband but also for each combination of subbands. It is usually reasonable to assume that the reliability of a certain combination of subbands can be estimated from the reliability of each of its component subbands, and that subband reliabilities are independent. We thus derive the weight for each combination c_i ($i = 1, \dots, \mathcal{B}$) of streams j ($j = 1, \dots, B$) from the probability that all its constituent streams are reliable ($j \in c_i$) and all streams *not* in the combination ($j \notin c_i$) are unreliable:

$$b_i \Leftrightarrow (j \text{ reliable } \forall j \in c_i) \wedge (j \neg\text{reliable } \forall j \notin c_i) \tag{7.26}$$

with event b_i as defined in Section 5.6. Assuming that the noise in each stream is independent, $P(b_i|x)$ can then be approximated by

$$w_i = P(b_i|x) \simeq \prod_{j \in c_i} P(j \text{ reliable}) \prod_{j' \notin c_i} P(j' \neg \text{ reliable}) \tag{7.27}$$

The experiments employing these SNR-based weights in our multi-band FC HMM/MLP hybrid system are presented in Section 8.4.3.

7.4.2 Adaptive Maximum-Likelihood weights

In Section 7.3.4, we have seen that it is possible to estimate the combination weight values to maximize the likelihood of some training data. Consequently, as for any likelihood-based training system, it is possible to develop an adaptive version where the weights are adapted, in an online and unsupervised way, during recognition.

In our case, we would like a weights estimate which is able to adapt to rapidly changing noise conditions. However, the shorter the interval of time N used for updating the local weights estimate, the less reliable it will be. By combining the local with the previous estimate, these inaccuracies can, to some extent, be smoothed out. A local estimate can be obtained using the offline ML-based weights from (7.18). Let $w_{i,k}^{N}(t)$ denote this local weight estimate for expert i ($i = 1, \ldots, \mathcal{B}$), and phoneme q_k ($k = 1, \ldots, K$), which was estimated using N data frames leading up to the current time frame t. Let $w_{i,k}(t-1)$ denote the previous weight estimate from time frame $t-1$. We can combine these in a weighted sum

$$w_{i,k}(t) = \alpha\, w_{i,k}(t-1) + (1 - \alpha)\, w_{i,k}^{N}(t) \tag{7.28}$$

where α is in $[0, 1]$ and $w_{i,k}(0)$ is initialized from the fixed ML weights (7.18) or as equal weights. These weights are evaluated in our multi-band HMM-GMM system in Section 8.4.4.

7.5 Summary

Performance of a multi-band or multi-stream system can be further enhanced when the stream probability estimates are weighted in the recombination process according to their respective reliability. This weight reflects the confidence we can have in a probability estimate at the output of an expert. Weights can either be trained offline or estimated during recognition.

In this chapter, we first developed a group of fixed weighting strategies. The first set of weights was estimated by minimizing, over the training data, the mean squared error between the posterior estimates of all streams and the respective phoneme, and the desired output for that phoneme (at a given frame). The second set of weights represent relative frequency measures, where the number of frames of training data are counted for which an expert has the largest posterior probability, over all experts, for a given phoneme. Next, we derived for likelihood-based systems ML-based weights by applying the EM algorithm for the estimation of the combination weights. Finally, for evaluation purposes, "quasi-optimal" weights were defined which always choose the recognizer which has the highest output probability for the correct class.

The first set of adaptive weights investigated in this chapter apply to multi-band systems and were based on SNR estimates. In each subband, the SNR is measured and compared to upper and lower thresholds. From these reliability estimates for each subband, the reliability of each subband combination was calculated. The second set of online weights constitute an adaptive version of the ML-based weights, where the weights are continuously updated during recognition on several frames of the test data. These weights are also applicable to multi-stream processing.

The different weighting strategies will be evaluated experimentally in the next chapter in the framework of the new combination strategies which were developed in the preceding chapter.

Chapter 8

Experimental evaluation of multi-band processing

In this chapter, the experiments with multi-band processing are presented. First, the multi-band systems together with the baseline fullband recognizers are described. The database used for training and testing of the recognizers as well as the artificially added noise cases are illustrated next.

We are seeking multi-band systems (and multi-stream systems in general) which perform competitively with a state-of-the-art fullband recognizer in clean speech, but which provide higher noise robustness to a large variety of noises. Thus, in the next section, the newly proposed combination strategies are compared against each other, and also against the baseline fullband recognizer, and some previous multi-band combination strategies.

After this, we compare the performance of the new weighting schemes which were discussed in the previous chapter. Experiments are carried out on the clean condition and various noise conditions.

8.1 Description of multi-band systems

As further described later, all our experiments were done with different, state-of-the-art recognizers, HMM/MLP hybrid systems and HMM-GMMs, each using the acoustic features yielding the best performance.

The multi-band systems employed in this thesis consist of four subbands, the exact frequency definitions of which are given in **Table 8.1**. There are several reasons for choosing four subbands. First, the use of four subbands is motivated by the idea of including roughly one formant in each frequency subband. Second, the choice is historically founded as most of the subband work carried out at our institute and partner institutes successfully employed four subbands (Bourlard and Dupont, 1997; Hagen et al., 1998; Mirghafori and Morgan, 1998b;

Christensen et al., 2000). Moreover, comparison to subband systems with less or more subbands had for example been studied in (Bourlard et al., 1996b; Bourlard and Dupont, 1996) and indicated four subbands to be a good choice. Thus, to be consistent in the development of our work, we continue on the same track. Finally, the same four subbands can be used to realize the FC approach which, hence, incorporates the training of 15 classifiers[1].

As compared to earlier work carried out at our institute, only the exact position of the four subbands is changed. In earlier work, the subbands were often overlapping. As in some combination strategies employed in this thesis, such as the product of errors rule (cf. Section 6.4), independent frequency subbands are assumed, the definition of the subbands was changed to include each critical band in only one frequency subband. However, due to the filter characteristics of the critical bands, the resulting subbands still interleave to a small extent.

Choice of Features Two different sets of acoustic features were used for the HMM/MLP hybrid multi-band systems: PLP and J-RASTA-PLP features, which were introduced in Sections 4.2.1 and 4.2.4. Both are based on LPC analysis, the respective order (LPC ORDER) of which is given for each frequency subband in **Table 8.1** (and **Table E.1** in Appendix E for all possible combinations of subbands). The prediction coefficients are converted to cepstral coefficients for decorrelation. The number of cepstral coefficients (CC) is also indicated in the tables. The values of both parameters were chosen according to the general rule given in (Rabiner and Juang, 1993, p. 116), and in proportion to the size of each subband or subband combination. For the FC approach, the parameters for feature estimation for the subband combinations were directly derived from the parameters of the subband feature vectors included in the respective combination. An example is given in **Table 8.1** for band combination 134.

The PLP-cepstral coefficients (after J-RASTA filtering in case of J-RASTA-PLP features) are the input to our classifiers, together with first and second order derivatives (including energy), if not stated otherwise. The features were extracted from windows of 25 ms length, with a shift of 12.5 ms.

For the HMM-GMM multi-band system, we used MFCC features (as described in Section 4.2.1) which we found to work best in this setup. The features are extracted on the same frequency subbands as given in **Table 8.1**. We employed ten filters extracting six coefficients in each band.

Subband experts An expert is associated with each subband, and, to be explicit, in the case of FC processing, with each combination of subbands. Each expert estimates a vector of parameters, posterior probabilities in the case of MLP experts, and likelihoods in the case of GMM experts. The estimates from all subband experts (and possibly subband-combination experts) are recombined, and then used in an HMM-based recognizer to decode the speech input. Evaluation of the different recombination strategies discussed in Chapter 6 is the task of the experiments described in the next sections.

[1] $15 = 2^4 - 1$ as for the stream consisting of prior information only no MLP needs to be trained.

BAND NUMBER	CRITICAL BANDS	DEFINITION IN Hz	LPC ORDER	NUMBER OF CC	HU	MLP PARAM.
1	2-5	115.3-628.5 Hz	3	5	1000	189 000
2	6-9	565.3-1369.9 Hz	3	5	1000	189 000
3	10-12	1262-2292.4 Hz	2	3	666	89 910
4	13-15	2121.7-3768.8 Hz	2	3	666	89 910
134	2-5 10-15	115.3-628.5 Hz, 1262-3768.8 Hz	7	11	1620	568 620
FULLBAND	2-15	115.3-3768.8 Hz	11	12	1750	661 500

Table 8.1: Definition of the frequency subbands as employed in our multi-band systems, together with the parameters used in feature extraction. The number of parameters are the same for PLP and J-RASTA-PLP features. The full information including all combinations of subbands is given in **Table E.1** in Appendix E. LPC: LPC analysis order; CC: number of cepstral coefficients; HU: number of hidden units; MLP PARAM.: number of MLP parameters.

HMM/MLP hybrids Each MLP expert is provided with nine consecutive frames of input, centered around the current frame. Training procedure by error back-propagation (Hush and Horne, 1993) is the same for all MLP experts, as well as the following architecture: an input layer comprising the nine respective feature vectors, one hidden layer of a fixed number of hidden units (cf. **Tables 8.1** and **E.1**), and an output layer, the size of which corresponds to the number of (one-state) phonemes in the database, which is 27. The size of the hidden layer of each MLP is chosen proportional to the size of its feature vector, varying between 660 and 1750 hidden units.

One HMM state is associated with each MLP output with the HMM states corresponding to the phoneme classes. We used context-independent phoneme models consisting of one to three repetitions of a phoneme state for duration modeling. For this, the number of states is determined from the average length of the phoneme as found in the training data. The phoneme models have fixed transition probabilities of 0.5 for each transition. Word models are constructed by concatenation of the constituent phoneme models according to the (single-pronunciation) dictionary.

The recombined posterior probabilities theoretically need to be divided by the class prior probabilities to obtain (scaled) likelihoods for Viterbi decoding. As it has been found during experimental evaluation by our institute and others, this division does not always lead to improved performance, depending on the respective features, database and other conditions. For this reason we evaluated in preliminary experiments for which system the division by priors was necessary. This is the case for the J-RASTA-PLP-based recognizers. In the case of PLP features, no division by priors is carried out in any of the experiments.

HMM-GMM systems In the HMM-GMM systems we used Gaussian mixtures as continuous observation densities (Juang et al., 1986). There are 78 Gaussian Mixture Models (GMMs), each modeling a 3-state triphone using 64 Gaussian mixtures and diagonal covariance matrices. Each multi-band HMM-GMM comprises the same number of feature coefficients and GMM parameters in order to render their likelihood estimates more comparable. The triphone models in the HMM-GMM systems are necessary to render them competitive to HMM/MLP hybrids.

Viterbi Decoder Standard Viterbi decoding is used for both HMM/MLP hybrid and HMM-GMM systems. Most of the manually adjustable decoder parameters were kept at default values except the word entrance penalty which is, for each set of experiments and each recognizer, adapted on the *clean* test data to yield the lowest word error rate (WER), and kept constant for all tests on noise.

8.2 Description of the experimental setup

The experiments in this thesis utilize a telephone speech database to which different noise cases were artificially added, in order to study the damaging effects caused by corruption by *additive noise*. This setup is well suited when experimenting with new techniques developed for robust automatic speech recognition which can be better evaluated when the noise occurrences and characteristics are known.

8.2.1 NUMBERS95 database

The NUMBERS95 corpus (Cole et al., 1995) consists of naturally spoken connected digits, pronounced by American English speakers. Utterances were recorded over the telephone and hand-labeled with phonetic transcriptions by trained phoneticians.

The database is divided into two independent subsets: the training set (including a cross-validation set) and the test set. The training set consists of 3590 utterances comprising approximately three hours of speech and is used to train the MLP and GMM classifiers. This partition corresponds to the one also chosen by other institutes (Mirghafori and Morgan, 1998a; Dupont, 2000). Phonetic segmentation was provided by one of our project partners, the TCTS Lab at the Polytechnical University of Mons, Belgium[2]. It is based on the CMU dictionary 4.0 consisting of 46 phonemes (a subset of the TIMIT phonemes), 27 of which are used in our sets. An exact description of the creation of the segmentation can be found in (Dupont, 2000, p. 158). Roughly 10% of the training data is set aside for cross-validation during MLP training to prevent the parameter-heavier MLPs from over-fitting. In the training of the GMM classifiers, where hardly any over-fitting can occur due to smaller number of parameters, the cross-validation set is included in the training set. Our test set consists of 200 utterances taken from the larger development test set of 1206 utterances (Glotin, 2000). The large number of subband and subband-combination recognizers and the need to test on several different noise conditions led to the decision to choose a smaller test set. The vocabulary of our training and test sets comprises 30 connected digit words, such as “zero”, “eight”, “fifteen”.

Finally, it needs to be mentioned that our NUMBERS95 utterances are 1600 samples longer than the utterances of the OGI[3] release. This is because 100 ms of silence were artificially added

[2] Many thanks to Stéphane Dupont for providing the segmentation.

[3] Oregon Graduate Institute (OGI) School of Science and Engineering (`http://cslu.cse.ogi.edu`).

at the beginning and end of each utterance in order to avoid problems with the contextual input of the MLPs and with the time-constant of the RASTA filter (cf. (Hermansky and Morgan, 1994, p. 579)).

8.2.2 Noisy test data

To create the noise-corrupted test data, a range of noise conditions was added to the NUMBERS95 test set at different SNR levels. These noise cases comprise real-environmental wide-band noise conditions, such as car and factory noise, as well as artificial, stationary and non-stationary, narrow-band noise, which are described in more detail below.

Adding the noise artificially to the test data allows us to use the same speech data throughout the recognition experiments and judge the effects of each noise on the recognition task without having to consider differences that would stem from different speech corpora or different recordings. Scaled samples of the recorded or artificially created noises $N(f)$ are thus added sentence-by-sentence to the test data $S(f)$:

$$\begin{aligned} |Y(f)|^2 &= |S(f)+N(f)|^2 & (8.1) \\ &= |S(f)|^2 + |N(f)|^2 + S(f)N^*(f) + S(f)^*N(f) & (8.2) \end{aligned}$$

with $S(f)^*$ and $N^*(f)$ the complex conjugates of the Fourier transform. Assuming that the speech signal and the noise signal are independent and uncorrelated, the last two terms in (8.2) can be supposed to be zero. The relative scaling between speech and noise is specified to a desired SNR (globally for each sentence, silences excluded), adapting the gain factor g so that the desired SNR is obtained according to

$$\mathrm{SNR} = 10 \cdot \log_{10} \frac{\sum_f S^2(f)}{\sum_f gN^2(f)} \tag{8.3}$$

Artificial band-limited noise Artificial *stationary* narrow-band noise is created from Gaussian white noise which is passed through a set of band-pass filters, which are produced from two first order Butterworth filter sections (one high-pass and one low-pass). The noise bandwidth of 300 Hz is kept constant for each frequency subband[4]. The noise is then added at SNR levels of 0 and 12 dB to the middle frequencies of each of the four bands[5]. The narrow-band noise case for subband two can be seen in the upper panel of **Figure 8.1**. The clean spectrum of the same sentence is given in the lower panel. For a more detailed description see (Glotin, 2000; Hagen and Glotin, 2000).

Non-stationary narrow-band noise (also refered to as artificial siren noise) is created using segments of 100 ms taken from the four stationary band-limited noise cases. These segments were concatenated according to the following order of filters: 1, 2, 3, 4, 4, 3, 2, 1. The resulting noise is shown in the middle panel of **Figure 8.1**.

[4] However, this constant noise bandwidth means that leakage is greater between low frequency subbands than between high frequency subbands.

[5] Many thanks to Hervé Glotin for providing these noise cases.

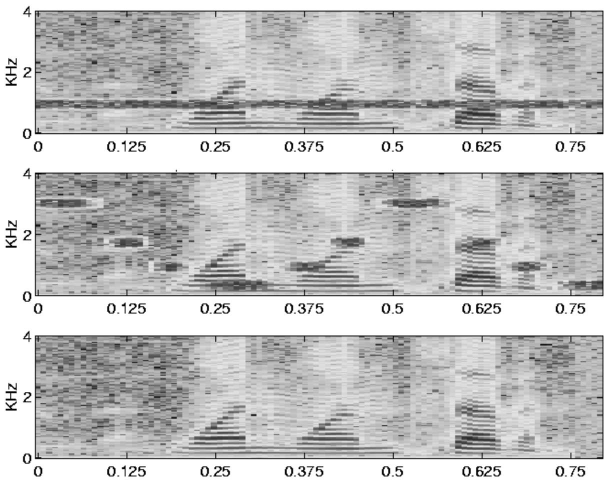

Figure 8.1: Illustration of a clean speech spectrum (lower panel) and of the same spectrum corrupted by stationary band-limited noise (upper panel) and non-stationary band-limited noise (middle panel). The digits spoken in this utterance are "one one seven".

	TITLE	DESCRIPTION
	clean	Uncorrupted speech
wide band	car	Added car noise from Daimler Chrysler
	factory	Added factory noise from the NOISEX92 database (noise 21)
band lim- ited noise	band 1	Added artificial band-limited noise in subband 1: 221.9-521.9 Hz
	band 2	Added artificial band-limited noise in subband 2: 817.6-1117.6 Hz
	band 3	Added artificial band-limited noise in subband 3: 1627.2-1927.2 Hz
	band 4	Added artificial band-limited noise in subband 4: 2795.3-3095.3 Hz
	siren	Concatenation of noise cases 'band 1', 'band 2', 'band 3', and 'band 4'

Table 8.2: Description of clean and noise conditions for our test set originating from the NUMBERS95 test database. Each noise is added at two different SNR levels: 12 and 0 dB. For further descriptions see text.

Real-environmental wide-band noise To evaluate our systems on more realistic, wide-band noise cases, two instances of real-environmental noise were added to the clean test set at the same SNR values as described for the band-limited noise in the last paragraph (i.e. SNR of 12 and 0 dB). The first noise case is factory noise (noise 21) taken from the NOISEX92 database (Varga et al., 1992). The second noise case is an in-house recorded car noise provided by our project partner Daimler Chrysler[6].

The different noise types are summarized in **Table 8.2**.

8.2.3 Evaluation by measure of word error rate

To evaluate the performance of a speech recognizer, the orthographic transcription, that is the words, which are obtained at the output of the recognizer are compared to the known transcriptions of the test database. With this, we receive the rate of wrongly inserted I, deleted D or substituted S words, and can evaluate the *word error rate (WER)* of the recognizer:

$$\mathrm{WER} = \frac{I + D + S}{N} \cdot 100 \tag{8.4}$$

with N the number of words in the respective test set. In our test set of 200 utterances the number of words is $N = 800$.

Significance Test To decide whether the word error rate estimate WER_i for test i is significantly greater than the best WER obtained ($\mathrm{WER}_{\mathrm{best}}$), we apply a test from (Mokbel, 1992). In this test, it is assumed that the total number of words N is large, that the true WER is approximately equal to the estimated WER_i, and that $\frac{\mathrm{WER}}{100}$ can be modeled as a fixed probability of error during N Bernoulli trials. In this case, the expected variance in the estimated WER_i is $\epsilon_i^2 = \mathrm{WER}_i \frac{100-\mathrm{WER}_i}{N}$, so that we can be 97.5% confident that true $\mathrm{WER}_i > \mathrm{WER}_i - 1.96\,\epsilon_i$. As ϵ^2 decreases as $\mathrm{WER} \to 0$, ϵ for the best (smallest) WER will be smallest. Therefore, for each WER_i we obtain ϵ_i^2 and decide that it is not significantly worse than $\mathrm{WER}_{\mathrm{best}}$ if $\mathrm{WER}_i - 1.96\epsilon_i < \mathrm{WER}_{\mathrm{best}}$.

8.3 Experimental evaluation of combination strategies

In this section, multi-band probability combination which was illustrated in **Figures 5.4** and **5.5** is evaluated.

In the framework of HMM/ANN hybrid systems, the newly proposed combination strategies, including all FC strategies, the product of errors rule (STD PoE and FC PoE), and the FC-ECPC scheme are compared to some of the most widely used recombination approaches which are usually found in the literature. These are the standard sum rule (STD SUM), the product rule (STD PRODUCT), and the independence assumption rule (STD INDEP ASMPT)[7]. We use the term "standard" (STD) to refer to all subband systems which employ the four single spectral subbands only and no combinations of subbands.

[6] Many thanks to Udo Haiber from Daimler Chrysler for supplying us with this noise, and to Christopher Kermorvant who added it to our test data.

[7] Their mathematical equations are, amongst others, summarized in the tables in Appendix D.

Each combination strategy presented in this section merely employs *equal weights*, which correspond to the fraction of one divided by the number of subband experts, *in order to evaluate the performance of all systems under as similar conditions as possible*. The different weighting strategies as described in Sections 7.3 and 7.4 are then evaluated for the best of the new systems in subsequent sections. As our goal is to establish an automatic speech recognition system which provides high performance in clean speech and degrades as little as possible in the case of unseen noise, each system is first tested on clean speech ("matched condition"). The next set of experiments is carried out on the noise conditions for which multi-band processing is known to be advantageous: band-limited noise restricted to one spectral subband. For this, we use the artificially created stationary and non-stationary band-limited noise cases as described above. We then turn to the more realistic noise conditions: wide-band car and factory noise.

8.3.1 Baseline systems

The baseline systems for all our experiments in this chapter constitute the fullband HMM/-MLP hybrid recognizers which were trained on the clean and entire frequency domain. For the Numbers95 database, this means the frequencies ranging from 115.3 Hz to 3768.8 Hz (due to the definition of the critical band filters). Lower and higher frequencies are disregarded as they are not present in telephone speech and thus only contain channel noise. The same features are employed as in the respective multi-band systems which are to be evaluated: PLP and J-RASTA-PLP features. Each (static) feature vector is appended with the first (delta) and second (delta-delta) order difference features. In the case of FC processing, when all combinations of subbands are used, the fullband recognizer is automatically part of the FC multi-band system.

8.3.2 FC and AFC experiments on clean speech and on speech with narrow-band noise

	PLP	J-RASTA-PLP
Fullband	7.1	7.8
STD Sum	14.8	17.9
STD Indep Asmpt	13.0	14.4
STD Product	12.9	11.2
STD PoE	17.1	21.8
FC Sum	7.4*	9.0*
AFC Sum	10.8	11.9

Table 8.3: WERs of the baseline fullband recognizers, the standard multi-band combination strategies, and FC Sum and AFC Sum in clean speech, employing PLP and J-RASTA-PLP features. * indicates that there is no significant difference to the best result in this column.

To recall, the corresponding equation for each posterior-based combination strategy is as follows: STD Sum is defined in (5.8), STD Indep Asmpt is defined in (5.12), STD Product is defined in (5.11), STD PoE in **Table D.1** in Appendix D, and FC Sum is defined in (6.3). For AFC Sum, the approximation of each *combination* probability is calculated according to (6.28) and normalization in (6.29). The approximated combination probabilities are then used together with the single-stream probabilities in (6.3).

Clean speech

The PLP features were chosen for their good performance in clean speech. The J-RASTA-PLP features, on the other hand, were selected due to their known robustness to additive and convolutive noise. The RASTA filtering, though, which would not be necessary under matched testing conditions leads to a slight degradation in clean speech as compared to the PLP features. This was also found in (Hermansky and Morgan, 1994, p. 580). As can be seen in **Table 8.3**, almost every system employing PLP features outperforms its counterpart using J-RASTA-PLP features when tested under matched conditions (with the exception of the STD Product).

Looking at the results for each feature set respectively, the performance loss in clean speech due to standard subband processing as compared to fullband processing is apparent. The standard combination strategies which only employ the four spectral bands miss correlation information between subbands which renders them inferior in clean speech. In FC subband processing, correlation information within each combination of subbands is explicitly modeled and provides competitive performance in clean speech. The difference between FC Sum and Fullband results is insignificant. AFC Sum , which approximates the posterior probabilities of the subband combinations, ranges between the two extremes of standard subband processing and FC.

PLP features in band-limited noise

The experiments on stationary band-limited noise in each of the frequency bands are presented next. Results for systems employing PLP features can be seen in **Table 8.4**. The last column indicates the mean value as taken over all 8 noise conditions in the table which eases performance comparison over the large number of conditions. As can be seen in the table the fullband baseline system has highest word error rate as compared to each of the multi-band systems in stationary band-limited noise. The standard combination strategies, STD Sum, STD Indep Asmpt and STD Product, are more robust to this kind of noise, with the STD Sum dominating. The STD PoE rule achieves results not significantly different from the STD Sum. In standard multi-band processing, the experts of the uncorrupted bands work under matched conditions and recombination results in higher performance than the fullband expert, even though one of the subband experts is unreliable and has high entropy. Better performance can still be achieved with some full combination rules, with AFC Sum outperforming FC Sum.

Very similar behavior can be observed on *non*-stationary band-limited noise (siren). The results for PLP features are presented in **Table 8.5**. Again, with corrupted feature components all over, the fullband recognizer degrades severely. The standard subband combination strategies achieve lower word error rates also in this non-stationary noise case when the band-limited noise alternates from one spectral subband to the next. Among the standard techniques, the

	Stationary Band-Limited Noise								
	Band 1		Band 2		Band 3		Band 4		Mean
	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	
FULLBAND	57.5	29.2	74.6	34.1	65.4	31.2	67.2	32.5	49.0
STD SUM	33.1	26.9	51.4	28.9	29.6	23.9	22.8	19.5	29.5
STD INDEP ASMPT	42.9	27.8	69.0	34.5	47.2	32.0	29.8	25.2	38.6
STD PRODUCT	43.1	27.6	66.2	34.2	47.6	31.1	29.6	25.1	38.1
STD POE	33.9	26.8	52.8	28.2	40.4	22.5	25.2	20.5	31.3
FC SUM	36.6	20.2	46.1	26.2	28.8	17.2	21.0	16.8	26.6
AFC SUM	31.2	20.6	27.4	17.1	22.9	17.0	17.9	15.8	21.2

Table 8.4: WERs of baseline fullband recognizer, standard subband combination strategies, FC and AFC in stationary, band-limited noise, employing PLP features.

STD PoE rule and STD SUM behave by far the most robust. The FC SUM does not lead to a further decrease in word error rate as compared to the STD SUM, but AFC SUM again achieves additional gain in robustness.

	Siren		
	0 dB	12 dB	Mean
FULLBAND	66.9	36.1	51.5
STD SUM	30.8	23.6	27.2
STD INDEP ASMPT	44.9	28.8	36.9
STD PRODUCT	44.0	28.1	36.1
STD POE	30.9	21.8	26.4
FC SUM	34.9	19.9	27.4
AFC	24.9	16.4	20.7

Table 8.5: WERs of baseline fullband recognizer, standard subband combination strategies, FC and AFC in non-stationary band-limited noise, employing PLP features.

Discussion In a multi-band recognizer when the noise is limited to only one frequency subband, the recognizers of all other subbands are (almost) unaffected. One would expect that only the subband recognizer of the noise corrupted frequency band will endure increased entropy which, in the worst case, results in equal posterior probabilities for all classes. In the recombination procedure, this corresponds to an addition (STD SUM) or multiplication (STD INDEP ASMPT, STD PRODUCT, STD POE) by a (almost) constant value for each class and, thus, does not affect the recognition task. In AFC, the corrupted bands will tend to have smaller probability estimates, and will therefore down-weight the combinations in which they are included, rendering the clean combinations dominant in the final recombining sum.

In the case of FC processing, the situation is two-fold. On the one hand, although there is noise only in one frequency subband, there will be several (in our case six) combination experts affected. On the other hand, there are still more combination experts which are free from noise, and, moreover, one of which models exactly the clean frequency domain. If we again assume the worst case, the corrupted experts should output equal probabilities, thus, the decision is dominated by the clean experts.

J-RASTA features in band-limited noise

As seen above in the case of PLP features, subband systems are significantly more robust than fullband systems in the case of stationary, as well as non-stationary, narrow-band noise. However, as we will see below, this advantage is lost in case of wide-band noise. On the other hand, (J-)RASTA-PLP is known to be particularly robust to slowly varying wide-band noise. We thus test our multi-band systems on J-RASTA-PLP parameters, where J-RASTA-PLP will remove the wide-band noise, followed by subband processing to address narrow-band noise. We start by testing the J-RASTA-PLP features in (stationary and non-stationary) band-limited noise.

	Stationary Band-Limited Noise								
	Band 1		Band 2		Band 3		Band 4		Mean
	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	
FULLBAND	30.6	11.4	48.0	16.0	35.2	18.4	24.5	19.2	25.4
STD SUM	35.0	24.1	38.8	25.5	29.2	25.4	24.5	23.6	28.3
STD INDEP ASMPT	34.6	20.6	41.1	21.8	33.1	23.1	25.5	23.0	27.9
STD PRODUCT	26.6	13.1	35.9	16.9	20.6	14.2	16.4	16.2	20.0*
STD POE	35.6	26.5	41.2	28.0	29.4	26.0	24.1	23.1	29.2
FC SUM	19.8	9.9	30.2	16.9	20.1	14.0	15.9	15.0	17.7
AFC SUM	26.6	13.5	33.6	17.6	22.9	16.9	18.0	17.1	20.8

Table 8.6: WERs of baseline fullband recognizer, standard subband combination strategies, FC and AFC in stationary band-limited noise, employing J-RASTA-PLP features. * indicates that there is no significant difference to the best result in this column.

The results on stationary narrow-band noise (see **Table 8.6**) show the increased noise robustness due to J-RASTA filtering for all systems (as compared to the results on PLP features in **Table 8.4**). Word error rate of the FULLBAND system is almost halved. As is well known, better systems are harder to improve and we have to observe that standard multi-band processing employing J-RASTA-PLP features results in higher noise robustness only for one combination scheme, the STD PRODUCT. The STD SUM and STD INDEP ASMPT cannot improve over the FULLBAND system for this kind of noise (though their results are not significantly worse as compared to the FULLBAND). On the other hand, both full combination strategies (FC SUM and AFC SUM) achieve significantly higher noise robustness than the FULLBAND system.

In fast changing non-stationary band-limited noise, all systems using J-RASTA-PLP features (cf. **Table 8.7**) degrade more than in stationary band-limited noise, showing that the

Table 8.7: WERs of baseline fullband recognizer, standard subband combination strategies, FC and AFC in non-stationary band-limited noise, employing J-RASTA-PLP features. * indicates that there is no significant difference to the best result in this column.

	Siren		
	0 dB	12 dB	Mean
FULLBAND	104.6	48.1	76.4
STD SUM	40.2	27.0	33.6
STD INDEP ASMPT	53.8	31.0	42.4
STD PRODUCT	48.4	22.1	35.3
STD PoE	39.1	26.5	32.8
FC SUM	40.0	19.9	30.0
AFC SUM	41.1	22.8	32.0*

J-RASTA-PLP features are unable to handle these fast changes. However, some of the additional degradation in this kind of noise as compared to the stationary noise case is due to the fact that by adding the noise to one subband after the other, more frames are actually corrupted than in the stationary noise case (although the same filters have been used), as described in (Hagen and Glotin, 2000).

In non-stationary band-limited noise all multi-band systems using J-RASTA-PLP features degrade significantly less than the J-RASTA-PLP-based FULLBAND, as some of the experts remain clean and reliable. Again, the FC SUM rule outperforms the standard multi-band combination rules, and the AFC SUM rule is not significantly worse than FC SUM.

As compared to the PLP features which are less disturbed by the changing noise condition, almost all systems result in higher word error rate using J-RASTA-PLP features in non-stationary band-limited noise. The only exception is the STD PRODUCT, though the difference is not significant.

Preliminary conclusions

From the above results we can so far conclude that:

- for clean speech, we need the FULLBAND recognizer or FC SUM (no significant difference between feature sets),
- for stationary band-limited noise, we need a multi-band system (FC SUM, STD PRODUCT or AFC SUM) employing J-RASTA-PLP features,
- for non-stationary band-limited noise, we need a multi-band system using PLP features (best would be AFC SUM, then STD PoE, STD SUM and FC SUM).

8.3.3 FC and AFC experiments on speech with real-environmental noise

As we have already experienced in the last paragraphs, the respective noise conditions play a significant role in the evaluation of each system. Let us hence turn to more realistic noise cases such as the car and factory noises described above. Both noises stem from real recordings and were artificially added to the clean test data.

	Car		Factory		Mean
	0 dB	12 dB	0 dB	12 dB	
FULLBAND	29.1	9.8	34.1	12.5	21.4
STD SUM	61.2	26.6	60.5	25.4	43.4
STD INDEP ASMPT	56.6	21.0	57.8	22.2	39.4
STD PRODUCT	42.5	13.8	45.0	15.1	29.1
STD POE	62.4	28.0	58.5	28.1	44.3
FC SUM	29.5	10.8	35.1	12.5	22.0*
AFC SUM	48.0	17.4	46.5	17.4	32.3

Table 8.8: WERs of baseline fullband recognizer, standard subband combination strategies, FC and AFC in wide-band (car and factory) noise, employing J-RASTA-PLP features. * indicates that there is no significant difference to the best result in this column.

In these experiments, the results stemming from J-RASTA-PLP and PLP features show similar behavior, we thus only discuss the results from one feature set and choose the J-RASTA-PLP features having overall lower error rate on wide-band noise. The results for the PLP features can be found in Appendix G.

On wide-band noise (cf. **Table 8.8**), the FULLBAND and FC SUM perform best with an insignificant difference between the two. They are followed by the STD PRODUCT and AFC SUM with no significant difference between these two. The STD SUM, STD INDEP ASMPT and STD POE combinations deteriorate significantly more on wide-band noise. It can be seen that the good performance of standard multi-band processing on frequency-selective noise is no longer warranted under these noise conditions. It is more difficult for a standard multi-band system to reach robust performance when several of the subband recognizers are affected by noise and each of them only works on one subband. In FC processing, on the other hand, all possible combinations of subbands are considered, which makes it more likely to also capture the most uncorrupted and reliable part of the data by one of the experts.

Preliminary conclusions

To conclude the first set of experiments, we can state that the FC SUM (employing J-RASTA-PLP features) is the only multi-band approach which has so far been able to provide highest noise robustness in *all* noise conditions tested, and, at the same time, being insignificantly different from the best system in clean.

Moreover, in full combination processing we have the possibility to further increase performance by the use of (non-equal) weights, which is not possible for the fullband recognizer. We will see in Section 8.4 below how the performance of the FC system can be further enhanced by the use of appropriate weighting strategies.

8.3.4 Experiments with FC Product, FC Indep Asmpt and FC PoE combination strategies

We now turn to the next set of combination strategies which also employ all possible combinations of subbands but recombine them in different ways: the FC Product of (6.23), the FC Indep Asmpt and FC PoE as defined in **Table D.2** of Appendix D. From now on only the results from J-RASTA-PLP-based systems will be discussed as the J-RASTA-PLP features provide in general higher noise robustness than the PLP features (except for non-stationary band-limited noise). The corresponding results using PLP features are summarized in Appendix G. The results in clean speech and the different noise conditions are presented in **Table 8.9**. For stationary band-limited noise, only the mean of the word error rates over all eight band-limited noise cases are given. For non-stationary band-limited noise the mean is calculated over the two SNR values.

	Clean	Band-Limited Noise		Wide-Band Noise				
		Stationary	Non-Stat.	Car		Factory		Mean
		Mean	Mean	0 dB	12 dB	0 dB	12 dB	
FC Indep Asmpt	8.0*	18.6◇	39.3	27.4	10.0	32.4	10.1	20.0°
FC Product	7.9*	19.1◇	41.8	25.4	9.1	31.9	10.8	19.3
FC PoE	15.9	19.4◇	31.8⋆	40.5	11.5	42.0	12.4	26.6

Table 8.9: WERs of FC Indep Asmpt, FC Product and FC PoE in clean and noise (band-limited and wide-band noise), employing J-RASTA-PLP features. * indicates that there is no significant difference as compared to Fullband in clean (7.8%), ◇ as compared to FC Sum in stationary noise (17.7%), ⋆ as compared to FC Sum in siren noise (30.0%), and ° as compared to best result in this column.

Recognition performance of FC Indep Asmpt and FC Product on clean speech does not differ significantly from the results achieved by the Fullband recognizer and the FC Sum (7.8% and 9.0% respectively, cf. **Table 8.3**). The FC PoE resulted in an almost double WER as it was also roughly the case for the STD PoE rule in clean speech (as compared to STD Product).

In stationary band-limited noise, there is, for all three systems, no significant difference to the best result, which was achieved by the FC Sum (17.7%, cf. **Table 8.6**).

In non-stationary band-limited noise, the FC Indep Asmpt and FC Product deteriorate more than most of the multi-band systems (cf. **Table 8.7**) though staying more robust than the Fullband recognizer (which has a WER of 76.4% on this kind of noise). The FC PoE rule outperforms both FC Indep Asmpt and FC Product achieving results almost as good as the best system (FC Sum with 30.0%).

In wide-band noise the FC PoE approach outperforms all standard multi-band combination strategies and AFC Sum, but it is not competitive to FC Sum or the Fullband recognizer. The FC Indep Asmpt and FC Product approaches gain the highest robustness in this kind of noise, but are not significantly better than the Fullband and FC Sum (21.4% and 22.0% respectively, cf. **Table 8.8**).

As can be seen in Appendix G, the results employing PLP features are similar, with the FC PoE rule being competitive even in clean speech. FC Indep Asmpt and FC Product using PLP features deteriorate more on stationary band-limited noise than it was the case when using J-RASTA-PLP features.

8.3.5 Experiments with FC-ECPC

In Section 6.5, we described how a recently introduced model for quantifying the influence of contextual information on human recognition performance (Bronkhorst et al., 1993) could be interpreted in multi-band based ASR. In Bronkhorst et al. (1993)'s model, the recognition probability of a word is derived from the probabilities of correctly or incorrectly recognizing each of its phonemes, multiplied by a correcting term to account for the mis-recognized phonemes. It is then summed over all possible combinations of correctly and incorrectly recognized phonemes. In multi-band processing, we derive the recognition probability of the whole frequency band in the same way from all possible combinations of correctly and incorrectly recognized subbands. We refered to this approach as Error Correction in Posteriors Combination (ECPC), as the error probabilities of the mis-recognized subbands are accounted for through the application of appropriate weights. Due to its similarity to FC, the ECPC approach was combined with our FC approach resulting in the FC-ECPC formula (6.34).

	Stationary Band-Limited Noise								Mean
	Band 1		Band 2		Band 3		Band 4		
	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	
Fullband	31.4	14.0	44.6	16.6	35.0	18.9	23.9	17.4	25.2
FC Sum	25.6	12.5	23.2	13.6	21.8	15.1	15.8	14.8	17.8*
FC-ECPC	24.1	11.9	22.6	13.9	21.5	14.5	14.8	13.6	17.1
FC-ECPC wght	32.0	13.6	39.1	15.0	37.4	19.1	22.4	18.5	24.6

Table 8.10: WERs of FC processing without (FC Sum) and with error correction (FC-ECPC), employing J-RASTA-PLP features. 'wght' refers to the proposed weights from Section 6.5. * indicates that there is no significant difference to the best result in this column.

In this section, the experiments employing FC-ECPC subband combination are presented (also cf. (Hagen and Bourlard, 2001)). The results for FC (without ECPC), which are given here for comparison to FC-ECPC, differ slightly from the FC ones presented in the last section (FC Sum) as we do not further convert the posterior probabilities to scaled likelihoods. In FC-ECPC, we work strictly with posterior probabilities. The differences in WER are insignificant for almost all conditions, with the exception of FC Sum in wide-band noise which deteriorates significantly more (from 22.0% down to 25.8%) when no division by priors is applied. For comparison, the Fullband results are also given without division by priors.

FC combined with the ECPC approach is refered to as FC-ECPC. This approach is first investigated with equal weights, to make direct comparison to FC (FC Sum) more apparent.

	Clean	Band-Limited Noise			Wide-Band Noise				
		Non-Stationary			Car		Factory		Mean
		0 dB	12 dB	Mean	0 dB	12 dB	0 dB	12 dB	
FULLBAND	8.0	89.4	36.9	63.2	32.8	10.6	34.6	11.4	22.4*
FC SUM	8.6*	39.5	19.4	29.5*	39.9	10.6	40.9	11.8	25.8
FC-ECPC	8.4*	37.9	18.1	28.0*	39.0	10.8	41.1	11.9	25.7
FC-ECPC WGHT	8.6*	46.0	33.8	39.9	31.2	11.1	33.8	12.0	22.0

Table 8.11: WERs of FC processing without (FC SUM) and with error correction (FC-ECPC), employing J-RASTA-PLP features. 'WGHT' refers to the proposed weights from Section 6.5. * indicates that there is no significant difference to the best result in this column.

Additionally, we evaluate FC-ECPC using the weights as derived in Section 6.5 based on the human model proposed by (Bronkhorst et al., 1993).

Results in clean speech are presented in the first column of **Table 8.11**. No significant performance difference between FC SUM, FC-ECPC (with and without weighting) can be observed.

In band-limited noise, both stationary (cf. **Table 8.10**) and non-stationary (cf. **Table 8.11**), results again stay almost the same with and without ECPC using equal weights, but deteriorate when the proposed weights are used. On the contrary, when applying FC-ECPC in wide-band car and factory noise, the proposed weights obtain a significant improvement in WER as compared to both FC SUM and FC-ECPC with equal weights. The weights which were employed in these tests are simply the prior probabilities of each phoneme class. If other weights could be found to more accurately reflect the error correction procedure of the FC-ECPC approach, the good results on wide-band noise could probably further be improved.

8.4 Experimental evaluation of weighting schemes

In this section, the different weighting strategies are evaluated in the multi-band system which so far achieved the best results, which is the FC SUM. Performance is compared to the baseline fullband recognizer.

8.4.1 Fixed weights in HMM/MLP hybrid systems

The fixed weights which are developed in this thesis are (i) the relative frequency (RF) estimates from (7.12), and (ii) the weights derived from LMSE estimation which are defined in (7.11). Both sets of weights are estimated on the clean training data, employing the same experts as used during recognition. Results in clean speech and the different noise conditions are presented in **Table 8.12**. The features which are used here are the J-RASTA-PLP features. Only mean values are given for the band-limited noise cases. Results for the PLP features can be found in Appendix G.

In clean speech, slight improvement in WER is achieved by both weighting schemes as compared to the FULLBAND recognizer and FC SUM using equal weights which, though, is not statistically significant. In stationary band-limited noise, the FC SUM rule using equal weights has been seen to outperform all other combination strategies. The same is true for FC SUM with RF and LMSE weighting but without further gain in robustness. The more difficult task of recognition in non-stationary band-limited noise reveals the performance advantage due to improved weights: although the FC SUM using equal weights had already proved highest robustness to this kind of noise, both the RF and the LMSE weights obtain further, significant improvement in robustness as compared to the use of equal weights. When testing on real-environmental wide-band noise, only small improvement due to RF and LMSE weights is achieved; the difference though is not significant as compared to the results obtained by the FULLBAND, FC SUM using equal weights, and FC PRODUCT which had lowest WER on this kind of noise (19.3% in **Table 8.9**).

	Clean	Band-Limited Noise		Wide-Band Noise				
		Stationary	Non-Stat.	Car		Factory		Mean
		Mean	Mean	0 dB	12 dB	0 dB	12 dB	
FULLBAND	7.8*	25.4	76.4	29.1	9.8	34.1	12.5	21.4*
FC SUM EQUAL	9.0*	17.7	30.0	29.5	10.8	35.1	12.5	22.0*
FC SUM RF	7.5*	18.3*	22.6	26.9	9.9	31.4	11.0	19.8
FC SUM LMSE	7.4	19.7*	25.0*	27.0	10.4	32.9	11.1	20.4*

Table 8.12: WERs of the FC SUM rule employing different weighting strategies and the FULLBAND recognizer, on J-RASTA-PLP features. * indicates that there is no significant difference to the best result in this column.

To conclude we can state that the use of RF and LMSE weights in the FC SUM employing J-RASTA-PLP features, consistently led to improved recognition performance on clean speech, siren and wide-band noise, though only the results obtained in non-stationary band-limited (siren) noise using RF weights were *significantly* better.

In the case of PLP features, RF weights employed in the FC SUM did neither significantly in- nor decrease performance as compared to using equal weights in all conditions, whereas LMSE weighting worsened performance in band-limited noise.

A comparison between the RF weights as illustrated in **Figure 8.2** and the LMSE weights as illustrated in **Figure 8.3** show that the RF weights better match the best streams for each phoneme in clean speech than the LMSE weights do. In the figures, the best three streams (in clean data) for each phoneme are indicated in the upper panel; in the middle panel, the three highest weight values are given for each phoneme and stream as calculated by RF or LMSE estimation, respectively. Finally, in the lowest panel, these two matrices are compared, and the difference is shown by white (too high weight values) and black (too low weight values) squares. The difference between the matrices is smaller for the RF weights than for the LMSE weights.

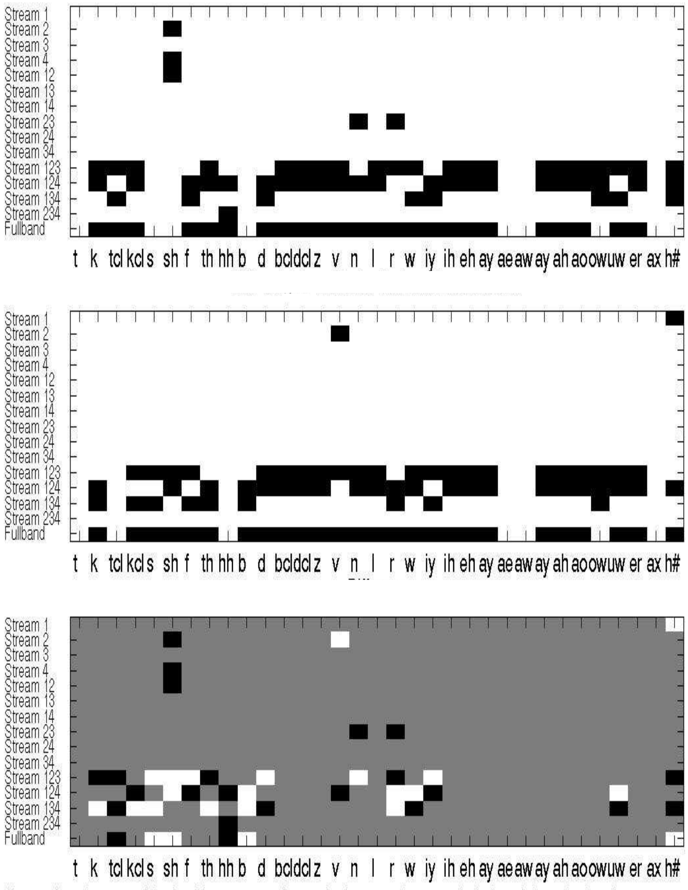

Figure 8.2: Evaluation of RF weights calculated on clean speech. The matrix of the best three streams for each phoneme is given in the upper panel, of the RF weights for the best three streams is given in the middle panel, and of the difference between the upper two is illustrated in the last panel. In the lowest panel, the white squares indicate too high weight values whereas black squares indicate too low weight values for the respective stream and phoneme.

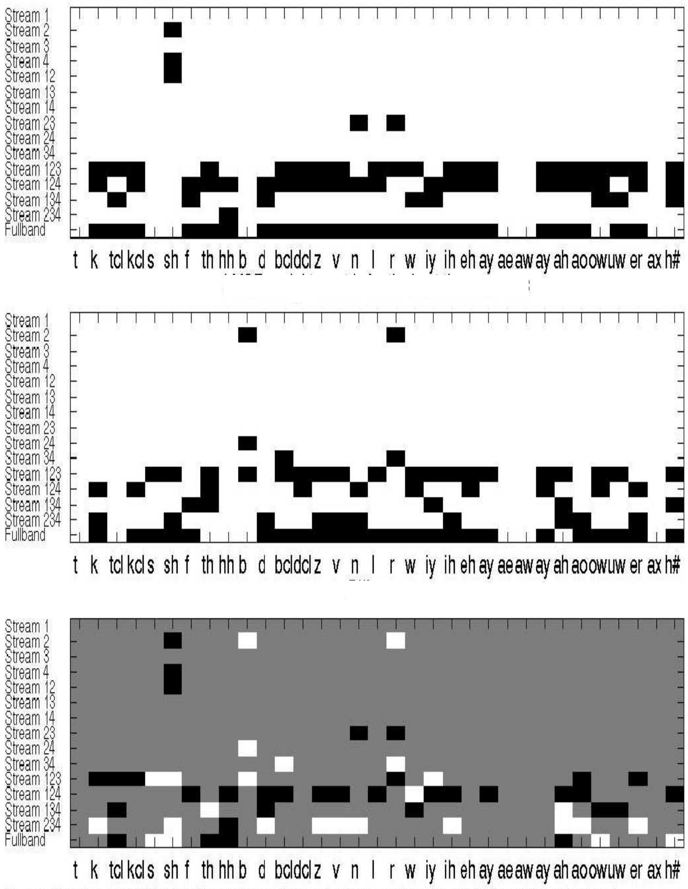

Figure 8.3: Evaluation of LMSE weights calculated on clean speech. The matrix of the best three streams for each phoneme is given in the upper panel, of the LMSE weights for the best three streams is given in the middle panel, and of the difference between the upper two is illustrated in the last panel. In the lowest panel, the white squares indicate too high weight values whereas black squares indicate too low weight values for the respective stream and phoneme.

8.4.2 Fixed weights in HMM-GMM systems

In this section, the multi-band systems employing *GMMs* to estimate the HMM emission probabilities are tested using the (fixed) maximum likelihood based (ML) weights (7.18) (together with (7.19) and (7.21)) which were introduced in Section 7.3.4. For the calculation of these weights, the parameters of all stream HMM-GMMs were fixed and only the combination weights were estimated using the EM algorithm. To recall, the corresponding recombination strategy using the (likelihood-based) FC Sum rule 2 is given in (6.9) together with (6.10), and for the STD Sum rule 2 it is presented in **Table D.1** of Appendix D.

GMM classifiers were trained for each frequency subband and combination of subbands as well as the fullband, on MFCC features. The ML weights are employed in both standard multi-band processing and FC processing. The results are compared to the same setup using equal weights and quasi-optimal weights as well as to the fullband recognizer. The quasi-optimal weights Opt in this case are zero for the noisy subband and equal for the clean bands.

For estimation of the offline ML weights, the test data was split into two sub-sets, the first one of which was used to calculate the weights, the second to carry out the tests.

In multi-band systems using either the single subbands only or the FC approach, we would expect the ML weights to show a clear advantage over equal weights when one of the bands is totally corrupted by noise. For this reason, experiments and a visual evaluation of the weights are first carried out on the stationary band-limited noise cases.

Visual analysis of the weights

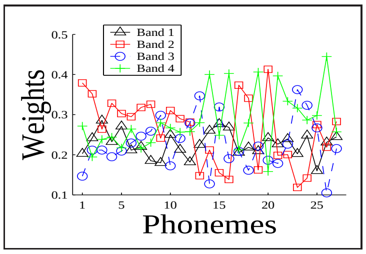

Figure 8.4: Illustration of offline adapted, fixed ML weights of (7.18) for clean speech.

The ML weights for the standard multi-band system are illustrated in **Figure 8.4** for clean data, and in **Figures 8.5** and **8.6** for the four stationary band-limited noise cases. The weights for the FC system are only given for stationary band-limited noise in subband 3 (see **Figure 8.7**). For clean speech the weights depend on both the subband and the respective phoneme and thus change from phoneme to phoneme. Investigating these weights in more detail to see whether for example fricatives received more weight in the higher subbands did not lead to any acoustic-phonetically consistent conclusions.

For noise-corrupted speech however, it can be seen how the noisy band gets consistently down-weighted as compared to the clean subbands, in the case of a standard multi-band system (**Figures 8.5** and **8.6**). In the FC system not only the respective subband but also all combinations which contain this subband are corrupted by noise. It can be seen in **Figure 8.7**, how also in this case the noisy combinations are well detected and down-weighted. However, we expected the larger subband combinations to obtain higher weight values than the single-stream experts. Unfortunately, rather the contrary was the case. In order to give equal weights to all clean combinations we introduced a threshold function which sets weights below the threshold to 0 and to 1 otherwise. Results are discussed below.

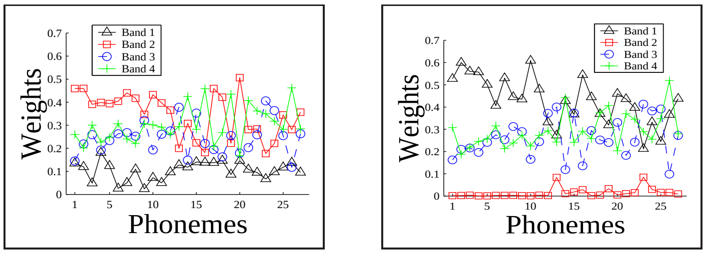

Figure 8.5: Illustration of fixed ML weights for noise in **subband 1** (left) and subband 2 (right).

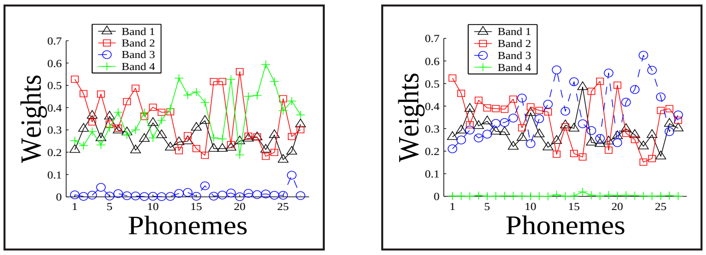

Figure 8.6: Illustration of fixed ML weights for noise in subband 3 (left) and subband 4 (right).

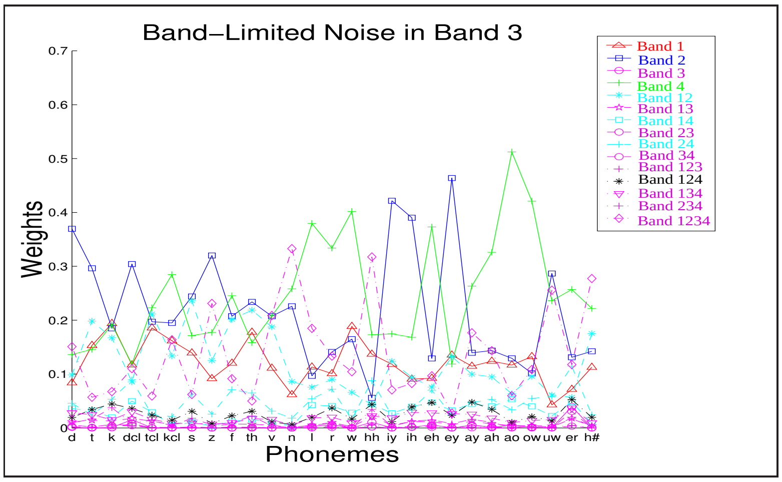

Figure 8.7: Illustration of offline adapted, fixed ML weights for the FC system for band-limited noise in subband 3. Noise corrupted combinations are illustrated in pink.

Experimental evaluation of the weights

	Clean	Stationary Band-Limited Noise				
		Band 1	Band 2	Band 3	Band 4	Mean
FULLBAND	13.5	85.7	63.4	51.6	44.2	61.2
STD SUM EQUAL	24.3	44.0	27.3	35.4	31.9	34.7
STD SUM ML	26.0	38.3	27.3	30.2	28.7	31.1
STD SUM OPT	-	35.1	26.5	26.8	27.8	29.1
FC SUM EQUAL	14.7	47.7	44.2	36.4	36.6	41.2
FC SUM ML	18.4	27.8	29.5	25.6	25.8	27.2
FC SUM OPT	-	23.6	27.0	25.1	19.4	23.8

Table 8.13: WERs of the baseline fullband HMM-GMM as compared to standard multi-band and FC multi-band processing using HMM-GMMs on MFCC features. The multi-band systems are tested with equal weights, ML weights and quasi-optimal OPT weights. Results are given for stationary band-limited noise at SNR=0 dB.

The results for the FULLBAND, the standard multi-band system as well as the FC multi-band system are given in **Table 8.13**. In clean speech the FULLBAND system results in lowest WER[8] though the difference to the FC SUM (EQUAL weights) is not significant. The standard multi-band system is not competitive in clean condition. Employing the ML weights in clean speech did not result in any performance gain neither for the standard multi-band system nor for the FC approach.

While ML weights in **Figures 8.5** to **8.6** for one-band experts clearly identify the noisy subband, in **Figure 8.7** weights for the largest clean combination are disappointingly small. This problem is reflected in **Table 8.13** where results using ML weights with FC SUM are only slightly better than with the STD SUM, while results using equal weights with FC SUM are worse than for the STD SUM (the only test in which STD SUM significantly outperforms FC SUM). At the time these tests were made, the FC ML weights estimation procedure (7.21) was not fully stabilized and each combination x_i was given the same number of coefficients in order to avoid the scale of $p(x_i|q_k)$ in (7.21) depending on the dimension of x_i. Correct normalization (dividing by $p(x_i)$) was introduced only after GMMs had already been trained on these x_i. Therefore, for the results reported here, larger combinations have a smaller than usual number of coefficients, although with hindsight we can see that this was not necessary. However, the use of the ML weights in the FC approach still achieved a significant improvement in WER for the band-limited noise cases as compared to the use of EQUAL weights. Further improvement can be expected in the case when larger combinations are given a higher number of coefficients.

The experiments employing quasi-optimal weights indicate how far the respective multi-band system can be improved if the noisy subband is simply ignored. They also show that, under these conditions, the lowest WER can be achieved by the FC approach. In two of the four band-limited

[8] As these experiments were carried out in the final stage of this Ph.D. no optimization of the HMM-GMM systems could be carried out as had been done for the HMM/MLP hybrid systems.

noise cases the FC approach employing ML weights is actually only insignificantly different from the "quasi-optimal" results when using Opt weights. The standard multi-band system achieved insignificantly different results employing the ML weights as compared to "quasi-optimal" results in three of the four noise cases.

Thus, in the case of full combination processing the use of (offline) ML weights (together with a hard-threshold to upweight larger clean combinations) significantly improved performance on stationary-band limited noise, outperforming both the fullband and the standard multi-band system employing the same kind of weights.

	Band-Limited Noise			Wide-Band Noise				
	Siren			Car		Factory		Mean
	0 dB	12 dB	Mean	0 dB	12 dB	0 dB	12 dB	
Fullband	79.6	43.5	61.6	76.2	34.2	87.2	34.2	58.0
STD Sum Equal	38.6*	23.6*	31.1*	88.0	68.3	87.2	68.6	78.0
STD Sum ML	35.9	21.1	28.5	88.0	64.1	86.2	67.3	76.4

Table 8.14: WERs of the baseline fullband HMM-GMM as compared to standard multi-band processing using HMM-GMMs on MFCC features. The multi-band system is tested with Equal and ML weights. * indicates that there is no significant difference to the best result in this column.

For the remaining noise conditions (cf. **Table 8.14**), experiments were only run using the Fullband and the standard multi-band system in order to evaluate the ML weights in these noise cases. As we will also observe in the multi-stream experiments in Section 10.2, (our) MFCC features result in poor performance in real-environmental wide-band noise. The standard multi-band system, moreover, degrades more than the Fullband recognizer in this kind of noise (which was also the case in this kind of noise for the (PLP and J-RASTA-PLP-based) standard multi-band HMM/MLP hybrid systems). Significant performance improvement in car noise could be achieved using the ML weights as compared to Equal weights, however only for the higher SNR-level. In factory noise, the improvement using ML weights is not significant.

In non-stationary band-limited noise, on the other hand, the MFCC features are relative robust (more than both PLP and J-RASTA-PLP as will be seen in **Table 10.3**). Moreover, standard multi-band processing significantly enhances performance. This could even be improved when the ML weights were employed, although the difference to the results employing Equal weights is not significant.

8.4.3 Adaptive SNR-based weights in HMM/MLP hybrid systems

For better noise adaptation of the HMM/MLP hybrid systems, we developed online-adaptive SNR-based weights as defined in (7.27). Estimation of the weights is described in Section 7.4.1. Results are presented for PLP features in **Table 8.15** (as well as more detailed results in Appendix G) and for J-RASTA-PLP features in **Table 8.16**.

The procedure, used to estimate SNR, first measures the noise spectrum as the average spectral power over the first 100 ms of each test utterance. It is thus best applicable to noise which is (almost) stationary within each utterance. Looking at the results of the PLP feature experiments presented in **Table 8.15**, we see that the highest gain in WER is effectively obtained on stationary band-limited noise. In the case of wide-band noise, the performance improvement is no longer significant. The same is true for the artificially created non-stationary noise case, where the noise is varying from one subband to the other for each utterance (see discussion below).

	Clean	Band-Limited Noise		Wide-Band Noise				
		Stationary	Non-Stat.	Car		Factory		Mean
		Mean	Mean	0 dB	12 dB	0 dB	12 dB	
FC SUM EQUAL	7.4	26.6	27.4	55.0	18.2	57.0	18.5	37.2∗
FC SUM SNR	7.4	23.6	26.9∗	52.0	18.0	54.4	16.6	35.3

Table 8.15: WERs of the FC SUM rule employing different weighting strategies and the FULL-BAND recognizer, on PLP features. ∗ indicates that there is no significant difference to the best result in this column (given for the Mean values only).

When SNR-based weighting is employed in our system based on the J-RASTA-PLP features, almost no improvement in robustness can be gained (see **Table 8.16**). For the wide-band and the stationary band-limited noise cases as well as in clean condition the difference in WER is not significant. The SNR-based weights are not capable of further improving the FC system employing the rather robust J-RASTA-PLP features in these noise cases.

The only (very surprising) exception constitutes the non-stationary band-limited noise case where significant performance improvement was achieved. Investigating the reasons for this, we found that the artificially created non-stationary noise was added to each utterance in such a way that the noise always occurred in the first subband at the beginning of each utterance. The variations in frequency thus occur in each utterance in the same way, and probably more often in the lower bands due to the rather short length of the utterances. As we pointed out above, the SNR-detection algorithm estimates the noise on the first 100 ms of each utterance. For this reason, in case of our artificial siren noise, it must have been the lowest subband which was consistently down-weighted. Comparing the performance of the individual experts in noise, we noticed that in the case of J-RASTA-PLP features, the subband recognizer working on subband $\{1\}$ degrades significantly more in noise than the subband expert working on subbands $\{2, 3, 4\}$ (relative to their respective performance in clean). For PLP features, on the contrary, the two subband experts degrade to a similar extent. For this reason, the J-RASTA-PLP-based system could be improved with these weights, where the lower subband was consistently down-weighted, when employed in this kind of non-stationary band-limited noise.

The adaptive weight estimation used in this section might have been oversimplified and could be improved. As with the fixed weighting of the last section, we could estimate separate weights for each phoneme. Moreover, to further improve the results in non-stationary noise, we

	Clean	Band-Limited Noise		Wide-Band Noise				
		Stationary	Non-Stat.	Car		Factory		Mean
		Mean	Mean	0 dB	12 dB	0 dB	12 dB	
FC SUM EQUAL	9.0*	17.7*	30.0	29.5	10.8	35.1	12.5	22.0*
FC SUM SNR	8.5	17.2	23.2	28.2	10.4	35.6	12.8	21.8

Table 8.16: WERs of the FC SUM rule employing different weighting strategies and the FULL-BAND recognizer, on J-RASTA-PLP features. * indicates that there is no significant difference to the best result in this column.

would need an SNR estimation procedure which does not assume that the noise is stationary (Martin, 1993; Hirsch and Ehrlich, 1995; Dupont and Ris, 1999).

8.4.4 Adaptive weights in HMM-GMM systems

	Band-Limited Noise			
	Band 1	Band 2	Band 3	Band 4
EQUAL ($\alpha = 0$)	44.0	27.3	35.4	31.9
$\alpha = 0.1$	38.1	**27.3**	**31.4**	**28.7**
$\alpha = 0.2$	38.6	27.8	32.4	30.0
$\alpha = 0.3$	**37.8**	29.0	33.2	30.7
$\alpha = 0.5$	39.6	30.0	34.6	31.2
$\alpha = 1$	41.3	30.2	32.9	32.7

Table 8.17: WERs for the multi-band HMM-GMM system (employing MFCC features) of four subbands using equal and online ML-weights on band-limited noise at 0 dB SNR. Recombination by STD SUM.

We now turn to the experiments employing adaptive ML-weights in the framework of HMM-GMM systems. The online estimated weights (as defined in (7.28)) were updated every $N = 100$ frames (1250 ms). The weights were initialized to have equal values. Results are presented in **Table 8.17**. The α-value indicates how fast the new weights are updated from the weights of the previous iteration, as was described in Section 7.4.2. It can be seen that for smaller α-values the weighting results in similar performance improvement as obtained with the offline estimation where much more data for the weight estimation was available. The performance improvement of the adaptive ML weights as compared to equal weighting is significant for band-limited noise in band 1, 3 and 4.

8.5 Summary

To sum up the experiments presented in this chapter we can conclude that (employing J-RASTA-PLP features[9]), best results were achieved

- on clean: by the FULLBAND and the full combination strategies FC SUM, FC PRODUCT, FC INDEP ASMPT, FC SUM with all weighting strategies tested (RF, LMSE, SNR) as well as all FC-ECPC. (Standard approaches not competitive).
- on stationary band-limited noise: by FC SUM with *no* significant difference to FC PRODUCT, FC INDEP ASMPT, FC POE, FC SUM with the different weighting schemes, and FC-ECPC.
- on non-stationary band-limited noise: by FC SUM employing RF, LMSE or SNR weights.
- on wide-band noise: by FC PRODUCT with *no* significant difference to (in decreasing order): FC SUM with RF and LMSE weights, FC INDEP ASMPT, FC SUM with SNR and EQUAL weights, and FULLBAND.

Hence, none of the standard multi-band approaches ranks among the best systems for any condition. The new FC strategies are competitive in clean and all noise cases as compared to the fullband recognizer. Moreover, the FC strategies are significantly and consistently better than the FULLBAND in case of band-limited noise, both stationary and non-stationary. In wide-band noise, the FC approaches and the FULLBAND do not perform significantly different. Thus, for J-RASTA-PLP features we can conclude that the use of FC SUM (with either of the here tested three weighting schemes) guarantees best performance in any condition.

In the likelihood-based systems, we tested the offline and online ML weights. In clean speech, no performance improvement could be achieved with the ML weights. In band-limited noise, both the standard and the FC multi-band systems significantly improved in performance, with the latter outperforming the standard multi-band system (with the exception of noise in band 2). The results of ML-weighting in this noise were usually close to the best-achievable results illustrated by the quasi-optimal weights. In high-SNR car noise the ML weights led to a significant performance improvement (only tested with the standard multi-band system), whereas in siren and factory noise the improvements where insignificant (as compared to equal weights). Also the adaptive ML weights led to significant performance improvement on band-limited stationary noise. In future work, adaptive ML weighting needs to be tested on non-stationary and wide-band noise.

We see in the next chapter how the FC approach originally developed for multi-band processing, can be applied also in multi-stream processing to optimally combine feature and probability combination. This way, possible correlation between the fullband feature streams is exploited and, the same time, the different feature streams are modeled independently in case one of them is severely corrupted by noise.

[9]For same evaluation on PLP features see Appendix G.

Chapter 9

Multi-stream speech recognition

In this chapter, we generalize what we have seen previously to multiple data streams towards improving the robustness of speech recognition systems.

In *multi-band* processing the different streams are constituted by different frequency subbands and possibly combinations of these. In *multi-stream* processing, as applied in this thesis, different *fullband* feature streams (and their combinations) are processed instead. The combination strategies and most of the weighting strategies presented in earlier chapters are, without any changes, also applicable to multi-stream processing.

In this chapter, we present the general multi-stream paradigm for ASR. Different motivations for multi-stream processing are discussed, including psychoacoustic and engineering motivations, and state-of-the-art research employing multi-stream processing is presented. Due to the fact that all streams stem from the same source, possible correlation information should be considered. This is realized by the FC approach which had been introduced in Section 5.6.

We investigated in the framework of FC multi-stream processing, the use of same- and multi-time scale feature streams.

9.1 Introduction

As in multi-band processing several input streams are processed in parallel. This time, the difference between streams consists in *different representations of the same source*, instead of (usually the same) representations extracted from different spectral regions as it is the case in multi-band processing. Multi-stream recognition is, amongst others, based on the observation that different representations of the speech signal often lead to different kinds of recognition errors or the same errors occurring at different points. The different streams are thus expected to complement each other at the recombination stage and to lead to a more powerful and robust performance of the combined multi-stream system.

As we will see below, multi-stream processing is also motivated from psychoacoustic research where it was found that at several stages of human auditory processing, multiple representations of the speech signal and appropriate time-scales are employed to render the final representation as robust as possible. Integration of both short- and long-term information as well as multiple representations of the input data can also be easily realized in a multi-stream automatic speech recognizer to render the ASR system more robust. Moreover, engineering motivations such as error reduction through reduced variance in a multiple classifier system, and better exploitation of training data through several, possibly smaller models sustain the multi-stream approach.

An extreme example of a multi-stream system is the ROVER system (Fiscus, 1997), where final hypotheses of complete speech recognition systems are combined. The ROVER system was able to show 30% relative error rate reduction over the best system in a NIST Broadcast News evaluation (Fiscus, 1997).

9.1.1 The multi-stream paradigm

Diversity of the streams in a multi-stream system can be obtained in various ways, such as by

- different sensory modalities, such as audio and visual data streams (Tomlinson et al., 1996; Dupont and Luettin, 1998; Rogozan and Deléglise, 1998);
- different feature processing techniques, of which multi-band processing is a special case. This also includes difference in pre- and post-processing, feature extraction, time window size and shift, derivative window size and shift, and many more (Wu et al., 1998a; Kirchhoff, 1998; Hagen et al., 2000);
- different classifiers or same classifiers using different training schemes or environments, and/or model configurations (Tumer and Ghosh, 1996; Shire, 2000);
- any combination of the above.

As one of the motivations of our work on multi-stream processing stems from the limitations we encountered in our multi-band work, we are mainly interested in extending the principles of the latter to the new approach of multi-stream processing. For this reason, our work on multi-stream processing concentrates on the same processing schemes as were employed in multi-band processing. In this thesis, multi-stream processing adheres to the second of the above categories: diversity of the multiple streams through diversity in the feature streams.

Just as in multi-band processing we can distinguish also in multi-stream processing between feature combination and probability combination.

Feature Combination In feature combination, the feature vectors are combined before acoustic modeling. This is what is usually done in most of the state-of-the-art ASR systems, where for example the first and second order time derivative features are concatenated to the instantaneous features (with possibly following LDA). Also in audio-visual speech recognition, the audio and visual feature streams are often combined in one stream which is refered to as "early integration" (Chen and Rao, 1998; Lucey et al., 2001).

Probability Combination In probability combination, the different streams are processed by specific acoustic models, the outputs of which are then recombined. In audio-visual speech recognition, this is refered to as “late integration”.

After processing the different streams for feature extraction and probability estimation, the same probability combination strategies can be applied as in multi-band processing, as can be seen in **Figure 9.1**.

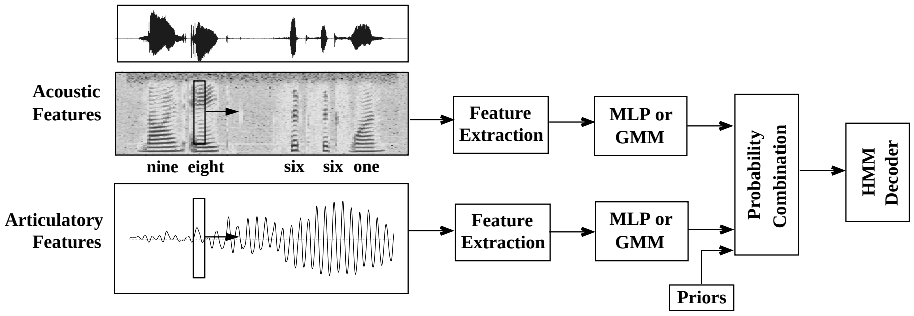

Figure 9.1: Illustration of probability combination in multi-stream ASR on two streams using different feature sets.

Special case of feature combination in the “tandem approach” A similar approach to multi-stream processing is also pursued using the recently introduced “tandem approach” (Hermansky et al., 2000; Ellis et al., 2001). In the “tandem approach”, the outputs of a neural network classifier (usually an MLP) are used as the input *features* for GMMs. The output of the GMMs provide the likelihoods for the different speech units needed in the HMM decoder. Such a tandem system thus effectively has two levels of acoustic models.

In order to render the posterior probabilities at the output of the MLP suitable for modeling by GMMs, the final non-linearity in the MLP is omitted (for the softmax non-linearity this is similar to taking the logarithms of the MLP outputs (Hermansky et al., 2000)). Then, a KL transform is applied to further decorrelate the outputs. These net outputs are interpreted as features and passed on to the GMMs.

To apply feature combination in this framework, several acoustic model ANNs are trained on different acoustic features. The net outputs, i.e. the logarithmic posterior probabilities, are then simply averaged before they are used as input features to the GMMs.

Thus, in tandem-based stream combination, the recombination of the streams is carried out on the posterior probabilities which though are interpreted as “features” for the GMM. This approach could thus be seen as either feature or probability combination.

Discussion on feature versus probability combination As feature combination of two or more streams usually leads to a rather large feature vector, which thus also demands larger models and more training data, pure probability combination might be preferred if training data is sparse. Moreover, the expected increase in robustness might be achieved more easily, if the different feature representations can each be fully exploited by one specific recognizer instead of forcing one recognizer to work on all representations. This can be illustrated as follows.

Let us assume that in the training data, several conditions occur for which one feature stream provides reliable information, whereas another feature stream does not provide any useful information. In the case of feature combination, such specific conditions are learned by the recognizer which works on the combined stream and which, thus, will have difficulties to generalize in the case where one stream offers well-known data but the other stream is in a harmful way different from the data represented in the training data.

By contrast, in probability combination, the respective stream-recognizers learn the specific good data of that stream and are not disturbed by unrecognized data in the other feature streams. They do not need to be trained on every conjunction of possible good and harmful data but will be able to generalize – up to a certain extent – to previously unseen conjunctions of feature sets as at least the probabilities from one stream-recognizer will be reliable and hopefully dominate in the posterior combination scheme.

On the other hand, for a given phoneme, different feature representations could be highly correlated. In order to benefit from this additional information which is not accessible in regular (single-stream) fullband processing, it would be better to model the combined representation by one classifier.

The above hypotheses are difficult to verify and specific tests would need to be carried out for each new feature set. As it cannot be known without extensive testing which combination approach, whether feature or probability combination or some combination of both, is best suited for a given task, we propose to follow the full combination scheme also in multi-stream processing, in which both approaches are sensibly combined.

9.1.2 FC multi-stream processing

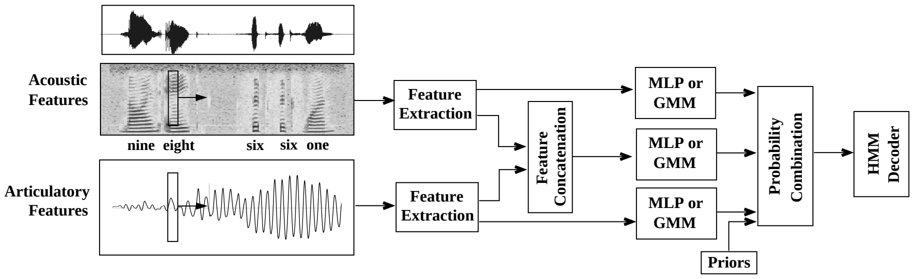

Figure 9.2: Illustration of "full combination" in multi-stream ASR on two streams using different feature sets.

In the multi-band FC approach, as described in Section 5.6, we suppose that at each instant one combination of subbands x_i (with $i = 1, \ldots, \mathcal{B} = 2^B$ combinations for B spectral subbands) carries the clean data and is best suited for identifying the current phoneme (Morris et al., 1999). As it is not known which combination of subbands comprises clean data features we integrate over all possible combinations of subbands. Likewise in multi-stream processing it is not known which stream or combination of data streams comprises clean data. In order to integrate over all possible combinations of streams, there are two steps involved in full combination multi-stream processing, as illustrated in **Figure 9.2**.

- First, the feature vectors from all streams have to be concatenated into all possible combinations of feature vectors.
- Second, each stream combination is processed independently by different experts.

In this thesis, the phoneme probability estimates from all experts are then combined by one of the probability combination strategies described in Section 5.4.1 and Chapter 6, before being passed on to the decoder. They are thus not further discussed in this framework.

9.2 Motivations

In this section, we discuss several psychoacoustic studies which have instigated, among others, the use of various different streams in ASR as this is what seems to be done also in certain aspects of human speech recognition. Moreover, from the engineering point of view several aspects are proposed which sustain the usefulness of parallel processing.

9.2.1 Psychoacoustic motivations to multi-stream processing

Listening experiments with human subjects reveal some of the "tricks" used by humans to render the human auditory system as powerful as it is. Some of these aspects are not too difficult to realize also in an automatic speech recognizer and might be able to provide some of the higher performance and noise robustness available to humans.

Multiple time scales and frequency regions In listening experiments on continuous speech from the TIMIT database, (Arai and Greenberg, 1998) found that in conditions with high SNR or non-reverberant speech, the low-frequency channels (< 1.5 kHz) seem to carry most of the speech information focusing on relatively *short-term* properties of the signal. For *long-term* information, such as syllable segmentation, on the other hand, high-frequency channels (> 3 kHz) appear to be of most importance. With a change in the acoustic environment, the 'perceptual weights' associated with a certain frequency region thus seem to adapt dynamically to upweight more reliable regions. The authors concluded by saying that the robustness of spoken language may lie in its broad and diverse capacity for encoding linguistic information across a *wide span of time scales and frequency regions*.

These findings suggest that in order to render ASR systems more powerful in noise, we should maybe look at the speech information from several *different time scales*. As short-term information is already included in almost every speech recognizer (cf. discussion on short-term analysis in Section 3.1), it is the longer time scales which might be missing in current systems.

Further experimental proof which sustains the idea of missing long-term information in conventional short-term processing of the speech signal is provided by the following study.

Missing longer time scale information In an evaluation of peripheral auditory models for mimicking human performance in the context of speech recognition, Ghitza (1994) compared speech recognition performance using their auditory model (referred to as "ensemble interval histogram") to human performance. Ghitza argues that models which are capable of mimicking human performance can provide a basis for realizing effective automatic speech processing. His system first simulates auditory nerve firing patterns which are then processed according to heuristically observed principles of the actual auditory nerve response[1]. Performance was improved when cepstral speech representation was substituted by the auditory model but was still lagging behind human performance. It was claimed that *missing integration over durations of 50-100 ms* were responsible for the shortcoming. The "tiling" experiment thus investigated human usage of such perceptually related integration rules, by interchanging time/frequency tiles from one word with the same tile of the other word in a word pair. It was found that humans utilize different time-frequency tiles to discriminate different phonemes, and that there is a direct mapping between phonemic/articulatory features and a particular tile. Ghitza (1994) further showed that humans seem to use not only different frequency bands but also different *time scales* to capture *short-term* and *long-term* information simultaneously.

Multi-band ASR investigating the use of different feature processing strategies in each subband to account for the specific phonemes which were mostly represented in that subband did not lead to conclusive results (Christensen et al., 2000). On the other hand, as we will see in the experiments, applying several different feature extractors in parallel to the *whole* frequency domain, some of which might be better suited for certain phonemes whereas others might account better for a different set of phonemes, achieved performance increase. A possible way of enhancing the integration over longer time spans in ASR where features usually are only extracted from short-term time windows is the use of *"variable window size" features* or *difference features* in separate streams, which will be discussed in Section 9.4.2.

Redundant and multiple-scale representations The speech signal is known to be highly redundant and so seems to be the human auditory system (Greenberg, 1997), as a large set of auditory experiments proposes. Such experiments include, as we already saw in Section 2.3, high- and low-pass filtering of the speech signal, filtering modulation energies (Arai and Greenberg, 1998) and desynchronizing speech energy channels (see above paragraph). In (Yang et al., 1992), the primary auditory cortex is described to contain a collection of (sequentially) *repeated representations* of the acoustic spectro-temporal information at *various scales*. To come from one representation to the next, various transformations are carried out, the first of which is described as an affine wavelet transform. This transform performs a multi-scale decomposition exhibiting progressively broader bandwidths at higher frequencies. Thus, not one, but a range of window durations are used to analyze the speech signal. More rapidly varying signals are analyzed with shorter windows. This redundancy in acoustic representation within the auditory system is one important way to guarantee high robustness if some part of the representation is corrupted by noise.

Importance of syllable-length information The importance of longer time-scale information is emphasized by (Greenberg, 1999). In this study, Greenberg (1999) systematically analyzed the phonetic properties (i.e. pronunciation variations) of spontaneous American English speech. He found that a large number of phonetic segments were either missing or changed in

[1] as exact knowledge of the latter is not available at present.

nature, i.e. transformed into other phonetic segments. These deletions and substitutions seem arbitrary but become systematic when placed within the framework of the syllable. Syllable onsets, for example, are usually well preserved which coincides with the characteristic of the human auditory system to be especially sensitive and responsive to the beginning of sound, be it speech, music or noise.

As ASR approaches nowadays usually model individual words as a sequence of phonetic elements, each of which receives equal importance, deletions and substitutions of these elements are devastating. If we could provide automatic speech recognizers with higher-level information than solely phoneme-based information, (formerly unseen) pronunciation variations could be better accounted for.

9.2.2 Engineering motivations to multi-stream processing

The paradigm of using an ensemble of trained classifiers instead of simply using only one classifier has been widely proposed in the literature (Hansen and Salamon, 1990; Jacobs et al., 1991; Jordan and Jacobs, 1994; Bishop, 1995). The idea behind using multiple classifiers is that, in the absence of the "true" model, the apparent best classifier can be improved upon by employing several classifiers to solve the classification task independently and then construct a final decision by making use of the individual scores (Hashem, 1997). This is motivated by the fact that different classifiers exhibit distinct recognition characteristics and, thus, also commit different types of errors.

Error reduction through reduced variance In Bishop (1995), it was shown that for a (weighted) average combination of classifiers the expected square error can theoretically be reduced by a factor equal to the number of experts in the committee, under the assumption of the errors having zero mean and being uncorrelated. In practice, errors are highly correlated and error reduction is thus much smaller. The gain which is still encountered can be ascribed to the reduction in variance due to the averaging over many solutions. Combining the outputs from several classifiers therefore usually results in higher performance than any one of the classifiers by itself (Bishop, 1995).

Better exploitation of finite training data and hardware As we are usually confronted with the problem of finite training data and time, but, on the other hand, have access to multiple machines, multiple (smaller) classifiers can easily be trained in parallel and then used in combination (Janin et al., 1999). Moreover, hardware limits often already preclude the increase in size of a single recognizer (such as an MLP with a larger hidden layer). Instead, an even higher benefit can usually be reached by training several smaller models. Hereby, better results are not necessarily achieved by the combination of classifiers with better individual performance. Often, classifiers were found to complement each other well, although one of the classifiers had poor performance by itself (Rogova, 1994; Sharma et al., 2000).

Decomposition of acoustic models In the hierarchical mixture of experts approach (Jordan and Jacobs, 1994) and the ACID/HNN[2] architecture (Fritsch, 1998), the acoustic modeling task is decomposed into smaller, modular tasks in a data-driven manner. Separately trained neural network classifiers estimate class posterior probabilities on different sub-sets of the acoustic input, employing uniform feature extractors for each acoustic unit.

Another approach is modular decomposition based on fixed class units, such as phonemes. So-called phonemic neural networks were used in (Waibel et al., 1989; Auda and Kamel, 1998). It is, among other things, argued that the usually used stopping criterion evaluated on all classes (in a monolithic ANN) does not guarantee the best stopping point for any particular class. Training one network for each class circumvents this problem.

One of the disadvantages of the modular decomposition approaches lies in the fact that the modeling of (phoneme) boundaries is repeated in each modular network. Moreover, the combined system contains, in total, a high number of parameters, though it is argued that these parameters are not extra degrees of freedom but rather represent a repetition of the same modeling.

Complementary classifiers for ASR In the specific case of automatic speech recognition, the performance of the ensemble is found to improve if, for example, speaker-specific effects (such as gender, age, dialect or speaking style (Mirghafori et al., 1994)), environmental effects (such as background noise, microphone, etc.) or speaker variability can be accounted for by complementary parallel processing. This is achieved through diversity obtained from training of different classifiers on each of the different conditions (Tumer and Ghosh, 1996), such as one system trained on male speakers and one trained on female speakers. Similarly, different classifiers can be trained for a variety of noise characteristics in order to being able to account for as many noise environments as possible (Shire, 2000).

Such effect-specific training is unfortunately often not applicable as the range of variability is too high. Especially in the case of noisy speech, the exact noise characteristics can never be foreseen. Appropriate training for each kind and level of noise, in the multi-stream framework is therefore not possible. Other ways apart from effect-specific training for diversity need to be found.

Further possibilities for achieving diversity Due to the large amount of training algorithms which are now available, distinctiveness among classifiers, working on the same output space, could already be achieved by simply employing different *training algorithms*. Alternatively, different types of training data (i.e. subsets of the training data), classifier architectures and/or features could be used for the training of each component classifier, which will be discussed in the following.

As *training data* can be expensive and time-consuming to create, it can generally be considered as a sparse resource. Good generalizability to unseen data, however, demands sufficient training data. For this reason, the alternative of splitting up the training data into several smaller sub-sets should be discarded in this framework.

[2]This described an Agglomerative Clustering algorithm based on Information Divergence (ACID), which automatically designs Hierarchies of Neural Networks ACID/HNN.

The right choice of *classifier architecture* is usually hard to foresee and often undergoes several trial-and-error approaches. When suitable architectures are finally found, such as the right number of hidden layers and hidden units in the case of ANN classifiers, it is often observed that different kinds of features lead to more independent results leading to higher error complementarity than different kinds of architectures (Rogova, 1994).

With the wide variety of *feature extractors* available for automatic speech recognition, the most promising approach, thus, seems to be to base the diversity of ensemble classifiers on the diversity of the input streams. Acoustic features are often known to work especially well under certain environmental conditions, whereas others show their advantage under completely different conditions (which might be known or not). Alternate feature processing strategies, which contain overlapping information, also contain abundant information about speech cues which is *not* available when only employing one of the preprocessing strategies.

The use of several different feature streams not only comprises employing different *kinds* of features but also different processing strategies for the same kind of features. More precisely, acoustic features can (i) be extracted over different time scales, thus providing more local or more contextual information, as well as (ii) undergo various pre- or post-processing strategies such as compression or derivation.

The diversity of available feature streams and their complementarity in different application domains can be well exploited through the use of multiple recognizers, each of which is trained on a different feature stream. The ensemble of all recognizers then more easily generalizes to a wider range of applications. This approach is followed in this thesis and will be discussed in more detail below.

9.3 State of the art

In parallel to the application and extension of our multi-band work to the area of multi-stream research, other research labs moved in the same direction. In this section, we therefore illustrate multi-stream research as carried out at other institutes.

Phonemic Neural Networks with diverse features Separately trained phonemic neural networks were used in (Antoniou and Reynolds, 1999) to estimate each phone's posterior probability, matching network resources and training to the needs of each phone respectively. In (Antoniou and Reynolds, 2000), this approach was extended to also incorporate diverse feature streams in the acoustic model for each phone. Three acoustic front-ends (MFCC-, PLP- and LPC[3]-features) were employed per phone, training a "primary" MLP classifier on each of them. The three "primary" MLPs were then recombined by a "posterior net" (recombining MLP) to estimate the phone posterior probability. To see whether the improvement in WER actually stemmed from the combination of front-ends, rather than the pure combination of classifiers, an ensemble of networks was set up using four nets trained with the same features but different weight initializations. This ensemble resulted in a lower improvement of the WER.

These experiments showed that the combination of different acoustic front-ends leads to

[3]LPC denotes Linear Predictive Coding.

higher performance improvement than several networks trained on the same features but with different weight initializations.

Combination of non-linearly transformed feature streams In (Sharma et al., 2000), a set of four different feature streams was investigated: PLP, RASTA-PLP-like features with LDA, modulation-filtered spectrogram (MSG) features (Greenberg and Kingsbury, 1997) and TempoRAl Patterns (TRAP) features (Hermansky and Sharma, 1999). Only the spectral-based PLP features were extracted on short-term information of 10 ms whereas the other 'temporal-based' feature sets were calculated on 1 ms of input speech so that each set was expected to result in complementary errors. Non-linear transformation of each feature set was carried out through the application of an MLP of which the softmax non-linearity had been removed so that the distribution of the features came closer to a Gaussian distribution. The outputs were then diagonalized through a KL transform for the following modeling by an HMM using diagonal covariance matrices. Their systems were thus similar to HMM/MLP hybrids. The non-linearly transformed features led to considerably improved robustness on the matched AURORA (version 1.0) (Hirsch and Pearce, 2000) task (training and testing in noise). Several combinations of the four feature sets were tested, where combination was carried out through averaging of the non-linearly transformed features (i.e. net outputs) prior to orthogonalization. Significant average reduction in WER as compared to the AURORA baseline (using MFCCs) fullband system could be achieved with each of the combinations.

From the above set of experiments, it can be seen that combining different (non-linearly transformed) feature sets, especially as complementary feature sets as those based on spectral and temporal processing, yields significant improvement in performance.

Averaging logarithmic probabilities from different streams In the framework of HMM/-MLP hybrid systems, Janin et al. (1999) investigated the performance of different sized MLP experts and their combinations, trained on PLP and MSG features. With the goal of equalizing the learning 'capacity' of the nets with differently sized inputs through the adaptation of the size of the hidden layer (instead of the number of parameters), several sets of nets with different sizes of hidden layer were trained on each feature stream. Performance comparison on the BROADCAST NEWS corpus (Cook et al., 1999) between nets of smaller and larger numbers of hidden units showed that bigger hidden layers gave consistently better performance. Moreover, combinations of the fullband streams through averaging of their logarithmic probabilities resulted in lower WERs when MLPs trained on *different* feature streams were combined than combining MLPs trained on the same feature stream but with different random starting points.

The main conclusions from this work are:

- Contrary to other reports, bigger neural networks gave consistently better performance,
- once again, MLPs trained on different feature streams proved more powerful when their outputs were recombined than MLPs trained with different starting points.

Feature and probability combination In the framework of HMM/MLP hybrid systems, Ellis (2000) carried out performance comparisons among feature and/or probability combinations on four sets of features: PLP static and delta features, MSG features covering modulation frequencies of 0-8 Hz and of 8-16 Hz, respectively. Probability combination was realized by averaging of the logarithmic posteriors at the outputs of the MLPs. In order to decide which feature streams were better suited for feature combination and which for posterior combination, a simplified version of the MI (cf. Section 3.2) was used to calculate the statistical dependence between the feature streams. It was argued that the higher the dependence between feature streams (and, thus, higher conditional MI) the better feature combination would be suited. Unfortunately, this hypothesis was only weakly supported by the experiments carried out under the matched multi-conditional task of the Aurora database. No consistent relation between higher (lower) MI and performance improvement due to feature (probability) combination could be shown. For probability combination it seemed the most appropriate to combine the "most different" features but the overall best systems were achieved by a mixed combination of both feature and probability combination.

This work of research shows that the decision between feature and probability combination is not easy to make. This often results in several trial-and-error approaches which finally lead to a good system for the specific testing conditions without any guarantee that under different conditions the same mixed combination would also achieve best performance.

Training acoustic front-ends under different acoustic environments Also in (Shire, 2000), multi-stream recognition performance was investigated employing multiple front-end acoustic modeling stages whose acoustic probability estimates were merged on the frame level for further processing. Each of the front-end stages was trained to enhance phone classification in a specific acoustic environment. The front-end systems were then combined in a multi-stream setting. Tests on homogeneous feature streams (RASTA-PLP with different LDA filters) showed that (i) training in noise degrades performance in clean (already for a single-stream system); (ii) when combining the posteriors from two MLP estimators, one of which is trained in clean conditions, the other in noise, the combined system degrades on matched conditions but improves on mismatched conditions.

Employing dual feature streams (PLP and MSG) trained on the *same* environment, the combined streams (posterior combination) resulted in significant improvement in most of the cases for both matched and mismatched conditions. Using the same streams (PLP and MSG, respectively) trained on *different* acoustic environments gave improvement only for mismatched data. No further improvement could be achieved when both MLPs were trained on both acoustic environments (4-stream combination).

Two weighting strategies based on frame-level confidence measures were set up, based on (i) maximum posterior values and (ii) information theoretic measures. These frame-entropy measures, however, did generally not prove reliably useful when compared to equal weights.

The following conclusions can be drawn from this work of research:

- Multi-style training in different acoustic environments does not provide a solution if good performance under *all* conditions is sought;
- Using two different feature streams, training in the same environment more often achieved performance gains when the stream probabilities were combined than when the streams

were trained in different environments.

- It is difficult to improve the performance of equal weights with new weighting strategies also in multi-stream processing.

Combining recognizers based on syllable and phone time scales Wu et al. (1998a) investigated the usefulness of syllable time scale information in ASR by developing a speech recognition system which focuses on information encoded over syllabic durations and comparing its performance to a standard (short-term) recognizer which focuses on information at the phonetic segment scale. Moreover, a third recognizer was included in the comparison which combined the syllable-based and phone-based recognizers into a single system. The phone-based recognizer employed log-RASTA-PLP features. The syllable-based recognizer employed MSG features which are described to blur envelope fluctuations at the phonetic segment scale (due to narrower filters) while emphasizing changes at the syllabic scale. Combination of the two classifiers was carried out on the utterance level using N-best lists. While the baseline phone-based recognizer performed better than the syllable-based recognizer in both clean speech and reverberation noise, the combined system demonstrated significantly higher performance than the baseline system on both test conditions.

In a further study, Wu et al. (1998b) investigated combination of phone- and syllable-based systems at three distinct levels: the frame, the syllable and the entire utterance level. Frame-level recombination was carried out by multiplication of phone probabilities, syllable-level combination by HMM-recombination (cf. Section 5.1) and utterance-level combination by merging and re-scoring N-best lists. Although each recognizer performed well by itself, combining the syllable-based systems with the (phone-based) baseline resulted in significant lower WERs in both clean speech and reverberation noise. Of the three recombination methods, syllable-based combination displayed the largest improvement, closely followed by frame-level combination which has lower implementation cost.

These results show that knowledge derived from both syllable- and phone-length time scales employed jointly in an automatic speech recognizer can yield performance superior to that obtained using information derived from either time scale alone (although the syllable-based recognizer performed worse than the (phone-based) baseline on its own). Moreover, it could be seen that frame-level combination resulted in almost as high performance as syllable-based combination with the advantage of significantly lower computational cost.

Combining recognizers based on acoustic and articulatory features In addition to an acoustic signal representation heuristically defined articulatory features describing manner and place of articulation were used as a supplementary source of information in (Kirchhoff, 1998). The use of articulatory features is motivated by the assumptions that (i) coarticulation can be modeled more easily in the production based domain than in the acoustic domain, and (ii) articulatory parameters are more robust to cross-speaker variation and signal distortions. The acoustic (log-RASTA-PLP and MSG) and articulatory feature based HMM/MLP hybrid recognizers were tested on the Numbers95 corpus under different environmental conditions by themselves and in combination. MLP classifier combination was carried out on the posterior probabilities using the STD Sum and STD Indep Asmpt. Whereas the acoustic systems performed slightly better in the case of clean speech and noise at high SNR, the articulatory systems showed a distinct advantage in the presence of noise at low SNR. Combination of both systems improved the WER.

In (Kirchhoff et al., 2000), these experiments were extended to a large-vocabulary conversational speech recognition task, this time using HMM-GMM systems on MFCC features. Recombination was carried out at the state, the word and the feature level, of which state-level combination yielded the best results, with the STD INDEP ASMPT rule outperforming the MINIMUM, MAXIMUM and STD SUM rule (cf. Chapter 6 for description of these combination strategies).

In (Kirchhoff and Bilmes, 2000) several new combination rules were proposed these being continuous and differentiable extensions of well-known combination strategies such as STD SUM, STD PRODUCT, MINIMUM and MAXIMUM rule, which can, thus, also be used as objective functions in a joint classifier training scheme. None of the joint training results outperformed the product rule combination of individually trained classifiers. Only in the case of joint embedded training and product rule combination, significant improvement in WER could be achieved.

The main conclusions which can be drawn from these works are:

- Probability combination of streams using different acoustic and articulatory based features also improves performance.
- Probability combination out-performed feature combination also on this task, and state-level combination out-performed word-level combination.

Heterogeneous features in multi-band and multi-stream processing In (Christensen et al., 2000), diverse feature information (from PLP, J-RASTA-PLP and MFCC features) was employed in both multi-band and multi-stream HMM/MLP hybrid systems to investigate the question as to whether diversity of the feature streams could achieve improved performance and/or noise robustness. In the multi-band framework, this was obtained by using different feature extraction techniques for each subband. A significant decrease in WER could be achieved by the final best set-up as compared to the multi-band systems using the same kind of feature in each subband.

In the multi-stream framework, feature concatenation as well as probability combination (using STD SUM and STD PRODUCT) were performed. Most, but not all, combinations resulted in higher performance on clean data as compared to the single-stream recognizers, showing that the correct choice of features and combination strategy is essential for performance improvement in clean speech. As the multi-stream systems have a higher number of parameters, different numbers of hidden units were tested for the single-stream recognizers to investigate whether it was the increase in parameters which led to improved performance. None of the monolithic recognizers could achieve as good WER performance as obtained by some multi-stream systems. This shows that it is not the size of the recognizer, but the right combination of feature or probability streams which allows for better performance.

Similar results were obtained by Meinedo and Neto (2000) who investigated multi-stream processing employing such diverse features as PLP (or log-RASTA-PLP) and MSG features. The multi-stream systems were found to improve over the single-stream recognizers in most of the cases, where special care was taken to develop (single- and multi-stream) systems of similar number of parameters. Investigating the possible additional gain when a smaller multi-stream constituent model was substituted by one of the larger baseline systems, only a small reduction in error rate was observed.

From these pieces of work we can see that

- performance improvement in clean speech is possible in multi-stream (and multi-band) processing by increasing the heterogeneity of the input features, though finding the right feature or probability combination is difficult;
- larger MLP classifiers using only one feature stream cannot achieve as low WERs as combination of smaller MLPs employing diverse feature streams.

Auditory and visual stream combination For several years now (Tomlinson et al., 1996; Dupont and Luettin, 1998; Rogozan and Deléglise, 1998; Teissier et al., 1999; Glotin, 2000; Heckmann et al., 2001), audio and visual streams are being combined in order to render the automatic speech recognizer more robust. Several studies have shown that the use of information on lip movement, in addition to the acoustic information, can significantly improve the recognition performance in the case of noisy speech. The audio and visual stream have the advantage that, although they stem from a common source (i.e. from one speaker) they still exhibit largely autonomous and complementary information. For instance, discrimination between the two phonemes /t/ and /p/ is accomplished more easily with visual information than with acoustic information (Dupont, 1997).

Heckmann et al. (2001) investigate audio and visual feature streams in the framework of the FC approach. The audio feature stream consists of (log) RASTA-PLP features, and the visual feature stream of 6 lip parameters, such as outer lip width, inner lip width, outer lip height, etc. Although the visual stream has only a limited, though noise-robust, performance capacity, the combined system degrades much slower and more gracefully than each of the streams by itself.

9.4 FC multi-stream employing diverse acoustic feature streams

In this section, the FC multi-stream work investigated in this thesis is presented. In this framework, we employ three different sets of features. The first set of features consists of diverse acoustic features extracted on a single time scale from different feature extraction techniques, the other two sets employ multiple time scale features, which stem from the same extraction technique (such as PLP or J-RASTA-PLP) but imply either different parameterization (such as window size) or different post-processing schemes (such as time differentiation).

Let $X = \{x_1, \ldots, x_t, \ldots, x_T\}$, $Y = \{y_1, \ldots, y_t, \ldots, y_T\}$, and $Z = \{z_1, \ldots, z_t, \ldots, z_T\}$ denote the acoustic vector sequence of three parallel streams, with x_t, y_t and z_t higher-dimensional acoustic vectors at time frame t ($t = 1, \ldots, T$). In a state-of-the-art recognizer, these three streams are usually concatenated and jointly processed, such as in the case of the instantaneous, first and second order time derivative features, resulting in one feature vector (x_t, y_t, z_t) for each time frame. In the framework of our FC multi-stream system, however, each stream is processed by itself as well as concatenated with every other feature stream. This way, the usually necessary search for the best feature and/or probability combination (cf. discussion in Section 9.1) is avoided, as all combinations are considered. Unreliable combinations will have little effect in the combination process as they tend to have high entropy. FC multi-stream processing employing three feature streams is illustrated in **Figure 9.3**. The probability estimates from the different experts are then recombined according to the FC formulae described in Sections 6.1 to 6.5.

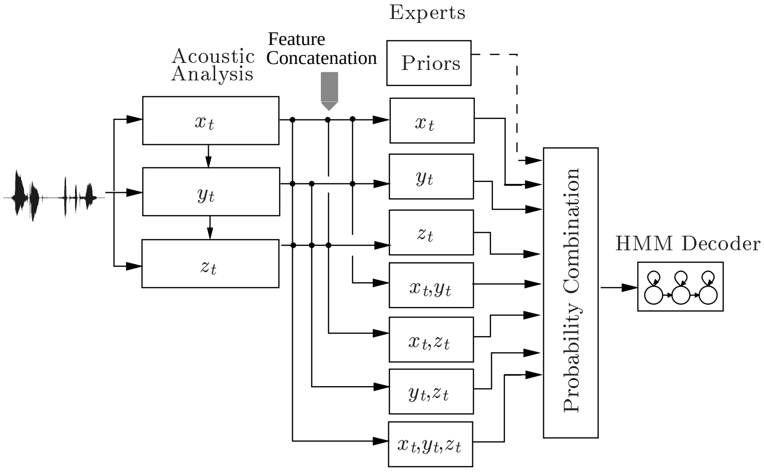

Figure 9.3: Illustration of recognizers combination according to the "full combination" approach, using three different feature streams (x_t), (y_t) and (z_t) as well as all possible combinations of feature streams $((x_t, y_t), (x_t, z_t), (y_t, z_t), (x_t, y_t, z_t))$ in the framework of an HMM/MLP hybrid system.

In multi-stream processing in this thesis, we exclusively work with HMM/MLP hybrid systems. It is therefore made use of the posterior-based FC recombination strategies. Denoting the set of acoustic vectors as S, we can write the FC formulae without any changes, using s_i with $i = 1, \ldots, \mathcal{B}$ to denote a certain stream at time t, as follows. The FC Sum (6.3) is

$$P(q_k|s) \quad = \quad \sum_{i=1}^{\mathcal{B}} P(q_k|s_i) P(b_i|s)$$

the FC Geom Mean (6.24)

$$P(q_k|s) \quad = \quad \Theta_k \, \Theta \, \frac{\prod_{i=1}^{\mathcal{B}} P^{w_i}(q_k|s_i)}{P^{(\sum_i w_i)-1}(q_k)}$$

and the FC Indep Asmpt (5.12) when applied to FC processing

$$P(q_k|s) \quad = \quad \Theta \prod_{i=1}^{\mathcal{B}} P^{w_i}(q_k|s_i)$$

with Θ a normalization constant, such that $\sum_{k=1}^{K} P(q_k|s) = 1$.

9.4.1 Single time scale feature streams

For the set of single time scale features, we chose acoustic features which are well-known for their good performance and, thus, widely used in speech processing. Firstly, these are the PLP features which are particularly powerful in clean speech which could be seen in the baseline fullband experiments carried out in Section 8.3.2 (**Table 8.3**). Secondly, we selected the J-RASTA-PLP features for their high-noise robustness, although they often imply slight degradation in clean speech. The J-RASTA-PLP features are filtered PLP features as was described in Section 4.2.4. The filtering suppresses low modulation frequencies which usually stem from channel characteristics or other non-speech artifacts. Thirdly, we chose MFCC features which are also widely used and known for their high noise robustness (cf. Section 4.2.1).

The experiments for multi-stream processing employing these feature streams in the framework of the FC approach are reported in Section 10.2, and compared to results using standard combination schemes.

9.4.2 Multiple time scale feature streams

Due to the inertia of the human speech production apparatus, the speech signal can be assumed to be short-time stationary in segments of 10 to 30 ms, allowing for feature extraction at these intervals. This short-term processing of the speech signal, however, also leads to loss of time contextual information. Local context effects such as variations within an utterance due to coarticulation and other suprasegmental factors such as stress and emphasis can sometimes be accounted for through the use of context dependent phonetic models (such as biphones or triphones). This, on the other hand, increases the size of the MLPs or number of GMMs in the models considerably which causes problems due to sparsity of training time and data. Moreover, the difference in word-internal or cross-word context dependency often needs to be modeled explicitly.

For these reasons, the possibility of employing context information by including larger time-scale information at the pre-processing stage is an attractive alternative allowing us to circumvent these problems. In some recognizers, such as MLP experts, a long input window can be employed. Another alternative is the use of first and second order derivative features which reflect the development of the acoustic features over time and thus give insight into longer temporal information of the speech signal.

As we have seen above, psychoacoustic studies show that the human auditory system integrates information from several, and also much larger, time spans than the temporal duration of the usually used time window in ASR. The maximal time span is often reported as approximating the length of a syllable which is around 100 to 250 ms long (Wu et al., 1998a).

We thus investigate the use of *multiple time scale* features also in ASR and more specifically in the framework of multi-stream processing. By modeling these characteristics found in human auditory processing, we hope to approach the high noise robustness that the human auditory system offers. Multiple time scale features seem especially promising in the multi-stream framework as they extract different information in each time scale which can be exploited by the different streams. The shorter time scales are often found to provide good performance in clean speech whereas the longer time scales usually show higher robustness to noise (McCourt et al., 1998; Wu et al., 1998b; Hagen and Bourlard, 2000; Sharma et al., 2000; Weber, 2000).

We investigate two different sets of multiple time scale features. The first set consists of features which are extracted over variable sized windows of three and five times the original window size. For the second set, we take as separate input streams the commonly used difference features, that is the first and second order time derivatives of the instantaneous features, as introduced in Section 4.2.4.

In the following, we show for both methods how the FC approach or an approximation of it can be used to combine features from different time scales.

Variable window size features

In this approach, acoustic features are calculated using windows covering different time spans of the speech signal. These features are then combined to form a single feature vector which is processed in the usual way by a single speech recognizer. We refer to the combined features as the *"variable window size" features.*

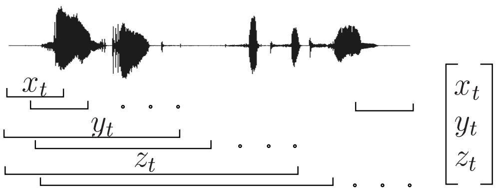

Figure 9.4: Illustration of multiple time scale features extracted from a regular-sized ($X = \{x_1, \dots, x_t, \dots, x_T\}$), triple- ($Y = \{y_1, \dots, y_t, \dots, y_T\}$) and quintuple-sized ($Z = \{z_1, \dots, z_t, \dots, z_T\}$) data window. They are then concatenated to yield the multiple time scale feature vector.

In our case, the first time scale features are extracted from regular short-term segments of 25 ms windows, shifted every 12.5 ms, and are denoted $X = \{x_1, \dots, x_t, \dots, x_T\}$. The second set of features is extracted on segments which span three times the original window size, covering 75 ms of speech ($Y = \{y_1, \dots, y_t, \dots, y_T\}$). The third set of features ($Z = \{z_1, \dots, z_t, \dots, z_T\}$) span 125 ms by using five times the original window size. Window shift of 12.5 ms is equal for all sets. This procedure is illustrated in **Figure 9.4**. Since these multi-scale features are clearly correlated, and not very likely to function well by themselves, they are respectively concatenated to the instantaneous feature vector (x_t), resulting in feature vectors (x_t, y_t), (x_t, z_t) and (x_t, y_t, z_t), which are used as independent feature streams in multi-stream processing. For each feature combination, one MLP expert is trained which is expected to model the correlation across input features.

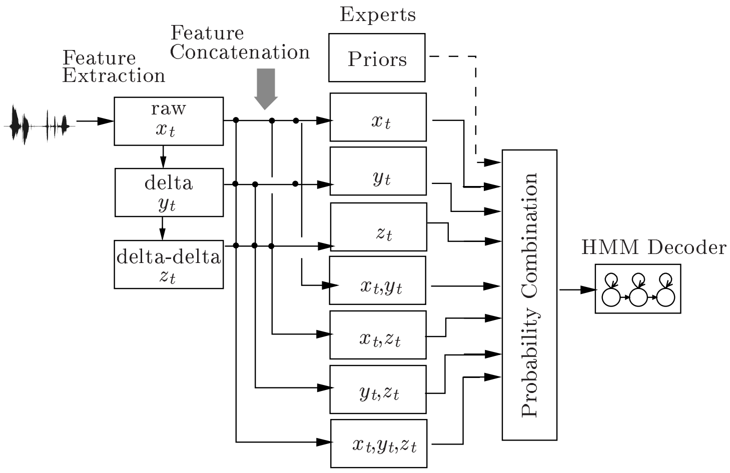

Figure 9.5: Illustration of recognizers combination according to the "full combination" approach, using raw ($X = \{x_1, \ldots, x_t, \ldots, x_T\}$), delta ($Y = \{y_1, \ldots, y_t, \ldots, y_T\}$) and delta-delta ($Z = \{z_1, \ldots, z_t, \ldots, z_T\}$) features as individual input streams as well as all possible combinations of feature streams in the framework of a HMM/MLP hybrid system.

Difference features

Other independent features which are often used to complement instantaneous features in state-of-the-art automatic speech recognizers are the temporal first and second order derivative features, which cover larger time scale information than the instantaneous features. They are thus often found to be more robust to noise.

Huang et al. (1993), for example, used first-order difference features extracted from both 40 ms and 80 ms derivative windows. They found that augmenting their LPC cepstral feature vectors only consisting of static and delta features (including energy) by delta features calculated on a longer derivative window (80 ms) as well as adding delta-delta features (including energy) reduced errors by over 15% as compared to the baseline system. In (Furui, 1986b), it was observed that delta (cepstral) features were able to reduce distortions from convolutive noise. The delta features are usually appended to the static feature components, which, unfortunately, makes the concatenated feature vector again less robust to noise. We therefore propose to use them as separate streams in the multi-stream framework.

In this thesis, the first order difference features (also refered to as delta features) are calculated from over five instantaneous features $y_t = \Delta\, x_t = [-2x_{t-2} - x_{t-1} + x_{t+1} + 2x_{t+2}]$, and the second order difference features (also refered to as delta-delta features) over seven instantaneous features ($z_t = \Delta\Delta\, x_t = [2x_{t-3} + x_{t-2} - 2x_{t-1} - 2x_t - 2x_{t+1} + x_{t+2} + 2x_{t+3}]$). They thus cover mid-term and long-term speech information of roughly syllable length as compared to the static features which usually only account for short-term speech portions. As these difference features are more or less independent of the instantaneous features they can more appropriately be treated as separate, higher time scale feature streams than the previously introduced "variable window size" features.

For FC processing, these three feature streams are first concatenated to give all possible combinations of feature streams. In our case, this amounts to a total of seven streams. For each of these seven streams a separate recognizer is then trained. The probabilities at the output of the experts are combined according to the FC recombination strategies. This procedure is illustrated in **Figure 9.5**.

The experiments which were carried out for multi-stream processing employing the feature streams presented in this chapter are discussed in Chapter 10.

9.5 Summary

The general approach to multi-stream processing is to process several information streams, stemming from the same source, in parallel up to a certain point where their representations are recombined. This often leads to higher noise robustness than processing by a single recognizer.

Previous research on multi-stream processing showed that diversity in the type of features in each stream usually leads to higher performance of the combined system than diversity in acoustic data (such as different noise environments) or in the acoustic models.

When different streams are employed, it needs to be decided whether they are best recombined at the feature or at the probability level. The common solution to this problem is to search by "trial-and-error" which feature streams are best modeled jointly ("early integration") and which ones by separate acoustic models ("late integration") to provide higher noise robustness. In order to circumvent this problem we propose the FC approach for multi-stream processing, which integrates over all possible combinations of streams, and thus considers both early and late integration.

In the framework of FC multi-stream processing, we described the use of same and multiple time scale features. An improved way to combine the instantaneous and time difference feature streams was proposed by integrating over all possible stream combinations, which results in improved noise robustness as compared to the usually used simple concatenation of these feature streams, as will be seen in the next chapter.

Chapter 10

Experimental evaluation of multi-stream processing

In this chapter, the results of our multi-stream experiments are presented. Our main goal in multi-stream processing is to find robust speech recognition techniques which can improve recognition performance also on real-environmental noise cases which has not been the case when using multi-band processing. For this reason, most of the experiments are carried out on the wide-band noise cases.

In this chapter, we report the experimental results achieved by using our new FC multi-stream approaches by (i) making use of different feature extraction techniques in each stream, (ii) complementing the usual instantaneous features with features extracted from different sized analysis windows and (iii) better use of temporal derivatives.

10.1 Baseline systems

In this section, we introduce the HMM/MLP hybrid fullband recognizers used in our multi-stream systems.

HMM/MLP hybrids Two of the three multi-stream HMM/MLP hybrid systems used in the first set of experiments are the baseline fullband hybrids as employed in the experiments on (posterior-based) multi-band processing (namely the PLP and J-RASTA-PLP-based fullband recognizers). The third fullband hybrid recognizer employs the same MLP setup as described for the other two fullband recognizers in Section 8.1 with the only difference that it is trained on MFCC features. The MFCC features comprise 12 coefficients extracted over 26 filters, and the energy. Just as for the other two feature sets, the first and second order time derivatives of the 13 coefficients are appended to the static feature stream. For the MLPs trained on the concatenated feature vectors (such as PLP-MFCC), the size of the hidden layer was chosen proportionally to the number of features in the vector (see **Table J.1** of Appendix J).

Similarly, in the set of experiments using multiple time scale features the size of the hidden layer of the MLPs was chosen proportionally to the size of the input feature vector. The exact numbers of hidden units and MLP parameters are given in **Tables J.2 and J.3** of Appendix J. The multiple time scale features are described in more detail in subsequent sections.

Training of all fullband classifiers was carried out on the same database (Numbers95) and the same set of (clean) utterances as described in Section 8.2, but with (i) different feature extractors (such as for PLP and MFCC features) or (ii) different pre-processing strategies (such as for PLP and J-RASTA-PLP features).

Posteriors or scaled likelihoods Whether to divide the net outputs by the prior probabilities or not has, this time, to be decided for all fullband streams in common. As the PLP and MFCC features result in better performance without division, the J-RASTA-PLP features, on the other hand, with division we decide for the least loss in performance, that is when outputs of the J-RASTA-PLP recognizers are also not divided by the priors. The results of the J-RASTA-PLP fullband recognizer in this section can thus not be directly compared to the J-RASTA-PLP fullband results in Section 8.3 where results are presented on scaled likelihoods (but to the ones used in the ECPC experiments in Section 8.3.5). The PLP fullband recognizer corresponds to the one used in all PLP-based multi-band experiments (see also Appendix G).

Full combination processing As for FC multi-band processing, each stream combination is processed independently by a different expert. The phoneme probability outputs from all experts are then combined via the FC formulae, before being passed on to the decoder. For three feature streams as employed in our multi-stream experiments, we thus have eight stream experts[1].

System evaluation We have to bare in mind that in a standard (fullband) recognizer which only uses one feature stream, this feature stream is chosen once and then fixed. As it can never be foreseen which features work best in which recognizer and under which conditions, this choice is often rather arbitrary (as deciding between MFCC and J-RASTA-PLP features, for example). For the evaluation of our multi-stream results we now have access to the performance of three standard fullband recognizers which (i) shows the different performance of the features in different noise cases, and (ii) makes the rating of the multi-stream system harder as we are suddenly confronted with three (if not even eight) "baseline" systems. It thus needs to be decided to which "baseline" system the multi-stream system is to be compared. Taking the average over all noise cases (including clean speech) which we examined, we found the J-RASTA-PLP features to be the most robust and thus, decided on the J-RASTA-PLP fullband recognizer for our baseline system for the experiments on single-scale feature streams.

For the experiments employing multi-scale feature streams, it is the fullband recognizer which works on the usually used time scale and thus was also the baseline fullband recognizer in the multi-*band* experiments which constitutes the baseline.

[1] of which the stream "expert" representing prior information only was excluded.

10.2 Single time scale features

In the first set of multi-stream experiments we use diverse features estimated on the same temporal scale, which corresponds to the time scale used in the multi-band experiments (window length of 25 ms, shifted every 12.5 ms, as described in Section 8.1). Three different kinds of acoustic features are employed: PLP, J-RASTA-PLP and MFCC features.

In this FC multi-stream system, we have three recognizers, which work only on one feature stream each, and the recognizers trained on each possible combination of feature streams which amounts to another set of four recognizers. Results are presented for each of the seven recognizers (the combination which includes no features being excluded) by itself, and for the FC multi-stream system using the FC SUM (with equal and relative frequency (RF) weights from (7.12)) and FC PRODUCT combination strategies (cf. Equations (6.3) for FC SUM, and (6.23) for FC PRODUCT).

The FC multi-stream systems are compared to standard multi-stream approaches where combination is carried out by STD SUM or STD PRODUCT (according to (5.8) and (5.11)).

10.2.1 Experiments in clean speech

Recognition performance in clean speech is competitive for almost all streams tested separately and all combination systems, with no significant differences as compared to the FC SUM which has lowest WER, as can be seen in **Table 10.1**. Only the fullband recognizer using MFCC features is significantly worse than the FC SUM. No improvement could be achieved using FC SUM with RF weighting or FC PRODUCT. For comparison, in **Tables H.1** to **H.3** in Appendix H

	Clean
J-RASTA-PLP	8.0
PLP	7.1
MFCC	8.6
PLP-J-RASTA-PLP	8.0
PLP-MFCC	6.9
MFCC-J-RASTA-PLP	7.1
PLP-J-RASTA-PLP-MFCC	6.9
STD SUM	6.9
STD PRODUCT	7.0
FC SUM EQUAL	6.7
FC SUM RF	6.9
FC PRODUCT	7.5

Table 10.1: WERs for multi-stream processing in clean speech, using PLP, J-RASTA-PLP and MFCC features. FC processing is compared to standard multi-stream processing where only the three single-feature-based probability streams are recombined as well as to the J-RASTA-PLP baseline. Results for each of the seven constituent recognizers which are used in FC processing are also given (Lines 1 to 7). RF refers to relative frequency weighting. Only MFCC results (8.6%) are significantly different from the best value (6.7%).

posterior combination by STD SUM and STD PRODUCT is presented for each of the possible two and three stream combinations. These results show that in the case of matched (clean) speech (**Table H.1**), feature and probability combination yield no significantly different results. The use of several additional feature or probability streams can, in this case, not improve significantly over the already good results of the J-RASTA-PLP feature based recognizer.

10.2.2 Experiments on speech with narrow-band noise

The next set of experiments is carried out on artificial (stationary) narrow-band noise. The results in **Table 10.2** show no significant improvement of the FC multi-stream system over the baseline system (which is the J-RASTA-PLP feature based recognizer in the first line), comparing the mean values in the last column. Standard multi-stream processing using the STD SUM results in the same average performance, whereas the STD PRODUCT significantly degrades performance. Employing RF weights in the FC SUM could not lead to any improvement on this kind of noise, just as it was the case in multi-band processing (cf. **Table 8.12**). (No tests using FC PRODUCT are carried out as no improvement is to be expected on this kind of noise (due to results presented in **Table 8.9**)). In the case of artificial *non*-stationary band-limited noise

	Stationary Band-Limited Noise								
	Band 1		Band 2		Band 3		Band 4		Mean
	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	
J-RASTA-PLP	31.4	14.0	44.6	16.6	35.0	18.9	23.9	17.4	25.2*
PLP	57.5	29.2	74.6	34.1	65.4	31.2	67.2	32.5	49.0
MFCC	57.5	28.7	50.6	29.5	43.8	23.4	29.1	15.4	34.8
PLP-J-RASTA-PLP	48.6	24.0	58.5	25.5	56.6	27.4	43.8	26.1	38.8
PLP-MFCC	65.2	34.8	59.9	29.0	42.8	23.2	50.9	22.5	41.0
MFCC-J-RASTA-PLP	52.0	27.2	36.9	18.6	28.4	16.0	18.4	12.9	26.3*
PLP-J-RASTA-PLP-MFCC	53.4	27.4	46.2	21.2	41.1	19.6	22.6	16.9	31.1
STD SUM	35.2	16.1	41.6	16.4	34.3	17.7	23.1	16.2	25.1
STD PRODUCT	41.8	18.4	49.6	23.5	34.2	17.5	24.7	15.8	28.2
FC SUM EQUAL	39.5	19.2	41.3	19.5	28.9	17.1	20.3	14.6	25.1
FC SUM RF	41.0	19.2	42.0	18.9	31.3	17.4	21.0	14.6	25.7*

Table 10.2: WERs for multi-stream processing in stationary band-limited noise, using PLP, J-RASTA-PLP and MFCC features. FC processing (using FC SUM) is compared to standard multi-stream processing where only the three single-feature-based probability streams are recombined as well as to the J-RASTA-PLP baseline. Results for each of the seven constituent recognizers which are used in FC processing are also given (Lines 1 to 7). RF refers to relative frequency weighting. * indicates that there is no significant difference to the best result in this column.

(cf. **Table 10.3**), on the other hand, the FC multi-stream system significantly outperforms the baseline recognizer due to the severe degradation of the J-RASTA-PLP features in this kind of noise as had also been found in multi-band processing (cf. **Table 8.7**). Unfortunately, contrary to multi-band processing, where the use of RF weights significantly improved performance on non-stationary band-limited noise (cf. **Table 8.12**), no robustness gain could be achieved with these weights in FC multi-stream processing in this noise. The standard multi-stream approaches (STD SUM and STD PRODUCT) still outperform the baseline system but are not competitive as compared to FC processing.

	Siren		
	0 dB	12 dB	Mean
J-RASTA-PLP	89.4	36.9	63.2
PLP	66.9	36.1	51.5
MFCC	57.4	34.5	46.0*
PLP-J-RASTA-PLP	68.2	32.9	50.6
PLP-MFCC	61.6	29.9	45.8*
MFCC-J-RASTA-PLP	74.6	30.2	52.4
PLP-J-RASTA-PLP-MFCC	67.0	28.9	48.0
STD SUM	67.9	30.4	49.2
STD PRODUCT	66.9	30.3	48.6
FC SUM EQUAL	60.3	25.9	43.1
FC SUM RF	60.8	26.5	43.7*

Table 10.3: WERs for multi-stream processing in non-stationary band-limited noise, using PLP, J-RASTA-PLP and MFCC features. FC processing (using FC SUM with EQUAL and RF weights) is compared to standard multi-stream processing where only the three single-feature-based probability streams are recombined as well as to the J-RASTA-PLP baseline. Results for each of the seven constituent recognizers which are used in FC processing are also given (Lines 1 to 7). * indicates that there is no significant difference to the best result in this column.

10.2.3 Experiments on speech with real-environmental noise

The main goal is to test the multi-stream system on the real-environmental noises where the multi-band FC systems could not achieve any significant performance improvement. We thus turn to the wide-band test environments of car and factory noise which correspond to the noise cases described for the multi-band systems in Section 8.2. Results are presented in **Table 10.4**.

The baseline recognizer employing the robust J-RASTA-PLP features outperforms the standard multi-stream systems as well as the FC approaches. RF weighting used in FC SUM did not achieve a significant improvement. As FC PRODUCT obtained lower WER than the FC SUM in multi-band processing in wide-band noise, we employ FC PRODUCT also to multi-stream processing in this kind of noise. Significant performance improvement as compared to FC SUM is achieved but not resulting in competitive performance as compared to the baseline system.

	Car		Factory		Mean
	0 dB	12 dB	0 dB	12 dB	
J-RASTA-PLP	32.8	10.6	34.6	11.4	22.4
PLP	50.5	13.8	52.6	14.6	32.9
MFCC	63.4	22.6	68.8	21.2	44.0
PLP-J-RASTA-PLP	50.1	13.6	48.0	14.4	31.5
PLP-MFCC	50.9	14.0	54.1	14.5	33.4
MFCC-J-RASTA-PLP	48.4	14.1	47.4	14.6	31.1
PLP-J-RASTA-PLP-MFCC	48.1	13.4	49.8	13.5	31.2
STD SUM	42.4	11.4	42.0	12.3	27.0
STD PRODUCT	40.6	12.1	40.8	11.9	26.4
FC SUM	47.5	12.6	46.9	14.0	30.3
FC SUM RF	46.9	12.7	47.0	13.0	29.9
FC PRODUCT	41.3	11.7	39.4	13.8	26.6

Table 10.4: WERs for multi-stream processing using PLP, J-RASTA-PLP and MFCC features. FC processing (using FC SUM with EQUAL and RF weights, and FC PRODUCT) is compared to standard multi-stream processing where only the three fullband probability streams are recombined as well as to the J-RASTA-PLP baseline. Results for each of the seven constituent recognizers which are used in FC processing are also given (Lines 1 to 7).

10.2.4 Preliminary conclusions

To sum up, significant performance improvement as compared to the J-RASTA-PLP baseline system is achieved by FC processing only on non-stationary band-limited noise. In this case, also the standard multi-band approaches deteriorate significantly more than FC SUM. For all other noise conditions and clean speech, no significant improvement is obtained using FC multi-stream processing as compared to the J-RASTA-PLP fullband recognizer. In real-environmental wide-band noise, FC PRODUCT is needed to improve FC processing on this kind of noise.

In a parallel study by Christensen et al. (2000) multi-stream experiments using PLP, MFCC and J-RASTA-PLP features on noise-corrupted data (from the NUMBERS95 database with added factory, car and lynx noise from the NOISEX92 database) also gave no significant performance improvement. Probability combination was carried out by STD SUM and STD PRODUCT. Investigation of the respective frame errors, however, showed that there was a significant performance difference in the ability of the streams to classify the different phonemes. It therefore has to be concluded that performance improvement in noisy speech is not guaranteed through the use of any heterogeneous feature sets, even when the features show different abilities in classifying the different phonemes on the frame level.

More diverse information streams are needed to improve over a single, good classifier such as in our case the J-RASTA-PLP feature based classifier. In the following, we therefore test the full potential of the FC multi-stream system using streams covering more diverse speech information, using acoustic data from *different time scales*.

10.3 Multiple time scale features

In the next set of multi-stream experiments we change from diverse features which were extracted from the same time scale to the same kind of features (i.e. either PLP or J-RASTA-PLP features) which, however, are extracted from multiple time scales. Two kinds of features are investigated: (i) *"variable window size" features* where the window for feature extraction is increased for each new feature stream, and (ii) *static and difference features* employed in independent streams and processed according to the FC approach.

In the multi-stream experiments presented in the previous section as well as in our multi-band experiments described in Section 8.4, equal weighting generally gave among the best results when using HMM/MLP hybrid systems. For this reason, we use equal weighting also in the following multi-stream experiments.

10.3.1 Variable window size features

Three different time scales are used in the estimation of the "variable window size" PLP features: 25 ms, 75 ms and 125 ms, respectively (all shifted every 12.5 ms), as was illustrated in **Figure 9.4**. The first time-scale PLP features correspond to the PLP features also employed in the multi-stream system in the previous section, as well as in the PLP-feature based baseline system in multi-band processing (see Section 8.1 for description). The larger time scale features use the same number of coefficients as described in **Table 8.1** for the short-time features, which was found to work best. In **Tables 10.5** and **10.6**, the multi-scale features are respectively denoted as (1), (3) and (5), referring to the respective multiple of the original window size.

Since these multi-scale features are clearly correlated, and not very likely to function well by themselves, they are respectively concatenated to the instantaneous feature vector (1), resulting in feature vectors (1-3), (1-5) and (1-3-5). For each feature combination, one MLP expert was trained on 9 frames of contextual input. The number of MLP parameters are given in **Table J.2** of Appendix J.

	Stationary Band-Limited Noise								
	Band 1		Band 2		Band 3		Band 4		Mean
	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	
1	57.5	29.2	74.6	34.1	65.4	31.2	67.2	32.5	49.0
1-3	68.2	37.4	66.0	30.8	58.1	29.9	73.4	35.6	49.9∗
1-5	63.0	37.8	71.5	33.4	60.9	33.2	83.9	44.4	53.5
1-3-5	65.5	37.5	79.4	35.6	59.8	33.5	84.2	42.4	54.7
STD Sum	63.6	34.9	75.1	33.6	57.1	30.0	71.5	33.4	49.9∗
STD Product	65.9	35.7	73.6	33.7	56.7	29.3	74.6	35.2	50.6∗

Table 10.5: WERs for the "variable window size" system tested in stationary band-limited noise. Features extracted on 25 ms are refered to as (1), on 75 ms as (3) and on 125 ms as (5). Longer time scale features (3) and (5) are concatenated to the short-term features (1). Multi-stream combination is carried out by STD Sum and STD Product. ∗ indicates that there is no significant improvement in WER as compared to the best value in that column.

Results for the "variable window size" system using PLP features can be seen in **Tables 10.5** and **10.6**. The baseline system – referred to as (1) – corresponds to the PLP fullband HMM/MLP hybrid system as also used in multi-band processing on PLP features (and in the multi-stream experiments of the preceding section). The multiple-time scale system consists of 4 MLPs, each of which was trained on one of the different time scales (1), (1-3), (1-5), and (1-3-5). The posterior probabilities at the output of the MLPs are then combined via sum and product.

The results in clean speech are presented in the last column of **Table 10.6**. It can be observed that in the longer time scale based recognizers the information germane to phonetic identity is most probably smeared across time and thus results in less accurate recognition. No improvement was obtained by posterior combination of the multi-scale streams.

In stationary band-limited noise (cf. **Table 10.5**), no performance improvement as compared to the baseline system could be achieved when using the recombined multi-stream system incorporating all three time scales (for both the STD Sum and STD Product). Degradation, though, is not significant either.

The results on non-stationary band-limited noise and wide-band noise are shown in **Table 10.6**. Some, though no consistent and no significant improvement using the multiple time scale systems as compared to the baseline fullband system is obtained. As already observed previously, the product combination degrades more than the sum combination in non-stationary band-limited noise but performs better than the latter in the case of wide-band noise.

	Band-Limited Noise			Wide-Band Noise					Clean
	Siren			Car		Factory		Mean	
	0 dB	12 dB	Mean	0 dB	12 dB	0 dB	12 dB		
1	66.9	36.1	51.5°	50.5	13.8	52.6	14.6	32.9*	7.1
1-3	68.8	34.5	51.7°	49.1	16.1	55.6	15.6	34.1	7.9◇
1-5	68.1	33.9	51.0°	50.8	16.1	53.6	17.0	34.4	8.6◇
1-3-5	77.6	37.8	57.7	48.9	15.9	54.8	16.9	34.1	9.4
STD SUM	65.8	34.5	50.2	47.5	14.6	49.8	15.5	31.9*	7.8◇
STD PRODUCT	69.7	34.8	52.3°	45.5	14.2	49.3	14.7	30.9	8.1◇

Table 10.6: WERs for the "variable window size" systems tested in non-stationary band-limited (siren) noise and wide-band car and factory noise, as well as clean speech. Features extracted on 25 ms are refered to as (1), on 75 ms as (3) and on 125 ms as (5). Longer time scale features (3) and (5) are concatenated to the short-term features (1). Multi-stream combination is carried out by STD SUM and STD PRODUCT. $*$, $\diamond$ and $\circ$ indicate that there is no significant improvement in WER as compared to the best value in that column.

As no significant performance improvement was achieved, no tests on J-RASTA-PLP features were pursued.

10.3.2 Static and difference features

The first and second order time *difference features*, were implicitly used in (almost) all the previous systems as described in Section 8.3.1. The delta features are calculated over five instantaneous features according to $\Delta\, x_t = [-2x_{t-2} - x_{t-1} + x_{t+1} + 2x_{t+2}]$ and the delta-delta features over seven instantaneous features according to $\Delta\Delta\, x_t = [2x_{t-3} + x_{t-2} - 2x_{t-1} - 2x_t - 2x_{t+1} + x_{t+2} + 2x_{t+3}]$ (which corresponds to a simple subtraction of the preceding delta-value from the following delta-value, i.e. $\Delta\Delta\, x_t = [-\Delta\, x_{t-1} + \Delta\, x_{t+1}]$). They cover respectively 75 ms and 100 ms of speech data. As these difference features are more or less independent of the instantaneous features they can more appropriately be treated as separate, higher time scale feature streams. In the following, we use these features (in PLP and J-RASTA-PLP processing) as multiple time scale feature streams, which are recombined according to the FC approach. For each of the seven streams an MLP recognizer was trained, the number of parameters of which are given in **Table J.3** of Appendix J. The posterior probabilities at the output of the MLPs are combined via FC SUM (6.3), FC INDEP ASMPT (as defined in **Table D.2** of Appendix D) and FC PRODUCT (6.23), using equal weights.

Experiments were carried out for both PLP and J-RASTA-PLP features.

PLP features

The results of the PLP baseline system (here referred to as RAW-D-DD[2]), which is the same as in the preceding sections, are shown in line 7 of **Tables 10.7** to **10.9**. Lines 1 to 6 give the recognition performance of each of the other six constituent streams for comparison.

In clean speech (see **Table 10.9**), none of the three combination schemes performs significantly different from the baseline recognizer. When looking at the performance of the constituent streams, we see that all combinations including the static (RAW) features perform equally well as the baseline. Only the single streams and the combination of DELTA and DELTA-DELTA features deteriorate when used by themselves. Similar results were found in (Macho et al., 1999, p. 113) on isolated digit recognition where the "acceleration filter" (i.e. the DELTA-DELTA features) also hurt performance in clean speech. The authors argue that degradation in clean speech of the DELTA-DELTA features could be due to the complete cancellation of the zeroth modulation frequency. Thus, static features are needed to achieve good performance in clean speech.

In stationary band-limited noise, FC SUM and FC PRODUCT perform significantly better than all other streams, including the baseline system, as can be seen in **Table 10.7**. The FC INDEP ASMPT results in a weak (though significant) improvement over the baseline. It is interesting to note that, in these noise conditions, the DELTA and DELTA-DELTA streams as well as their combination outperform any other constituent stream.

In non-stationary band-limited noise (see **Table 10.8**), on the contrary, these three streams deteriorate the most, whereas all streams which include the static features are more noise robust. This performance pattern shows the similarity of the difference features to the RASTA-filtered features which, as could be seen in **Tables 10.2** and **10.3** (line 1), led to the same performance pattern on these two narrow-band noise cases. In non-stationary band-limited noise, no significant performance difference is observed between the RAW-D-DD baseline and the FC SUM, though the FC PRODUCT and FC INDEP ASMPT deteriorate significantly, with the latter being unable to decode in case of very low SNR.

The experiments on real-environmental wide-band noise (cf. **Table 10.9**) led to very different results than observed for the other noise cases. Here, it is FC INDEP ASMPT which outperforms all other streams, including the baseline system and the other two FC strategies. Neither FC SUM nor FC PRODUCT can improve over the RAW-D-DD baseline system.

These results show an important improvement from the first set of multiple time scale features. Using the static and difference features in FC, good improvement was achieved in the two band-limited noise cases (with FC SUM) and the wide-band noise cases (using FC INDEP ASMPT) as compared to the baseline system. The use of the difference features is especially appealing as these features, firstly, are more independent than the "variable window size" features and, secondly, only need to be extracted once which speeds up calculation time.

As compared to the results of the multi-band systems of Section 8.3.1 (**Table 8.4**), no improvement was achieved on band-limited noise. Both the multi-band FC SUM and the multi-

[2]This refers to the concatenated features used at the input to the stream MLP : "RAW" denotes the static features, "D" the delta features, and "DD" the delta-delta features.

	Stationary Band-Limited Noise								Mean
	Band 1		Band 2		Band 3		Band 4		
	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	
RAW	63.9	42.0	71.1	36.9	65.1	38.0	81.8	43.6	55.3
DELTA (D)	54.6	33.5	67.9	34.5	56.5	33.9	48.0	38.2	45.9
DDELTA (DD)	62.0	37.1	65.4	29.9	55.9	36.5	44.8	35.6	45.9
RAW-DELTA	72.4	38.6	69.8	34.9	56.0	28.5	72.8	32.5	50.7
RAW-DD	60.0	32.6	66.1	33.0	62.5	32.1	78.2	34.9	49.9
D-DD	62.1	37.8	66.4	31.8	56.4	36.5	45.0	35.1	46.4
RAW-D-DD	57.5	29.2	74.6	34.1	65.4	31.2	67.2	32.5	49.0
FC SUM	49.1	26.8	57.8	26.1	41.8	24.1	39.1	23.5	36.0
FC INDEP ASMPT	78.5	41.5	54.0	23.6	43.8	25.6	58.9	30.8	44.6
FC PRODUCT	53.1	29.8	59.9	27.8	46.4	26.4	45.9	23.4	39.1*

Table 10.7: WERs on stationary band-limited noise for each of the static and difference (PLP) feature streams (pure or concatenated "-"), as well as the FC multi-stream systems (combination by FC SUM, FC INDEP ASMPT and FC PRODUCT), using equal weights. * indicates that there is no significant difference to the best result in that column.

	Siren		
	0 dB	12 dB	Mean
RAW	65.4	37.4	51.4*
DELTA	101.5	57.2	79.3
DDELTA	95.5	50.9	73.2
RAW-DELTA	67.9	34.6	51.3
RAW-DD	68.8	34.2	51.5*
D-DD	91.5	49.6	70.6
RAW-D-DD	66.9	36.1	51.5*
FC SUM	67.8	32.0	49.9
FC INDEP ASMPT	112.8	53.6	83.2
FC PRODUCT	76.0	36.4	56.2

Table 10.8: WERs on non-stationary band-limited noise for each of the static and difference (PLP) feature streams (pure or concatenated "-"), as well as the FC multi-stream systems (combination by FC SUM, FC INDEP ASMPT and FC PRODUCT), using equal weights. * indicates that there is no significant difference to the best result in that column.

	Clean	Wide-Band Noise				
		Car		Factory		Mean
		0 dB	12 dB	0 dB	12 dB	
RAW	9.6	55.0	18.0	67.6	18.4	39.8
DELTA	10.4	81.4	25.8	83.4	24.6	53.8
DDELTA	12.4	89.0	31.1	91.1	32.6	61.9
RAW-D	7.9*	49.1	13.4	52.0	15.9	32.6
RAW-DD	7.0*	54.0	14.5	58.6	14.2	35.3
D-DD	12.4	89.6	32.8	92.1	34.0	62.1
RAW-D-DD	7.1*	50.5	13.8	52.6	14.6	32.9
FC SUM	7.2*	68.8	15.5	68.5	16.2	42.3
FC INDEP ASMPT	6.9	34.8	11.8	49.9	12.2	27.2
FC PRODUCT	7.0*	61.8	13.8	59.2	16.2	37.8

Table 10.9: WERs on clean speech and wide-band noise for each of the static and difference (PLP) feature streams (pure or concatenated "-"), as well as the FC multi-stream systems (combination by FC SUM, FC INDEP ASMPT and FC PRODUCT), using equal weights. * indicates that there is no significant difference to the best result in that column.

band AFC SUM obtained lower WERs on stationary (26.6% WER and 21.2% WER, respectively) and non-stationary (27.4% and 20.7% WERs, respectively) narrow-band noise. In wide-band noise, though, the multiple time scale FC INDEP ASMPT achieved significantly higher accuracy than any of the multi-band systems (cf. Appendix G, **Table G.3**).

J-RASTA Features

With the good results of this multiple time scale FC system using PLP features we now turn to using J-RASTA-PLP features. As we have already observed in the experiments of the preceding systems, J-RASTA-PLP-based systems are harder to improve as their overall performance is already higher.

In this section, we only report the results from the FC strategies, and compare their performance to the J-RASTA-PLP fullband baseline system. All systems using J-RASTA-PLP features employ scaled likelihoods in these experiments. Results for the different noise cases and clean speech can be seen in **Tables 10.10** and **10.11**. The results of the constituent streams, which are discussed in the next paragraph, can be found in Appendix I.

In clean speech, only the DELTA and DELTA-DELTA feature streams (used by themselves) led to significantly degradation for the J-RASTA-PLP features, whereas in the case of PLP features, also the RAW-stream by itself was not competitive (see Appendix I). In stationary band-limited noise, we can again observe that the RAW features by themselves degrade most. For the J-RASTA-PLP features, this degradation, though, is less than for PLP features. In case of non-stationary narrow-band noise (cf. **Tables I.3**), the first derivative features give worst performance for both feature sets, whereas the second-order derivatives degrade significantly less. Combinations with this feature stream even led to the most robust streams in the case of J-RASTA-PLP processing, whereas for PLP features it was the combinations with the RAW features which stayed the most powerful in this kind of noise. In wide-band noise, on the other hand, it is the DELTA-DELTA feature stream which cannot handle this noise corruption. In this kind of noise, the RAW features are needed in both feature sets to enhance performance of the constituent streams. This shows that each of the three feature streams (RAW, DELTA and DELTA-DELTA) is powerful on a different kind of noise condition. We try to better exploit this characteristic by applying FC processing instead of the usually used simple concatenation of the three feature sets.

Turning to the FC strategies, we see that in clean speech (see **Table 10.11**) the baseline system and the FC systems give competitive performance also for J-RASTA-PLP features.

Using static and difference J-RASTA-PLP features in multiple time scale FC, only slightly decrease WER over the baseline system in stationary band-limited noise, but not significantly. In non-stationary band-limited noise, on the contrary, performance was considerably improved by all multiple time scale FC systems and especially in the case of FC SUM.

	Stationary Band-Limited Noise								Mean
	Band 1		Band 2		Band 3		Band 4		
	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	
FULLBAND	30.6	11.4	48.0	16.0	35.2	18.4	24.5	19.2	25.4*
FC SUM	28.9	12.5	40.4	15.0	27.8	15.5	24.4	19.1	23.0*
FC INDEP ASMPT	28.4	11.9	41.5	14.8	29.1	15.8	23.0	17.6	22.8
FC PRODUCT	28.7	12.6	42.9	15.5	28.9	15.7	23.5	18.0	23.2*

Table 10.10: WERs of the FC systems using RAW, DELTA and DELTA-DELTA (J-RASTA-PLP) features as different time scale streams and equal weighting, as well as the fullband baseline employing the three features after concatenation. Tests carried out in stationary band-limited noise. * indicates that there is no significant difference in WER as compared to best value in that column.

In wide-band noise, recognition performance of the baseline and the FC systems, are comparable with no significant improvement using multiple time scale FC. (In the case of the PLP features, it was the FC INDEP ASMPT which significantly outperformed all other streams in this kind of noise).

Discussion To sum up, in clean speech, FC processing did not result in any significantly different performance as compared to the FULLBAND, for both PLP and J-RASTA-PLP. In stationary band-limited noise FC SUM and FC PRODUCT were significantly better than the FULLBAND when using PLP features, but no difference in performance was observed for the J-RASTA-PLP features. In non-stationary band-limited noise, the J-RASTA-PLP based FC SUM, FC PRODUCT, and FC INDEP ASMPT gave significantly improved performance, whereas for PLP features the

	Clean	Band-Limited			Wide-Band Noise				
		Siren			Car		Factory		Mean
		0 dB	12 dB	Mean	0 dB	12 dB	0 dB	12 dB	
FULLBAND	7.8∗	104.6	48.1	76.4	29.1	9.8	34.1	12.5	21.4⋄
FC SUM	8.9∗	83.8	34.4	59.1	27.8	9.6	31.3	11.7	20.1⋄
FC INDEP ASMPT	7.5	100.0	39.9	70.0	28.1	8.8	30.8	10.1	19.5
FC PRODUCT	8.1∗	100.0	39.9	70.0	29.7	9.7	31.9	11.1	20.6⋄

Table 10.11: WERs of the FC systems using RAW, DELTA and DELTA-DELTA (J-RASTA-PLP) features as different time scale streams and equal weighting, as well as the fullband baseline employing the three features after concatenation. Tests carried out in clean speech, stationary band-limited and wide-band noise. ∗ and ⋄ indicate that there is no significant difference in WER as compared to best value in that column.

improvement with FC SUM was not significant. In real-environmental wide-band noise, the FC INDEP ASMPT using PLP features significantly improved over the PLP fullband, though for J-RASTA-PLP features the improvement was not significant.

As it could be seen, the relative improvement is in general less when employing J-RASTA-PLP features. This can be due to the fact that (J-)RASTA processing is very similar to the calculation of the difference features and that less gain is achieved when both are used jointly. Temporal derivatives are equal to an FIR filter which attenuates slow-changing frequency components (i.e. lower modulation frequencies). RASTA processing uses an IIR filter (i.e. band-pass) which is usually a bit broader than the pass-band used in the estimation of the difference features (Hermansky and Morgan, 1994, p. 586). The frequency response of a RASTA filter is supposed to let the most relevant portions of the speech signal pass[3], whereas the DELTA feature filter has a slightly more selective frequency response emphasizing a smaller range of changes in the frequencies and attenuating the rest. This only small difference between both filtering schemes might be responsible for the smaller gain we achieved with the multiple time scale FC systems using J-RASTA-PLP features as compared to PLP features.

Comparison to multi-band systems As compared to the J-RASTA-PLP FC multi-band system, performance of the J-RASTA-PLP-based multiple time scale FC systems cannot compete on stationary band-limited noise (WER of 22.8% as compared to 17.7% in **Table 8.6**), neither on non-stationary band-limited noise (59.1% average WER as compared to 30.0% in **Table 8.7**). In real-environmental wide-band noise, the multiple time scale FC systems degrade less than the FC multi-band systems though the difference is not significant (19.5% as compared to 22.0% in **Table 8.8**). Moreover, all three FC systems gave no significant improvement over the J-RASTA-PLP fullband system in this kind of noise.

[3] Hermansky and Morgan (1994) refer to experiments conducted by Green (1976) where it was found that human hearing is more sensitive to modulation frequencies around 4 Hz than to lower or higher modulation frequencies.

10.4 Summary

The use of multiple, diverse feature streams consisting of PLP, J-RASTA-PLP and MFCC features, in the framework of FC multi-stream processing achieved significant performance improvement in clean speech over the FC *multi-band* system employing J-RASTA-PLP features only. As compared to the J-RASTA-PLP baseline system, though, the difference was not significant. For the different noise cases, the FC multi-stream system employing single-scale features could only significantly reduce WER in the case of non-stationary band-limited noise (as compared to the baseline), but performance was not competitive with that obtained with the J-RASTA-PLP FC multi-band system on this kind of noise.

In *multiple time scale* multi-stream processing, we first investigated combined features extracted from windows covering different time spans. Using these "variable window size" features did not result in any significant performance improvement, neither degradation, in any conditions of our experiments.

Our best success in using multi-stream processing was achieved by the use of static and difference features as different time scales. In the case of PLP features, FC Sum and FC Product significantly enhanced recognition in stationary band-limited noise, and the FC Indep Asmpt in wide-band noise, as compared to the raw-d-dd baseline recognizer. The same experiments carried out using J-RASTA-PLP features gave smaller improvements due to the already noise-robust J-RASTA-PLP features. However, in non-stationary band-limited noise, the FC Sum significantly reduced the WER as compared to the J-RASTA-PLP baseline system.

Comparing these results to the ones obtained using multi-band processing, we can see a tendency of the multi-band systems to be more competitive in band-limited noise, whereas the multi-stream systems are more competitive in clean speech. In wide-band noise, results are less conclusive.

- In stationary band-limited noise, it was the multi-band FC-ECPC system (17.1%) which achieved lowest WER, tightly followed by FC Sum (17.7% with Equal and 18.3% with RF weights) and FC Indep Asmpt (18.6%) in J-RASTA-PLP multi-band processing.
- In non-stationary band-limited noise, the best system was constituted by the multi-band recognizer employing PLP features and the AFC Sum combination strategy (20.7%). The next best result on this kind of noise is achieved by J-RASTA-PLP-based multi-band FC Sum using RF weights (22.6%).
- Turning to the wide-band noise cases, lowest WER was also obtained by a (J-RASTA-PLP) multi-band system (FC Product (19.3%)), and the next best system constituted a multi-stream system (with J-RASTA-PLP static and difference features) using the FC Indep Asmpt (19.5%).
- In clean speech, it is the multi-stream systems employing diverse (single-scale) features (PLP together with J-RASTA-PLP, and J-RASTA-PLP together with MFCC, see Appendix I) which outperformed (6.3%) all other single or multiple stream systems, though the difference was not significant to the next most competitive set of recognizers. The next-best systems are still all multi-stream recognizers, first the (single-scale) FC Sum

(6.7%) followed by the recognizers (6.9%) employing (i) FC Indep Asmpt on static and difference PLP features, (ii) STD Sum in single-scale features, and (iii) simple feature concatenation of these features (PLP-MFCC and PLP-J-RASTA-PLP-MFCC).

Chapter 11

Conclusion

11.1 General summary

In this thesis, the two paradigms of multi-band and multi-stream processing for robust ASR were investigated. The main goal was to advance the development of multiple stream systems which are trained in clean speech and achieve high performance in both matched and mismatched conditions.

Background information on human speech processing was presented, in order to use this understanding to illustrate several automatic processing schemes, introduced later, which are inspired by some aspects of human speech processing. We presented early psychoacoustic findings by Fletcher who experimented on human hearing of CVC syllables. His experiments on high–and low-pass filtered speech suggested the existence of auditory frequency bands which seemed to be processed rather independently. More extensive research several years later by Steeneken and Houtgast, who also employed band-pass filtered speech, disproved this assumption. It was shown that correlation between neighboring frequency bands exist and that there was high information redundancy in the speech signal which is exploited by human listeners.

After describing these human processing schemes which are highly robust to noise, automatic processing schemes to enhance noise robustness where presented. Many of these, including certain feature extraction techniques and the approaches of MD, multi-band and multi-stream processing, comprise characteristics of HSP.

A multi-band system offers increased noise robustness especially to band-limited noise due to the separate processing of each frequency subband. On the other hand, it had previously been observed in multi-band processing that an additional fullband recognizer was needed with clean speech and wide-band noise to render performance of the multi-band system competitive.

Thus, in order to account for the correlation between neighboring frequency bands and exploit the redundancy in the speech signal to render multi-band processing more powerful also to other than band-limited noises, the "full combination" (FC) approach was introduced. In this approach, all possible combinations of frequency subbands are considered, usually by training one expert for each combination. The implementation of the FC approaches for both the sum and product rules were derived for posterior- and likelihood-based systems. In this framework, an interesting relation of the likelihood-based FC approach to missing data processing was found. Employing the marginalization approach from MD, it would be possible in the (likelihood-based) FC approach to integrate over all possible combinations of feature coefficients without having to train a separate expert for each stream (coefficient). In MD, on the other hand, only one configuration of (reliable and unreliable) feature coefficients is chosen at a time. In order to reduce training effort also for posterior-based systems, an approximation to the FC approach was proposed, where the single-stream experts are used to approximate all combination-experts before recombination is carried out.

Although early multi-band approaches relied on Fletcher's assumption of subband independence, none of these approaches actually came anywhere near to achieving the performance described by his "product of errors" rule which states that the overall recognition is correct, if any subband is correct. As error probabilities are known to be directly related to posterior probabilities, the descriptive "product of errors" rule was implemented as a prescriptive method for expert combination. Surprisingly, this showed competitive performance in noise but degradation in clean speech.

Another posterior-based combination strategy was investigated which was also based on a model of HSP quantifying the influence of contextual information on human recognition performance as presented by Bronkhorst, Bosman and Smoorenburg. We investigated how the application of this approach, which we refered to as "error correction in posteriors combination" (ECPC), could be combined with our FC approach. Although only a preliminary approximation to error correction through prior information was used, improved noise robustness on low SNR wide-band noise could be achieved.

In multiple stream systems noise robustness can furthermore be enhanced through the use of reliability weights. Several fixed and adaptive weighting strategies were investigated. The stationary weights comprised relative frequency (RF) measures and weights calculated from the least mean squared error (LMSE) criterion, which both resulted in only small performance improvements. For HMM-GMM systems, (fixed and adaptive) ML weights were derived through the application of the EM algorithm. Here, significant improvement in WER was achieved for both standard and FC multi-band processing. The second set of adaptive weights was based on SNR estimates in each frequency subband.

Although the FC multi-band approaches, in most of the cases, significantly enhanced recognition performance as compared to the standard multi-band approaches much less improvement was achieved as compared to the fullband recognizer. We thus investigated the paradigm of multi-stream processing, based on different fullband feature streams. The same stream combination strategies as introduced for multi-band processing are also applicable here. The same is true for the weighting strategies (with the exception of the specific realization of the SNR-based weights). Single time scale and multiple time scale features were investigated. The multi-scale features, using static, delta and delta-deltas as separate feature streams, achieved highest improvement in noise robustness as compared to the baseline recognizer. It depended on the respective noise condition whether this was the case for the FC Sum, FC Product, or FC Indep Asmpt rule.

Statistics drawn from our tests could be summarized as follows:

- The product rules (both FC and standard) outperformed the sum rules in 69% of the clean cases, in 76% of the band-limited noise conditions, and 92% of the wide-band noise conditions.
- Probability combination (of PLP, J-RASTA-PLP and MFCC features) outperformed feature combination in 81% of all possible combinations (cf. Appendix H).
- Comparing FC processing to the respective standard combination strategy, it was found that for 83% of the cases it was FC processing which performed better.
- When looking at the performance achieved by PLP versus J-RASTA-PLP features, we can observe that PLP features usually gain lower WER in clean and non-stationary band-limited noise where as J-RASTA-PLP features performed better in stationary band-limited noise and wide-band noise.

11.2 Original contributions

Investigating multi-band processing We investigated the limitations of multi-band processing both regarding its original motivation from psychoacoustics and its effective realization in automatic speech recognizers. The first showed us that more recent psychoacoustic results reveal that humans do not seem to process uncorrelated and distinct frequency subbands but make use of dispersed information across the spectrum. Looking at the multi-band approaches as they were employed previously (one expert per subband) we saw that the formerly applied assumption that the events[1] were exhaustive was not valid, and that important correlation information was disregarded.

Based on the results from both investigations we were able to define a conforming set of mutually exclusive and exhaustive events establishing the "full combination" approach. In this approach, correlation information is preserved as much as possible (in the case of noise corrupted data) by integrating experts trained on all possible positions of clean data. The results showed that the FC processing schemes consistently ranged among the best approaches tested in both matched and mismatched conditions.

Development of new combination schemes In the framework of this thesis, we developed several new probability combination strategies. Where appropriate, these were developed for both posterior- and likelihood-based systems. These comprised, firstly, implementation of the already mentioned FC approach as well as an approximation scheme with which every standard subband system can be extended to an approximated FC setup. For the likelihood-based approach, a combination scheme was developed using the marginalization approach (drawn from MD) which allows to integrate over all possible combinations of reliable data using the full-band pdf only. This offers a significant advantage over MD processing where the reliable data needs to be detected, and, moreover, only *one* MD mask is used for each time frame, instead of integrating over all possibilities.

[1] event that "subband i must contain the best selection of data".

For posterior-based combination, two more strategies were investigated, the application of

- the "product of errors" (PoE) rule as proposed by Fletcher
- Bronkhorst et al.'s model, which includes the error probability of the mis-recognized constituent streams (within each combination) in the respective combination-probability.

The STD PoE and FC PoE are competitive in noise as compared to the standard and FC approaches, respectively, but cannot compete in clean speech. FC-ECPC is competitive for all conditions tested, resulting in significant performance improvement in low SNR wide-band noise when weights were employed.

Investigation of fixed and adaptive weighting strategies In order to further enhance our multiple stream systems, we investigated several fixed and adaptive weighting strategies.

The weighting strategies developed for posterior-based systems only led to small improvements. It therefore has to be stated that in this case, equal weighting offers one of the best and, most of all, computationally fastest weighting schemes.

In the case of likelihood-based processing, the conclusions are very different. Significant performance improvement was achieved using the ML weights (trained offline) in standard and FC processing, with the latter significantly outperforming the former. Results were close to those obtained by the "quasi-optimal" weights. Only preliminary experiments could be carried out employing online ML weights adaptation, though performance improvement as compared to equal weights could already be demonstrated.

Investigating multi-stream processing In multi-stream processing, it is usually unpredictable whether the streams are correlated and should thus be processed jointly or whether separate processing up to the probability level is preferable. Employing the FC approach, we can account for both at the same time.

This could be confirmed in multi-stream processing using diverse (single-scale) features, where FC processing achieved among the best results (averaged over all noise conditions) as compared to either pure feature or pure probability combination.

Looking for more diverse and more complementary feature streams, we investigated multiple time scale features. When the static, delta and delta-delta features were used as separate information streams within the FC approach, significant performance improvement could be achieved for PLP features. (Tests on J-RASTA-PLP features showed less significant improvements). This shows that complementary information can be obtained from different time scales, and that, especially in the case of the static and difference (PLP) features, it can be better exploited through FC multi-stream processing than through pure concatenation of these feature streams as usually done in ASR.

Experimental evaluation From the experiments conducted in this thesis, we can draw the following main conclusions:

- The multi-band FC approach is competitive in clean speech (which is not the case for *standard* multi-band processing) and consistently ranges among the best systems for all noise cases. Depending on the noise case, it was observed that the FC SUM rule generally obtains better results in band-limited noise, and the FC PRODUCT in wide-band noise.

- For performance improvement in *clean* speech, multi-stream processing should be applied, though none of our systems tested gained a significant improvement over the best (i.e. J-RASTA-PLP) baseline.

- For multi-stream processing in noise, the results are less conclusive, but again it was observed that the FC SUM rule obtains better results in band-limited noise, whereas in wide-band noise it is the FC INDEP ASMPT rule.

11.3 Future work

While significant progress has been made, there are many issues in multi-band and multi-stream modeling which require further research. This is the case for some of the weighting strategies. The SNR-based weights offer a potential to further improve noise robustness in non-stationary noise if the reliability and speed of SNR estimation can be increased. Thus, it can be hoped to improve these weights through an improved SNR estimator, such as proposed in (Dupont, 2000).

Moreover, ML weight adaptation needs to be tested in non-stationary band-limited noise as well as real environmental noise conditions, where the fixed ML weights have already led to improved performance. The optimal step size for the adaptation of the weights is hereby a crucial factor, which needs to be investigated.

Although the experiments in this thesis were carried out on a wide range of additive noise conditions as well as clean speech, it is now important to test whether the best set of multi-band and multi-stream systems developed in this thesis can also be applied to a large vocabulary continuous speech recognition task.

Finally, some new ideas which developed during the course of this thesis seem interesting for further development. They are described in the following.

Joint multi-stream and multi-band approach The two approaches of multi-band and multi-stream processing are usually considered as alternative paradigms to robust ASR. However, when looking at their respective advantages and disadvantages it becomes apparent that they should rather be seen as complementary. Multi-band processing is powerful in speech corrupted by band-limited noise. Moreover, through the use of narrow frequency bands, training in white Gaussian noise becomes possible in such a way that it will also account for other noise cases, as in narrow frequency bands each noise condition resembles white noise (Cerisara, 1999a; Dupont, 2000). In multi-stream processing, on the other hand, it was shown that training of one of the multiple classifiers in noise usually leads to decreased performance of the combined system and that better performance is rather obtained when using different feature representations (Shire, 2001). Multi-stream processing appears to possess advantages in clean speech as well as speech corrupted by some wide-band noise. Thus, to account for each system's short-coming, it seems promising to use both approaches jointly.

Combination of a multi-band and a multi-stream system has not yet been proposed and has the potential to achieve both good performance in clean speech and noise robust behavior in all noise conditions, that is band-limited and wide-band noise.

Researching for a hybrid combination strategy Moreover, in both approaches, multi-band and multi-stream, the question arises as to which level ("feature combination" or "probability combination") stream combination should take place, and in what way. The first question has, to some extent, been accounted for through the "full combination" (FC) approach which combines both approaches. In probability combination, however, the best probability combination strategy seems to depend on the test application (train/test mismatch). It has been found in this thesis, that the FC PRODUCT rule, which is a severe rule when errors are present, as a single classifier can inhibit a particular class by outputting a probability which is close to zero, often obtains best results in matched conditions and in speech corrupted with wide-band noise.

The FC SUM rule, on the other hand, is more error tolerant, as inaccurate probabilities from one classifier have a smaller effect on the final result. It achieves best performance in speech corrupted with band-limited noise.

For this reason, we propose that a combined strategy which is capable of either switching from one combination rule to the other or combining both in a hybrid rule should be able to handle all unseen testing conditions, clean speech and speech corrupted by diverse noise conditions. Such a combined rule needs to be researched.

The decision as when to switch from one rule to the other could for example be based on signal to noise ratio estimates in several frequency bands which detect the location and level of the noise.

Appendix **A**

Background to probability theory

In this chapter, we recall some basic knowledge from probability theory based on (Bronstein and Semendjajdew, 1989; Saporta, 1990; Papoulis, 1991).

Random events and probabilities A random experiment is a process, which has several possible outcomes, so that it cannot be known in advance which outcome will occur. Moreover, such an experiment can, in theory, be repeated as often as desired. In this context, the set of possible experimental outcomes, which preclude one another, is defined as the set of elementary mutually exclusive *events*. The total set of elementary events is the certain event E and can be represented as the sum of its n mutually exclusive (random) events A_j $(j = 1..n)$, i.e. $E = A_1 \cup A_2 \cup \ldots \cup A_n$ with $A_i \cap A_{i'} = \emptyset$ for $i \neq i'$.

The union $A_1 \cup A_2 \cup \ldots \cup A_k$ $(k < n)$ is a new event and occurs if at least one of the events $A_1, A_2, \ldots, A_k$ occurs (sum of events). The conjunction $A_1 \cap A_2 \cap \ldots \cap A_k$ $(k < n)$ also constitutes an event and occurs if all events $A_1, A_2, \ldots, A_k$ occur simultaneously. It is called the product of events.

Following Kolmogorow's axioms of probability theory (Saporta, 1990, p. 9), a real number $P(A)$ with $0 \leq P(A) \leq 1$ is assigned to each random event A which is called the *probability* of A. The probability of the certain event E is $P(E) = 1$.

Given a set of mutually exclusive events $A_1, A_2, \ldots, A_k$, the axiom of addition describes the relation

$$P(A_1 \cup A_2 \cup \ldots \cup A_k) = P(A_1) + P(A_2) + \ldots + P(A_k) \tag{A.1}$$

For the probability of a union of events, it holds the following rule:

$$\begin{aligned} P(\cup_i^n A_i) &= \sum_{i=1}^{n} P(A_i) - \sum_{i<j}^{n} P(A_i \cap A_j) + \sum_{i<j<k}^{n} P(A_i \cap A_j \cap A_k) - \\ &\quad \ldots (-1)^{n-1} P(A_1 \cap A_2 \cap \ldots \cap A_n) \end{aligned} \tag{A.2}$$

This is sometimes referred to as Sylvester's rule (Barth et al., 1986, p. 106).

Conditional probability and law of total probability The probability of a random event A usually changes, if it is known that a different random event B has already occurred. The probability of A under the condition that event B (with $P(B) \neq 0$) has already occurred is denoted as $P(A|B)$, the *conditional probability of A under the condition B*. The conditional probability is usually defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{with } P(B) \neq 0 \tag{A.3}$$

Solving this for $P(A \cap B)$ we obtain the multiplication rule of probabilities:

$$P(A \cap B) = P(A|B)P(B) \tag{A.4}$$

For any (random) event B it, moreover, holds $B = (B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_n)$, if $P(\cup_i^n A_i) = 1$, i.e. A_i are exhaustive, and, according to the axiom of addition, follows $P(B) = \sum_{i=1}^n P(B \cap A_i)$, if A_i are mutually exclusive. Using the multiplication rule of probabilities (A.4) we get the law of total probability: $P(B) = \sum_{i=1}^n P(B|A_i)P(A_i)$.

Mutual independence Two events are *mutually independent* if the occurrence of any one of them does not influence the occurrence of the other, i.e. if $P(A|B) = P(A)$. The multiplication rule of (A.4) then writes as

$$P(A \cap B) = P(A)P(B) \tag{A.5}$$

stating that the probability of the conjunction of two independent random events is equal to the product of their probabilities.

Random variable

> *A random variable is a process of assigning a number* $\mathbf{x}$ *to every outcome [of an experiment].* (Papoulis, 1991, p. 66)

The value of the random variable will vary from trial to trial as the experiment is repeated. There are two types of random variable - discrete and continuous. A discrete random variable only assumes a set of values, and these values describe its probability distribution[1]. A continuous random variable, on the other hand, is one which takes a continuous range of possible values. In the application in this thesis, the random variables which model the feature vector components are continuous.

(The following derivations are based on (Papoulis, 1991, p. 66ff)). The *cumulative distribution function* (cdf) of a (continuous) random variable $\mathbf{x}$ is the function

$$F(x) = P\{\mathbf{x} \leq x\} \tag{A.6}$$

defined for every x from $-\infty$ to ∞. A cdf has, among many others, the following properties $\lim_{x \to \infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$.

[1]The probability distribution is also sometimes called probability function or the probability mass function.

The derivative

$$f(x) = \frac{dF(x)}{dx} \tag{A.7}$$

of $F(x)$ is called the *probability density function* (pdf) of the (continuous) random variable $\mathbf{x}$. Integrating (A.7) from ∞ to 0, and using $F(-\infty) = 0$, we obtain

$$F(x) = \int_{-\infty}^{x} f(t)\, dt \tag{A.8}$$

Since $F(\infty) = 1$, the above yields

$$\int_{-\infty}^{\infty} f(x) dx = 1 \tag{A.9}$$

which is an important property of any pdf.

In the following, we use upper-case letters for probabilities and lower-case letters for pdfs. For continuous variables, the conditional probabilities introduced above, become conditional pdfs.

Appendix B

Implementation of the approximation to FC

The calculation of the approximated combination posteriors $P(q_k|x_i)$ of (6.28) for the $\mathcal{B}$ combinations of bands can be implemented efficiently by a recursive procedure which utilizes multiplications of bands which are included in several combinations instead of recalculating the same multiplications for similar combinations each time. In the case of a high number of bands, this procedure reduces considerably the number of calculations which have to be carried out. Instead of $B\times 2^{B-1}-2^B+1$ multiplications per frame and per phoneme, only 2^B-B-1 multiplications are necessary, which is a reduction of about $\frac{B}{2}$ when $B\rightarrow\inf$.

The procedure can be illustrated with the help of a binary tree in which the terms of the multiplication are the branches and each node in the tree constitutes a call to the recursive function. This function accumulates the preceding multiplications and then branches out for the next recursions. The value which is obtained in each leaf of the binary tree corresponds to the multiplications of the values on the branches along the path which was run through up to the respective leaf and constitutes $P(q_k|x_i)$[1]. For the case of $B=3$, this is illustrated in Figure B.1, where $p_i = P(q_k|x_i)\,\forall i \in B$ for simplicity. Although the gain in this example of 3 bands is not significant (4 multiplications instead of 5), in the case of 8 bands there will be only 247 multiplications to carry out instead of 769, and in 65519 instead of 458753 the case of 16 bands.

Illustration of an efficient algorithm for the AFC system to calculate all possible combinations of subbands from the probabilities $b_i = P(q_k|x_i)\,\forall i \in B$ of the one-band experts only. The value in each leave, calculated by multiplying all values on the branches from the root to the respective leave, corresponds – after division by the priors and normalization – to the posterior probability

[1] In the leaves, each combination probability is then divided by the respective power of priors, according to (6.28). Moreover, when a combination probability has been calculated for all phonemes, its values are added to obtain the normalization factor for this combination. The normalized probabilities $\hat{P}(q_k|x_{c_i})$ are then used in (6.3).

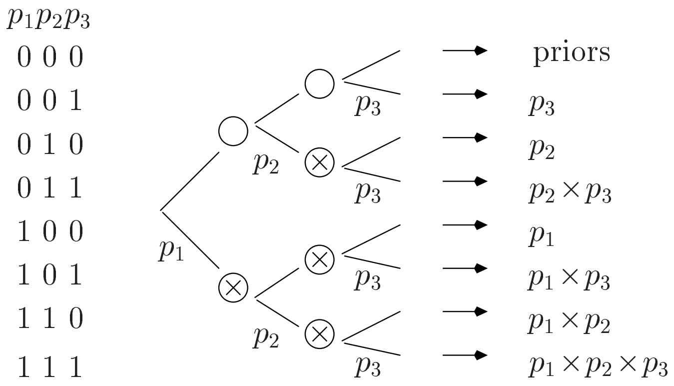

Figure B.1: Illustration to implementation of AFC

$P(q_k|x_{c_i})$ of an approximated combination. One can see that the number of multiplications needed in the binary tree (which are indicated by the symbol $\times$) is smaller (4) than the number of multiplications needed when each combination is estimated explicitly as shown in the column on the right (5). This difference increases proportionally to the number of bands considered.

Appendix C

A model for context effects in human speech recognition

In human perception, the availability of context enhances recognition and renders it more robust to noise. Even if not all phonemes in a word (or words in a sentence etc.) are correctly perceived, humans can fill in missing parts with the help of cues from the surrounding speech parts. This was proven in studies on human speech perception where recognition of words in sentences under noise was shown to outperform recognition of words in isolation or, even more drastically, of nonsense syllables under noise.

Such a model for quantifying the influence of contextual information on human recognition performance was recently proposed and is presented here. Its concept was shown to be applicable to ASR in Section 6.5.

Investigations on human speech perception are carried out to estimate the recognition probability of a certain stimulus by the human perceiver. In (Boothroyd, 1978), the recognition scores of elements (e.g. phonemes) and of wholes (e.g. words) were calculated and reported to follow the relationship[1]

$$p_h = p_e^i$$

with p_h (p_e) being the recognition probability of a whole (element) and i the number of independent elements in a whole, $1 \leq i \leq n$ (n total number of elements in a whole). This relation resembles Fletcher's premise of syllable "articulation" as the product of phone articulation (2.3), which is discussed in Section 2.3. Fletcher's approach though did not regard contextual information as it was only established for nonsense syllables. Boothroyd (1978) included *contextual information* in the model by assuming statistical independence of sensory information s and contextual information w, and stating the error probability of an element as $(1 - p_e) = (1 - p_s)(1 - p_w)$ where p_s (p_w) describes the probability of correctly identifying the element only using sensory (contextual) information.

A more recent model for quantifying the influence of contextual information on human

[1] This assumes equal recognition probability of all elements in the whole.

recognition performance was proposed by Bronkhorst et al. (1993). Although the authors state that it is a model describing recognition performance rather than a model for the recognition process itself, we see in Chapter 6 how this model can actually be used as a basis for recombining the scores from different recognizers (trained on different frequency subbands which are interpreted as the "elements") using contextual information as a method of error correction to obtain improved recognition of the whole stimulus (i.e. the whole frequency domain).

The model is based on a description of human perception as a two-stage process: A listener first tries to identify the stimulus by using sensory information only. Then, in the second part, he/she corrects the mis-classified[2] parts of the (incompletely perceived) stimulus by the use of contextual information:

1. Measure of sensory information

   The probability of occurrence of a (possibly incorrectly perceived) stimulus S_i is calculated from the recognition probabilities of the n elements in the (possibly incomplete) stimulus, which are denoted by p_i, $i = 1..n$. For correct recognition of an element and, thus, the calculation of p_i, only the information in the element itself is used. With this, the probability of correctly and independently identifying all elements of a stimulus is given by $p_1 p_2 ... p_n$ whereas the probability of identifying e.g. all elements but the last is $p_1 p_2 ... (1-p_n)$. An example can be seen in **Figure C.1**.

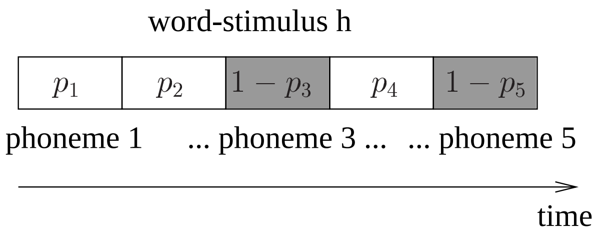

Figure C.1: Example of a word-stimulus consisting of 5 phonemes, 3 of which were correctly identified.

2. Measure of contextual information

   It is assumed that a listener has the chance w_i of correctly guessing, i.e. correcting, one of i mis-classified elements in a stimulus and that corrected elements are indistinguishable from correctly perceived elements. Thus, if a listener correctly perceives $n-i$ elements, he/she has the chance of $w_i w_{i-1} ... w_1$ of correcting the whole stimulus. The parameter w_i therefore provides a measure of the influence of context in the recognition process.

An estimate of the average recognition probability of the whole stimulus h can then be established by multiplying the probability of occurrence S_i of a "percept" with i errors by the chance of correcting this percept, and adding over all possible percepts:

$$p(h) = S_0 + w_1 S_1 + w_1 w_2 S_2 + ... + w_1 ... w_n S_n \tag{C.1}$$

[2] "mis-classified element" here means "element that was incorrectly recognized when using only sensory information".

with

$$
\begin{aligned}
S_0 &= p_1 p_2 ... p_n, \\
S_1 &= (1-p_1)p_2...p_n + ... + p_1...p_{n-1}(1-p_n) \\
&\vdots \\
S_n &= (1-p_1)(1-p_2)...(1-p_n)
\end{aligned}
\tag{C.2}
$$

Each S_i consists of $\binom{n}{i}$ summands which represent all possible permutations of i mis-classified elements in the set of n elements.

There are different possibilities for estimating the context parameters w_i. For example, in word recognition, for each number i of missed phonemes, the alternative words are counted in the lexicon. These numbers together with the total number of words in the dictionary can then be used to calculate each w_i.

Testing this model with different kinds of estimates of w_i, Bronkhorst et al. (1993) showed that recognition scores of consonant-vowel-consonant (CVC) words in auditory or orthographic presentation could be well predicted with this two-stage model.

Appendix D

Summarizing tables of combination strategies

In this chapter, the different combination strategies are summarized in tabular formate for both the posterior-based and the likelihood-based approaches. Table D.1 summarizes the standard combination strategies, Tables D.2 to D.3 give the full combination strategies introduced in this thesis for the posterior-based and likelihood-based case, respectively.

Note that the likelihood-based approach always results from the posterior-based approach from applying Bayes' rule, and vice versa.

COMBINATION STRATEGY	POSTERIOR-BASED	LIKELIHOOD-BASED
STD SUM	$P(q_k\|x) = \sum_{i=1}^{B} P(q_k\|x_i)P(b_i\|x)$	rule 1 $\frac{p(x\|q_k)}{p(x)} = \sum_{i=1}^{B} \frac{p(x_i\|q_k)}{p(x_i)} P(b_i\|x)$
STD SUM	–	rule 2 $\frac{p(x\|q_k)}{p(x)} = \sum_{i=1}^{B} \frac{p(x_i\|q_k)}{p(x_i)} P(b_i\|q_k)$
STD ARITHM MEAN	$P(q_k\|x) = \sum_{i=1}^{B} P(q_k\|x_i)P(b_i)$	$\frac{p(x\|q_k)}{p(x)} = \sum_{i=1}^{B} \frac{p(x_i\|q_k)}{p(x_i)} P(b_i)$
STD PRODUCT	$P(q_k\|x) = \Theta \frac{\prod_{i=1}^{B} P(q_k\|x_i)}{P^{B-1}(q_k)}$	$p(x\|q_k) = \prod_{i=1}^{B} p(x_i\|q_k)$
STD GEOM MEAN	$P(q_k\|x) = \Theta_k \frac{\prod_{i=1}^{B} P^{w_i}(q_k\|x_i)}{P^{(\sum_i w_i)-1}(q_k)}$	$p(x\|q_k) = \prod_{i=1}^{B} p^{w_i}(x_i\|q_k)$
STD PoE	$P(q_k\|x) = 1 - \prod_{i=1}^{B} (1 - P(q_k\|x_i))$	$\frac{p(x\|q_k)}{p(x)} = \frac{1}{P(q_k)} - \prod_{i=1}^{B} (\frac{1}{P(q_k)} - \frac{p(x_i\|q_k)}{p(x_i)})$
STD INDEP ASMPT	$P(q_k\|x) = \Theta_k \prod_{i=1}^{B} P^{w_i}(q_k\|x_i)$	same as STD GEOM MEAN
MINIMUM	$P(q_k\|x) = \Theta \min_{i=1}^{B} P(q_k\|x_i)$	$\frac{p(x\|q_k)}{p(x)} = \Theta \min_{i=1}^{B} \frac{p(x_i\|q_k)}{p(x_i)}$
MAXIMUM	$P(q_k\|x) = \Theta \max_{i=1}^{B} P(q_k\|x_i)$	$\frac{p(x\|q_k)}{p(x)} = \Theta \max_{i=1}^{B} \frac{p(x_i\|q_k)}{p(x_i)}$
MEDIAN	$P(q_k\|x) = \Theta \,\mathrm{med}_{i=1}^{B} P(q_k\|x_i)$	$\frac{p(x\|q_k)}{p(x)} = \Theta \mathrm{med}_{i=1}^{B} \frac{p(x_i\|q_k)}{p(x_i)}$
MINIMUM ENTROPY	$P(q_k\|x) = P(q_k\|x_i)$ $i = \arg\min_{i=1}^{B} - \sum_{k=1}^{K} P(q_k\|x_i) \log P(q_k\|x_i)$	$\frac{p(x\|q_k)}{p(x)} = \frac{p(x_i\|q_k)}{p(x_i)}$ $i = \arg\min_{i=1}^{B} - \sum_{k=1}^{K} P(q_k\|x_i) \log P(q_k\|x_i)$
VOTE	$P(q_k\|x) = \frac{\sum_{i=1}^{B} \Delta_{ki}}{B}$ $\Delta_{ki} = \begin{cases} 1 : \text{if } P(q_k\|x_i) = \max_{k'=1}^{K} P(q_{k'}\|x_i) \\ 0 : \text{otherwise} \end{cases}$	$\frac{p(x\|q_k)P(q_k)}{p(x)} = \frac{\sum_{i=1}^{B} \Delta_{ki}}{B}$ $\Delta_{ki} = \begin{cases} 1 : \text{if } \frac{p(x_i\|q_k)P(q_k)}{p(x_i)} = \max_{k'=1}^{K} \frac{p(x_i\|q_{k'})P(q_{k'})}{p(x_i)} \\ 0 : \text{otherwise} \end{cases}$
RECOMBINING MLP	$P(q_k\|x) = f(\Theta, P(q_{k'}\|x_i), \forall i, k')$	$P(q_k\|x) = f(\Theta, p(x_i\|q_{k'}), \forall i, k')$

Table D.1: Summary of "standard" combination strategies, the first four of which were used in this thesis for comparison to the "full combination" strategies. Θ and Θ_k are normalization constants, such that $\sum_{k=1}^{K} P(q_k|x) = 1$, where factor Θ_k depends on k and Θ does not depend on k. In the literature where the "standard" combination strategies can be found, the normalization constant is often ignored. This might be due to the fact that the normalization constant is sometimes (e.g. in the case of the STD GEOM MEAN rule) hard to calculate.

Combination Strategy	posterior-based
FC Sum	$P(q_k\|x) = \sum_{i=1}^{\mathcal{B}} P(q_k\|x_i)P(b_i\|x)$
FC-ECPC	$P(q_k\|x) = \Theta_k \sum_{i=0}^{\mathcal{B}} P(q_k\|x_i)(1 - P(q_k\|x'_i))\, w^{\|x'_i\|}$
FC Product	$P(q_k\|x) = \Theta \frac{\prod_{i=1}^{\mathcal{B}} P(q_k\|x_i)}{P^{\mathcal{B}-1}(q_k)}$
FC Geom Mean	$P(q_k\|x) = \Theta_k \frac{\prod_{i=1}^{\mathcal{B}} P^{w_i}(q_k\|x_i)}{P^{(\sum_i w_i)-1}(q_k)}$
FC Indep Asmpt	$P(q_k\|x) = \Theta_k \prod_{i=1}^{\mathcal{B}} P^{w_i}(q_k\|x_i)$
AFC Sum	$P(q_k\|x) = \sum_{i=1}^{\mathcal{B}} \hat{P}(q_k\|x_{c_i})P(b_i\|x)$ with $\hat{P}(q_k\|x_i) = \Theta \frac{\prod_{l \in x_i} P(q_k\|x_{(l)})}{P^{\|x_i\|-1}(q_k)}$
FC PoE	$P(q_k\|x) = 1 - \prod_{i=1}^{\mathcal{B}} (1 - P(q_k\|x_i))$

Table D.2: Summary of new combination strategies for posterior-based systems with $\mathcal{B} = 2^B$ stream-combinations for a system of B single-streams. Θ and Θ_k are normalization constants, such that $\sum_{k=1}^{K} P(q_k|x) = 1$, where factor Θ_k depends on k and Θ does not depend on k.

COMBINATION STRATEGY	LIKELIHOOD-BASED
FC SUM rule 1	$\frac{p(x\|q_k)}{p(x)} = \sum_{i=1}^{\mathcal{B}} \frac{p(x_i\|q_k)}{p(x_i)} P(b_i\|x)$
FC SUM rule 2	$\frac{p(x\|q_k)}{p(x)} = \sum_{i=1}^{\mathcal{B}} \frac{p(x_i\|q_k)}{p(x_i)} P(b_i\|q_k)$
FC SUM (MARG)	Same as FC SUM with: $p(x_i\|q_k) = \sum_{j=1}^{M} P(m_j\|q_k) \prod_{l\in s_i} p(x_{(l)}\|m_j, q_k)$
FC SUM (BNDED MARG)	Same as FC SUM with: $p(x_i\|q_k) = \sum_{j=1}^{M} P(m_j\|q_k) \prod_{h\in s_i} p(x_{(h)}\|m_j, q_k) \prod_{l\notin s_i} \frac{1}{x_{(l)}^{obs}} \int_{x_{(l)}=0}^{x_{(l)}=x_{(l)}^{obs}} p(x_{(l)}\|x_m, q_k) dx_{(l)}$
FC PRODUCT	$p(x\|q_k) = \prod_{i=1}^{\mathcal{B}} p(x_i\|q_k)$
FC GEOM MEAN	$p(x\|q_k) = \Theta_k \prod_{i=1}^{\mathcal{B}} p^{w_i}(x_i\|q_k)$
FC INDEP ASMPT	same as FC GEOM MEAN
FC PoE	$\frac{p(x\|q_k)}{p(x)} = \frac{1}{P(q_k)} - P^{\mathcal{B}-1}(q_k) \prod_{i=1}^{\mathcal{B}} (\frac{1}{P(q_k)} - \frac{p(x_i\|q_k)}{p(x_i)})$

Table D.3: Summary of new combination strategies for likelihood-based systems with $\mathcal{B} = 2^B$ stream-combinations for a system of B single-streams. Θ and Θ_k are normalization constants, such that $\int_x p(x|q_k)\, dx = 1$, where factor Θ_k depends on k and Θ does not depend on k.

Appendix E

Definition of full combination subbands

In this table, the parameters are given for both feature extraction and MLP training for each of the subband stream of our multi-band systems.

BAND NUMBER	CRITICAL BANDS	DEFINITION IN Hz	LPC	NUMBER OF CC	HU	MLP PARAM.
1	2-5	115.3-628.5 Hz	3	5	1000	189 000
2	6-9	565.3-1369.9 Hz	3	5	1000	189 000
3	10-12	1262-2292.4 Hz	2	3	666	89 910
4	13-15	2121.7-3768.8 Hz	2	3	666	89 910
12	2-9	115.3-1369.9 Hz	6	10	1485	481 140
13	2-5, 10-12	115.3-628.5 Hz, 1262-2292.4 Hz	5	8	1215	328 050
14	2-5, 13-15	115.3-628.5 Hz, 2121.7-3768.8 Hz	5	8	1215	328 050
23	6-12	565.3-2292.4 Hz	5	8	1215	328 050
24	6-9, 13-15	565.3-1369.9 Hz, 2121.7-3768.8 Hz	5	8	1215	328 050
34	10-15	1262-3768.8 Hz	4	6	945	204 120
123	2-12	115.3-2292.4 Hz	8	12	1700	642 600
124	2-9, 13-15	115.3-1369.9 Hz, 2121.7-3768.8 Hz	8	12	1700	642 600
134	2-5 10-15	115.3-628.5 Hz, 1262-3768.8 Hz	7	11	1620	568 620
234	6-15	565.3-3768.8 Hz	7	11	1620	568 620
FULLBAND	2-15	115.3-3768.8 Hz	11	12	1750	661 500

Table E.1: Definition of the frequency subbands and combination of subbands as employed in our multi-band systems, together with the parameters used in feature extraction and MLP training. The number of parameters are the same for PLP and J-RASTA-PLP features. LPC: LPC analysis order; CC: number of cepstral coefficients; HU: number of hidden units; MLP PARAM.: number of MLP parameters.

Appendix F

Performance of full combination HMM/MLP hybrids

SINGLE-SUBBAND	WER
1	35.6
2	28.4
3	34.8
4	51.1

Table F.1: WERs of the single-subband MLPs on J-RASTA-PLP features tested on the clean Numbers95 test data.

SUBBAND-COMBINATION	WER
12	15.6
A12	24.9
13	16.0
A13	22.5
14	15.9
A14	25.9
23	15.6
A23	28.0
24	15.4
A24	22.0
34	22.1
A34	34.4
123	10.2
A123	30.5
124	10.4
A124	28.6
134	11.2
A134	28.5
234	11.0
A234	35.8
1234	8.0
A1234	45.2

Table F.2: WERs of the subband-combination MLPs on J-RASTA-PLP features tested on the clean Numbers95 test data. The combinations marked with an 'A' denote the approximated combinations as used in the AFC approach.

Appendix G

Comparison of recombination strategies on PLP-features

In this chapter, the experimental results for the fullband and multi-subband systems employing HMM/MLP hybrid systems working on PLP features are summarized.

	Stationary Band-Limited Noise								
	Band 1		Band 2		Band 3		Band 4		Mean
	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	
Fullband	57.5	29.2	74.6	34.1	65.4	31.2	67.2	32.5	49.0
STD Sum	33.1	26.9	51.4	28.9	29.6	23.9	22.8	19.5	29.5
STD Product	42.9	27.8	69.0	34.5	47.2	32.0	29.8	25.2	38.6
STD Indep Asmpt	43.1	27.6	66.2	34.2	47.6	31.1	29.6	25.1	38.1
STD PoE	33.9	26.8	52.8	28.2	40.4	22.5	25.2	20.5	31.3
AFC Sum	31.2	20.6	27.4	17.1	22.9	17.0	17.9	15.8	21.2
FC Sum Equal	36.6	20.2	46.1	26.2	28.8	17.2	21.0	16.8	26.6
FC PoE	39.4	22.0	46.8	23.9	31.8	18.6	24.2	18.4	28.1
FC Geom Mean	45.2	23.4	64.5	31.1	54.1	24.4	46.9	28.0	39.7
FC Indep Asmpt	46.5	22.8	71.1	33.4	60.1	23.5	45.0	27.0	41.2
FC Sum RF	49.5	25.6	46.1	22.4	23.9	16.1	15.9	12.8	26.5
FC Sum LMSE	50.6	27.5	59.6	28.4	52.8	26.4	20.2	16.9	35.3
FC Sum SNR	36.6	20.2	41.4	23.8	23.2	15.5	15.5	12.4	23.6*

Table G.1: WERs of baseline fullband recognizer, standard subband combination strategies, and FC strategies in stationary band-limited noise, employing PLP features. * indicates that there is no significant difference to the best result in this column.

	Siren		
	0 dB	12 dB	Mean
Fullband	66.9	36.1	51.5
STD Sum	30.8	23.6	27.2
STD Geom Mean	44.9	28.8	36.9
STD Indep Asmpt	44.0	28.1	36.1
STD PoE	30.9	21.8	26.4
AFC Sum	24.9	16.4	20.7
FC Sum Equal	34.9	19.9	27.4
FC PoE	37.5	20.4	29.0
FC Geom Mean	53.1	26.9	40.0
FC Indep Asmpt	57.6	26.5	42.1
FC Sum RF	36.4	20.1	28.3
FC Sum LMSE	48.9	26.0	37.5
FC Sum SNR	33.4	20.4	26.9

Table G.2: WERs of baseline fullband recognizer, standard subband combination strategies, and FC strategies in non-stationary band-limited noise, employing PLP features.

	Wide-Band Noise					Clean
	Car		Factory		Mean	
	0 dB	12 dB	0 dB	12 dB		
Fullband	50.5	13.8	52.6	14.6	32.9	7.1
STD Sum	72.4	34.5	70.0	32.1	52.3	14.8
STD Geom Mean	62.8	31.6	60.4	27.8	45.7	13.0
STD Indep Asmpt	63.1	30.5	60.8	27.9	45.6	12.9
STD PoE	73.0	32.1	68.4	30.8	51.1	17.1
AFC Sum	67.2	25.1	70.2	27.8	47.6	10.8
FC Sum Equal	55.0	18.2	57.0	18.5	37.2	7.4*
FC PoE	53.0	17.8	62.2	18.5	37.9	8.4*
FC Geom Mean	55.6	17.9	54.1	17.2	36.2*	7.8*
FC Indep Asmpt	54.1	17.6	52.8	17.0	35.4*	7.2*
FC Sum RF	53.1	17.1	53.2	16.9	35.1*	7.1*
FC Sum LMSE	53.9	16.6	54.1	16.8	35.4*	7.5*
FC Sum SNR	52.0	18.0	54.4	16.6	35.3*	7.4*

Table G.3: WERs of baseline fullband recognizer, standard subband combination strategies, and FC strategies in wide-band (car and factory) noise and clean speech, employing PLP features. * indicates that there is no significant difference to the best result in this column.

For the PLP features employed in HMM/MLP hybrid systems in different conditions the best results were found for

- clean: by the Fullband and all FC strategies.
- stationary band-limited noise: by AFC Sum and FC Sum using SNR weights.
- non-stationary band-limited noise: by AFC Sum.
- wide-band noise: by the Fullband, FC Sum with all non-equal weighting schemes (RF, SNR, and LMSE weights), FC Indep Asmpt, FC Geom Mean.

Thus, also for these features, the FC Sum employing SNR weights is an overall good solution for clean speech and most of the noise cases with the exception of non-stationary band-limited noise.

Appendix **H**

Standard posterior combination of multi-stream HMM/MLP hybrids

Here, the results are presented for the combination of the single-stream (fullband) experts as used in multi-stream processing of Chapter 10.

	Clean
PLP + J-RASTA-PLP	6.8
PLP * J-RASTA-PLP	6.3
PLP + MFCC	7.2
PLP * MFCC	6.8
J-RASTA-PLP + MFCC	7.0
J-RASTA-PLP * MFCC	6.3
PLP + J-RASTA-PLP + MFCC	6.9
PLP * J-RASTA-PLP * MFCC	7.0

Table H.1: Posterior combination of the different feature fullband streams, combination by STD Sum (+) and STD Product (*). There is no significant difference between any of the WERs.

	Stationary Band-Limited Noise								
	Band 1		Band 2		Band 3		Band 4		Mean
	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	
PLP + J-RASTA-PLP	32.3	14.6	41.1	17.1	31.1	17.1	30.6	16.1	25.0
PLP * J-RASTA-PLP	37.8	16.3	53.7	22.2	38.1	18.5	25.4	16.8	28.6
PLP + MFCC	53.8	30.0	58.2	30.0	52.0	24.3	39.9	18.5	38.3
PLP * MFCC	56.1	31.6	62.2	28.6	50.6	23.5	31.5	19.1	37.9
J-RASTA-PLP + MFCC	31.8	15.1	31.8	15.3	23.4	14.7	16.8	12.2	20.2
J-RASTA-PLP * MFCC	33.3	17.0	35.2	19.1	25.6	13.3	16.7	14.2	21.8
PLP + J-RASTA-PLP + MFCC	35.2	16.1	41.6	16.4	34.3	17.7	23.1	16.2	25.1
PLP * J-RASTA-PLP * MFCC	41.8	18.4	49.6	23.5	34.2	17.5	24.7	15.8	28.2

Table H.2: Posterior combination of the different feature fullband streams, combination by STD Sum (+) and STD Product (*).

	Band-Limited Noise			Wide-Band Noise				
	Siren			Car		Factory		Mean
	0 dB	12 dB	Mean	0 dB	12 dB	0 dB	12 dB	
PLP + J-RASTA-PLP	67.6	23.4	45.5	42.9	10.9	42.7	12.0	27.1
PLP * J-RASTA-PLP	72.3	30.1	51.2	37.0	10.4	37.3	11.3	24.0
PLP + MFCC	58.9	31.4	45.2	55.8	16.7	62.7	17.2	38.1
PLP * MFCC	59.7	43.3	51.5	53.4	15.2	58.2	15.8	35.7
J-RASTA-PLP + MFCC	66.7	29.0	47.9	40.5	11.4	39.4	12.0	25.8
J-RASTA-PLP * MFCC	62.4	26.3	44.4	44.7	11.0	42.4	12.7	27.7
PLP + J-RASTA-PLP + MFCC	67.9	30.4	49.2	42.4	11.4	42.0	12.3	27.0
PLP * J-RASTA-PLP * MFCC	66.9	30.3	48.6	40.6	12.1	40.8	11.9	26.4

Table H.3: Posterior combination of the different feature fullband streams, combination by STD Sum (+) and STD Product (*).

Appendix I

Feature combination with J-RASTA-PLP static and difference features

The experiments employing static and first and second order difference features as separate streams in an HMM/MLP hybrid system are presented in Section 10.3. Here, the J-RASTA-PLP-based feature streams and each of their feature concatenations were tested separately (without posterior combination) in clean as well as noise corrupted speech.

	clean
RAW	9.5*
DELTA	10.8
DDELTA	12.8
RAW-DELTA	9.1*
RAW-DDELTA	8.5*
D-DD	9.5*
RAW-D-DD	7.8

Table I.1: WERs of the seven single streams of the multiple time scale FC multi-stream system employing J-RASTA-PLP static (RAW) and first (DELTA) and second (DDELTA) order and difference features in clean speech. * indicates that there is no significant difference to the best result in this column.

	Stationary Band-Limited Noise								
	Band 1		Band 2		Band 3		Band 4		Mean
	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	
RAW	39.5	19.8	53.2	22.6	52.6	25.9	35.4	28.5	34.7
DELTA	36.1	16.4	51.8	20.6	33.4	19.9	32.1	25.5	29.5
DDELTA	39.5	20.8	47.4	19.2	38.2	21.5	32.2	25.5	30.5
RAW-DELTA	35.0	15.0	53.4	20.2	40.1	21.0	29.9	22.8	29.7
RAW-DDELTA	29.6	14.0	48.1	18.0	34.8	18.8	25.1	19.6	26.0*
D-DD	31.5	15.4	44.9	15.9	27.1	17.5	27.4	22.1	25.2
RAW-D-DD	30.6	11.4	48.0	16.0	35.2	18.4	24.5	19.2	25.4*

Table I.2: WERs of the seven single streams of the multiple time scale FC multi-stream system employing J-RASTA-PLP static (RAW) and first (DELTA) and second (DDELTA) order difference features in (stationary) band-limited noise. * indicates that there is no significant difference to the best result in this column.

	Band-Limited Noise			Wide-Band Noise				
	Siren			Car		Factory		Mean
	0 dB	12 dB	Mean	0 dB	12 dB	0 dB	12 dB	
RAW	118.0	59.5	88.8	35.2	13.5	40.8	16.8	26.6
DELTA	123.2	66.9	95.1	34.5	12.8	36.0	14.8	24.5
DDELTA	95.5	51.5	73.5*	45.9	15.1	45.1	17.2	30.8
RAW-DELTA	142.1	51.5	96.8	36.9	12.5	30.1	11.8	22.8*
RAW-DDELTA	95.6	43.6	69.6	37.9	13.0	34.4	11.5	24.2*
D-DD	95.8	47.2	71.5	37.5	11.8	35.4	11.2	24.0*
RAW-D-DD	104.6	48.1	76.4	29.1	9.8	34.1	12.5	21.4

Table I.3:
WERs of the seven single streams of the multiple time scale FC multi-stream system employing J-RASTA-PLP static (RAW) and first (DELTA) and second (DDELTA) order difference features in non-stationary band-limited noise and wide-band car and factory noise. * indicates that there is no significant difference to the best result in this column.

Appendix **J**

Definition of multi-stream fullband recognizers

In these tables, the parameters are given for the MLPs as used in the multi-stream systems. There are 27 output units for each MLP.

FEATURES	NUMBER OF		
	INPUTS	HU	MLP PARAM.
PLP	351	1750	661 500
J-RASTA-PLP	351	1750	661 500
MFCC	351	1750	661 500
PLP-J-RASTA-PLP	702	1850	1 348 650
PLP-MFCC	702	1850	1 348 650
J-RASTA-PLP-MFCC	702	1850	1 348 650
PLP-MFCC-J-RASTA-PLP	1053	2200	2 376 000

Table J.1: Definition of the multi-stream MLPs as used in the experiments on heterogeneous features. INPUTS: number of input units; HU: number of hidden units; MLP PARAM.: number of MLP parameters.

FEATURES	NUMBER OF		
	INPUTS	HU	MLP PARAM.
1	351	1750	661 500
1-3	702	1400	1 020 600
1-5	702	1400	1 020 600
1-3-5	1053	2100	2 268 000

Table J.2: Definition of the multi-stream MLPs as used in the experiments on "variable window size" features as multiple time scale features. INPUTS: number of input units; HU: number of hidden units; MLP PARAM.: number of MLP parameters.

FEATURES	NUMBER OF		
	INPUTS	HU	MLP PARAM.
RAW	117	1000	144 000
DELTA	117	1000	144 000
DDELTA	117	1000	144 000
RAW-DELTA	234	1375	358 875
RAW-DDELTA	234	1375	358 875
D-DD	234	1375	358 875
RAW-D-DD	351	1750	661 500

Table J.3: Definition of the multi-stream MLPs as used in the experiments on static, delta and delta-delta (PLP and J-RASTA-PLP) features as multiple time scale features. INPUTS: number of input units; HU: number of hidden units; MLP PARAM.: number of MLP parameters.

Acronyms

dB	decibel: a logarithmic unit of sound intensity
pdf	probability density function
tiffing	time- and frequency filtering (approach)
r.p.m.	rotations per minute
ACID	Agglomerative Clustering algorithm based on Information Divergence
AFC	Approximation to FC
CVC	Consonant Vowel Consonant
FC	Full Combination
FC-ECPC	Full Combination with Error Correction in Posteriors Combination
AI	Articulation Index
ASR	Automatic Speech Recognition
ANN	Artificial Neural Network
DPCM	Differential Pulse Code Modulation
EBP	Error Back-Propagation (algorithm)
ECPC	Error Correction in Posteriors Combination
EM	Expectation Maximization
FC	Full Combination
FF	Frequency Filtering (features)
FFT	Fast Fourier Transform
FIR	Finite Impulse Response (filter)
GPD	Generalized Probability Descent (algorithm)
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HSP	Human Speech Processing
HSR	Human Speech Recognition
HMM/MLP hybrid	Hidden Markov Model/Multi-Layer Perceptron hybrid
HMM-GMM	Hidden Markov Model employing Gaussian Mixture Models
I(D)FT	Inverse (Discrete) Fourier Transform
IIR	Infinite Impulse Response (filter)
KL	Karhunen-Loeve (transform) = PCA
LDA	Linear Discriminant Analysis
LMSE	Least Mean Squared Error (criterion)

LP	Linear Prediction
LPC	Linear Predictive Coding
MAP	Maximum A Posteriori (criterion)
MCE	Minimum Classification Error (criterion)
MD	Missing Data (approach)
MFCC	Mel-Frequency Cepstral Coefficient
MI	Mutual Information (criterion)
ML	Maximum Likelihood
MLP	Multi-Layer Perceptron
MRE	Minimum Relative Error (criterion)
MSE	Mean Squared Error (criterion)
MSG	Modulation Spectrogram (features)
NLDA	Non-Linear Discriminant Analysis
OGI	Oregon Graduate Institute
PCA	Principal Component Analysis = KL
PLP	Perceptual Linear Prediction
PMC	Parallel Model Combination
RASTA	RelAtive SpecTrAl (filtering)
STI	Speech Transmission Index
SNR	Signal to Noise Ratio
SPL	Sound Pressure Level
TRAPs	TempoRAl Patterns (features)
WER	Word Error Rate

Notation

F(z)	z-transform of a signal
$\log(x)$	$\log_2(x)$, unless otherwise stated
$H(x)$	entropy of random variable x
$I(x,y)$	mutual information between random variables x and y
x_i	the i^{th} component of x
$D(P_1\|\|P_2)$	Kullback-Leibler divergence between probability distribution P_1 and P_2
x_t	(d-dimensional) acoustic vector x at time t
d	dimension of acoustic vector
t	time index
n	iteration index
$X = \{x_1, \dots, x_t, \dots, x_T\}$	acoustic vector sequence of length T
$X_i^t = \{x_i, \dots, x_t\}$	a subsequence of X from vector x_i to vector x_t
q_k	HMM state k
ω_k	a class
K	number of classes, HMM states, or MLP outputs
$\Omega = \{\omega_1, \dots, \omega_K\}$	set of K possible classes
q_k^t	HMM state q_k observed at time t
$Q = \{q^1, \dots, q^t, \dots, q^T\}$	an HMM state sequence of length T
W, W_j	Hidden Markov Model built up by concatenating elementary speech unit HMMs and which is a set of L states $\{q_1, \dots, q_l \dots, q_L\}$
$w_{ij}^{l-1,l}$	MLP weights between unit i (in layer $l-1$) and unit j (in layer l)
$2c+1$	width of the contextual acoustic information at the input of an MLP
$\{w_{ij}(1), .., w_{ij}(n), .., w_{ij}(N)\}$	sequence of weight vectors (connecting unit i to unit j) of length N
E	error function minimized for MLP training
$p(x\|q)$	a likelihood
$P(q\|x)$	a posterior probability
$P(q)$	a prior probability of class (or HMM state) q
$p(X\|W)$	likelihood of X given Markov Model W
$\overline{p}(X\|W)$	Viterbi approximation of the likelihood of X given the Markov Model W
$P(W\|X)$	posterior probability of a Markov Model W given the acoustic

	vector sequence X
$\overline{P}(W\|X)$	Viterbi approximation of the posterior probability of a Markov Model W given the acoustic vector sequence X
$\hat{\Theta}, \Theta$	old and new parameter estimates
μ_k	mean vector associated with ω_k or q_k in case of Gaussian distribution
σ_k	standard deviation associated with ω_k or q_k in case of Gaussian distribution
$P(m_l)$	$P((x$ is from) mixture component $m_l)$
Δx_t	delta feature vector $\Delta\, x$ at time t
$\Delta\Delta x_t$	delta-delta feature vector $\Delta\Delta\, x$ at time t
$n(t)$	noise signal
$s(t)$	speech signal
$h(t)$	impulse response
$x(t) = s(t) + n(t)$	speech signal with additive noise
$y(t)$	signal with additive and convolutive noise
$x_{i,t}$	coefficient i of x at time t
B	number of (frequency) subbands
$\mathcal{B} = 2^B$	number of all possible combinations of subbands
C	set of all possible combinations of B subbands
b_i	event that data in combination i is clean speech data, and data not in combination i is completely uninformative and can therefore be regarded as missing.

Bibliography

Acero, A. (1990). *Acoustical and Environmental Robustness in ASR.* PhD thesis, Carnegie Mellon University.

Acero, A. and Stern, R. (1990). Environmental robustness in automatic speech recognition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 849–852.

Allen, J. (1994). How do humans process and recognize speech? *Transactions on Speech and Audio Processing*, 2(4):567–577.

Antoniou, C. and Reynolds, T. (1999). Using modular/ensemble neural networks for the acoustic modeling of speech. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, Keystone, Colorado.

Antoniou, C. and Reynolds, T. (2000). Acoustic modeling using modular/ensemble combinations of hetergogeneous neural networks. In *Int. Conf. on Spoken Language Processing*, pages 282–285.

Applebaum, D. (1996). *Probability and Information; An integrated approach.* Cambridge University Press.

Arai, A. and Greenberg, S. (1998). Speech intelligibility in the presence of cross–channel spectral asynchrony. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2:929–932.

Auda, G. and Kamel, M. (1998). Modular neural networks: A comparative study. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing.*

Barker, J., Josifovski, L., Cooke, M., and Green, P. (2000). Soft decisions in missing data techniques for robust automatic speech recognition. In *Int. Conf. on Spoken Language Processing*, volume 1, pages 373–376.

Barth, F., und F. Nikol, P. M., and Wörle, K. (1986). *Mathematische Formeln und Definitionen.* J. Lindauer Verlag (Schaefer), München, Germany.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 41.

Berthommier, F. and Glotin, H. (1999a). A measure of speech and pitch reliability from voicing. In Klassner, F., editor, *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI)*, Computational Auditory Scene Analysis (CASA) workshop, pages 61–70, Stockholm.

Berthommier, F. and Glotin, H. (1999b). A new SNR-feature mapping for robust multistream speech recognition. In *Proc. Int. Congress on Phonetic Sciences (ICPhS)*, volume 1, pages 711–715.

Beyerlein, P. (1998). Discriminative model combination. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 481–484.

Bilmes, J. A. (1997). A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden Markov models. Technical report, ICSI-TR-97-021, ICSI, Berkeley, USA.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Blahut, R. (1990). *Principles and Practice of Information Theory*. Addison-Wesley.

Boite, R., Bourlard, H., Dutoir, T., j. Hancq, and Leich, H. (2000). *Traitement de la parole*. Presses polytechniques et universitaire romandes.

Boothroyd, A. (1978). *Auditory Management of Hearing*, chapter Speech Perception and Sensorineural Hearing Loss. M. Ross and T.G. Giolas.

Bourlard, H. (1999). Non-stationary multi-channel (multi-stream) processing towards robust and adaptive ASR. *Proceedings of the ESCA Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pages 1–9.

Bourlard, H. and Dupont, S. (1996). A new ASR approach based on independent processing and recombination of partial frequency bands. *Int. Conf. on Spoken Language Processing*, pages 426–429.

Bourlard, H. and Dupont, S. (1997). Subband-based speech recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, pages 1251–1254.

Bourlard, H., Dupont, S., Hermansky, H., and Morgan, N. (1996a). Towards sub-band based speech recognition. In *Proc. European Signal Processing Conference*, pages 1579–1582, Italy.

Bourlard, H., Dupont, S., and Ris, C. (1996b). Multi-stream speech recognition. IDIAP-RR 07, IDIAP.

Bourlard, H., Hermansky, H., and Morgan, N. (1996c). Towards increasing speech recognition error rates. *Speech Communication*, pages 205–231.

Bourlard, H., Konig, Y., and Morgan, N. (1995). REMAP: Recursive estimation and maximization of a posteriori probabilities (application to transition–based connectionist speech recognition. Technical report, International Computer Science Institute, Berkeley.

Bourlard, H. and Morgan, N. (1994). *Connectionist Speech Recognition. A Hybrid Approach*. Kluwer Academic Publishers, 101 Philip Drive, Assinippi Park, Norwell, Massachusetts 02061 USA.

Bourlard, H. and Morgan, N. (1997). Hybrid HMM/ANN systems for speech recognition: Overview and new research directions. In *Summer School on Neural Networks*, pages 389–417.

Bourlard, H. and Wellekens, C. (1989). Speech pattern discrimination and multi-layered perceptrons. *Computer Speech and Language*, 3:1–19.

Bronkhorst, A., Bosman, A., and Smoorenburg, G. (1993). A model for context effects in speech recognition. *Journal of the Acoustic Society of America*, 93(1):499–509.

Bronstein, I. and Semendjajdew, K. (1989). *Taschenbuch der Mathematik*. Verlag Harri Deutsch, Frankfurt/Main.

Cerisara, C. (1999a). *Contribution de l'approche Multi-Bande à la reconnaissance automatique de la parole*. PhD thesis, Institut National Polytechnique de Lorraine, Nancy, France.

Cerisara, C. (1999b). Dealing with loss of synchronism in multi-band continuous speech recognition systems. In Ponting, K., editor, *Computational Models of Speech Pattern Processing*, pages 90–95. Springer Verlag in cooperation with NATO ASI Series, Berlin.

Cerisara, C., Fohr, D., and Haton, J. (1999a). Robust behavior of multi-band paradigm. *Proceedings of the ESCA Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pages 187–190.

Cerisara, C., Fohr, D., and Haton, J.-P. (2000). Asynchrony in multi-band speech recognition. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2:1121–1124.

Cerisara, C., Haton, J.-P., and Fohr, D. (1999b). Towards a global optimization scheme for multi-band speech recognition. *Proc. European Conf. on Speech Communication and Technology*, 2:578–590.

Cerisara, C., Haton, J.-P., Mari, J.-F., and Fohr, D. (1998). A recombination model for multi-band speech recognition. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 717–720.

Chen, T. and Rao, R. (1998). Audio-visual integration in multimodal communications. *Proceedings of the IEEE*, 86(5):837–852.

Christensen, H., Lindberg, B., and Andersen, O. (2000). Employing heterogeneous information in a multi-stream framework. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, III:1571–1574.

Clemen, R. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 8:167–173.

Cole, R., Noel, M., Lander, T., and Durham, T. (1995). New telephone speech corpora at CSLU. *Proc. European Conf. on Speech Communication and Technology*, 1:821–824.

Cook, G., Christie, J., Ellis, D., Fosler-Lussier, E., Gotoh, Y., Kingsbury, B., Morgan, N., Renals, S., Robinson, A., and Williams, G. (1999). The sprach system for the transcription of broadcast news. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop, Herndon VA*.

Cooke, M., Green, P., Anderson, C., and Abberley, D. (1994a). Recognition of occluded speech by hidden Markov models. Technical Report TR-94-05-01, Sheffield University, Dept of Computer Science, Sheffield, UK.

Cooke, M., Green, P., and Crawford, M. (1994b). Handling missing data in speech recognition. *Int. Conf. on Spoken Language Processing*, pages 1555–1558.

Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable data. *Speech Communication*, 34(3):267–285.

Cooke, M., Morris, A., and Green, P. (1997). Missing data techniques for robust speech recognition. *Int. Conf. on Spoken Language Processing*, pages 863–866.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory.* John Wiley and Sons, New York.

Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-28(4):357–366.

Degan, N. and Prati, C. (1988). Acoustic noise analysis and speech enhancement techniques for mobile radio applications. *Signal Processing*, 15:43–56.

Deller, J., Proakis, J., and Hansen, J. (1987). *Discrete-Time Processing of Speech Signals.* Prentice Hall, New Jersey (USA).

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38.

Duda, R. and Hart, P. (1973). *Pattern Classification and Scene Analysis.* John Wiley & Sons.

Dupont, S. (1997). Using the multi–stream approach for continuous audio–visual speech recognition: Experiments on the M2VTS database. *TCTS, FPMs, Mons*, pages 1–15.

Dupont, S. (2000). *Études et Développement de Nouveaux Paradigmes pour la Reconnaissance Robuste de la Parole.* PhD thesis, Laboratoire TCTS, Université de Mons, Belgium.

Dupont, S. and Luettin, J. (1998). Using the multi-stream approach for continuous audio-visual speech recognition: experiments on the M2VTS database. *Int. Conf. on Spoken Language Processing*, 2:1283–1286.

Dupont, S. and Ris, C. (1999). Assessing local noise level estimation methods. *Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pages 115–118.

Ellis, D., Singh, R., and Sunil, S. (2001). Tandem acoustic modeling in large-vocabulary recognition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 517–520.

Ellis, D. P. (2000). Stream combination before and/or after the acoustic model. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 3:1635–1638.

Fant, G. (1960). *Acoustic Theory of Speech Production.* Mouton, The Hague.

Fiscus, J. (1997). A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (ROVER). In *Proceedings of IEEE ASRU Workshop, Santa Barbara*, pages 347–352.

Fletcher, H. (1953). *Speech and Hearing in Communication.* Krieger, New York.

Fontaine, V., Ris, C., and Boite, J.-M. (1997). Nonlinear discriminant analysis for improved speech recognition. *Proc. European Conf. on Speech Communication and Technology*, 4:2071–2074.

French, N. and Steinberg, J. (1947). Factors governing the intelligibility of speech sounds. *Journal of the Acoustic Society of America*, 19(1):90–119.

Fritsch, J. (1998). ACID/HNN - clustering hierarchies of neural networks for context connectionist acoustic modeling. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 505–508.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition.* Academic Press, Boston, second edition.

Furui, S. (1986a). On the role of spectral transition for speech perception. *Journal of the Acoustic Society of America*, 80(4):1016–1025.

Furui, S. (1986b). Speaker independent isolated word recognition based on emphasized spectral dynamics. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1991–1994.

Gales, M. (1998). Predictive model-based compensation schemes for robust speech recognition. *Speech Communication*, 25(1–3):49–74.

Gales, M. and Young, S. (1992). An improved approach to the hidden Markov model decomposition of speech and noise. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1:233–236.

Gales, M. J. F. (1995). *Model-based techniques for noise robust speech recognition.* PhD thesis, Gonville and Caius Golleage, University of Cambridge, UK.

Gemello, R., Moisa, L., and Laface, P. (2000). Synergy of spectral and perceptual features in multi-source connectionist speech recognition. In *Int. Conf. on Spoken Language Processing*, pages 843–846.

Ghitza, O. (1994). Auditory models and human performance in tasks related to speech coding and speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2(1):115–131.

Glotin, H. (2000). *Élaboration et comparaison de systèmes adaptatifs multi-flux de reconnaissance robuste de la parole: incorporation des indices d'harmonicité et de localisation.* PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France.

Glotin, H. and Berthommier, F. (2000). Test of several external posterior weighting functions fur multiband full combination ASR. In *ICSLP*, volume 1, pages 333–336.

Glotin, H., Berthommier, F., and Tessier, E. (1999). A CASA-labelling model using the localisation cue for robust cocktail-party speech recognition. In *Proc. European Conf. on Speech Communication and Technology (EUROSPEECH)*, volume 5, pages 2351–2354.

Gold, B. and Morgan, N. (2000). *Speech and Audio Signal Processing.* John Wiley and Sons, New York.

Goldberg, R. and Riek, L. (2000). *A Practical Handbook of Speech Coders*. CRC Press, Boca Raton, Florida.

Grant, K. and Braida, L. (1991). Evaluating the articulation index for auditory-visual input. *Journal of the Acoustic Society of America*, 89(6):2952–2960.

Green, G. (1976). *Temporal aspects of audition*. PhD thesis, Oxford University.

Green, K., Kuhl, P., Meltzoff, A., and Stevens, J. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception and Psychophysics*, 50(6):524–536.

Greenberg, S. (1997). Auditory function. In Crocker, M., editor, *Encyclopedia of Acoustics*, pages 1301–1323. John Wiley and Sons, Inc.

Greenberg, S. (1999). Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation. In *ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 47–56, Kekrade, Netherlands.

Greenberg, S. and Kingsbury, B. E. D. (1997). The modulation spectrogram: In pursuit of an invariant representation of speech. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1647–1650.

Haberstadt, A. and Glass, J. (1998). Heterogeneous measurements and multiple classifiers for speech recognition. *Int. Conf. on Spoken Language Processing*, 3:995–998.

Haeb-Umbach, R. and Ney, H. (1992). Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 13–16.

Hagen, A. and Bourlard, H. (2000). Using multiple-time scales in the framework of full combination multi-stream speech recognition. *Int. Conf. on Spoken Language Processing*, 1:349–352.

Hagen, A. and Bourlard, H. (2001). Error correcting posterior combination for robust multi-band speech recognition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 257–260.

Hagen, A., Bourlard, H., and Morris, A. (2001). Adaptive EM-weighting in multi-band recombination of Gaussian Mixture Models. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1:257–260.

Hagen, A. and Glotin, H. (2000). Études comparatives des robustesses au bruit de l'approche 'full combination' et de son approximation. *Journée d'Études sur la Parole, Aussois*, pages 317–320.

Hagen, A., Morris, A., and Bourlard, H. (1998). Subband-based speech recognition in noisy conditions: The full combination approach. IDIAP-RR 15, IDIAP.

Hagen, A., Morris, A., and Bourlard, H. (2000). From multi-band full combination to multi-stream full combination processing in robust ASR. *ISCA ITRW Workshop on Automatic Speech Recognition – Challenges for the new millenium (ASRU2000)*, pages 175–180.

Hansen, L. and Salamon, L. (1990). Neural network ensembles. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 12(10):993–1001.

Hanson, B., Applebaum, T., and Junqua, J.-C. (1996). *Spectral Dynamics for Speech Recognition Under Adverse Conditions*, chapter 14, pages 331–356. Ch.-H. Lee and F.K. Soong and K.K. Paliwal, Norwell, MA, USA.

Hashem, S. (1997). Optimal linear combination of neural networks. *Neural Networks*, 10(4):599–614.

Haton, J.-P., Cerisara, C., and Fohr, D. (1999). Improvement of multi-band speech recognition. *SPECOM*.

Heckmann, M., Berthommier, F., and Kroschel, K. (2001). Optimal weighting of posteriors for audio-visual speech recognition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 161–164.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752.

Hermansky, H., Ellis, D., and Sharma, S. (2000). Tandem connectionist feature extraction for conventional HMM systems. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1635–1638.

Hermansky, H. and Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589.

Hermansky, H., Morgan, N., Bayya, A., and Kohn, P. (1992). RASTA–PLP speech analysis technique. *IEEE Trans. on Signal Processing*, 1:121–124.

Hermansky, H. and Sharma, S. (1999). Temporal patterns (TRAPS) in ASR of noisy speech. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 289–292.

Hermansky, H., Tibrewala, S., and Pavel, M. (1996). Towards ASR on partially corrupted speech. *Int. Conf. on Spoken Language Processing*, pages 462–465.

Hirsch, H.-G. (1993). Estimation of noise spectrum and its application to SNR-estimation and speech enhancement. Technical report, ICISI, Berkeley, California, USA.

Hirsch, H.-G. and Ehrlich, C. (1995). Noise estimation techniques for robust speech recognition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 153–156.

Hirsch, H.-G. and Pearce, D. (2000). The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *ISCA ITRW Workshop on Automatic Speech Recognition – Challenges for the new millenium (ASRU2000)*, pages 181–188.

Houtgast, T. and Steeneken, H. (1985). A review of the mtf concept in room acoustics and its use for speech intelligibility. *Journal of the Acoustic Society of America*, 77:1069–1077.

Houtgast, T. and Verhave, J. (1991). A physical approach to speech quality assessment: Correlation patterns in the speech spectrogram. *Proc. European Conf. on Speech Communication and Technology*, pages 285–288.

Huang, X., Alleva, F., Hon, H.-W., Hwang, M.-Y., Lee, K.-F., and Rosenfeld, R. (1993). The SPHINX-II speech recognition system: an overview. *Computer Speech and Language*, 7(2):137–148.

Hush, D. and Horne, B. (1993). Progress in supervised neural networks. *IEEE Signal Processing Magazine*, pages 8–39.

Jacobs, R. A., Jordan, M. I., Nowland, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3:78–87.

Jancovic, P. and Ming, J. (2001). A multi-band approach based on the probabilistic union model and frequency-filtering features for robust speech recognition. *Proc. European Conf. on Speech Communication and Technology*, pages 579–582.

Janin, A., Ellis, D., and Morgan, N. (1999). Multi-stream speech recognition: Ready for prime time? *Proc. European Conf. on Speech Communication and Technology*, 2:591–594.

Jones, D. (1979). *Elementary Information Theory.* Oxford Applied Mathematics and Computing Science Series, Oxford.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixture of experts and the EM algorithm. *Neural Computation*, 6:181–214.

Juang, B. and Katagiri, S. (1992). Discriminative learning for minimum error classification. *IEEE Trans. on Signal Processing*, 40(12):3043–3054.

Juang, B., Levinson, S. E., and Sondhi, M. M. (1986). Maximum-likelihood estimation for multivariate mixture observations of Markov-chains. *IEEE Transactions on Information Theory*, 32(2):307–309.

Junqua, J.-C. and Haton, J.-P. (1996). *Robustness in Automatic Speech Recognition; Fundamentals and Applications.* Kluwer Academic Publishers, Boston, USA.

Katagiri, S., Lee, C., and Juang, B. (1991). New discriminative training algorithms based on the generalized probabilistic descent method. *Proceedings IEEE Workshop on Neural Networks for Signal Processing*, pages 299–308.

Kirchhoff, K. (1998). Combining articulatory and acoustic information for speech recognition in noisy and reverberation environments. *Int. Conf. on Spoken Language Processing*, pages 891–894.

Kirchhoff, K. and Bilmes, J. (2000). Combination and joint training of acoustic classifiers for speech recognition. *ISCA ITRW Workshop on Automatic Speech Recognition – Challenges for the new millenium (ASRU2000)*, pages 17–23.

Kirchhoff, K., Fink, G., and Sagerer, G. (2000). Conversational speech recognition using acoustic and articulatory input. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 3:1435–1438.

Kittler, J., Hatef, M., Duin, R., and Matas, J. (1998). On combining classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(3):226–239.

Koehler, J., Morgan, N., Hermansky, H., Guenter, H. G., and Tong, G. (1994). Integrating RASTA–PLP into speech recognition. *IEEE Trans. on Signal Processing*, 1:421–424.

Kryter, K. (1962). Methods for the calculation and use of the articulation index. *Journal of the Acoustic Society of America*, 34(11):1689–1697.

Kunt, M. (1996). *Traitement Numérique des Signaux*. 20. Presses Polytechniques et Universitaires Romandes.

Lecomte, I., Lever, M., Boudy, J., and Tassy, A. (1989). Car noise processing for speech input. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 512–515.

Lippmann, R., Martin, E., and Paul, D. (1987). Multi-style for robust isolated-word speech recognition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 705–708.

Lippmann, R. P. (1996). Accurate consonant perception without mid-frequency speech energy. *IEEE Transactions on Speech and Audio Processing*, 4(1):66–69.

Lockwood, P., Baillargeat, C., Gillot, J., Boudy, J., and Faucon, G. (1991). Noise reduction for speech enhancement in cars: non-linear spectral subtraction/Kalman filtering. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 83–86.

Lucey, S., Sridharan, S., and Chandran, V. (2001). An investigation of HMM classifier combination strategies for improved audio-viusal speech recognition. In *Proc. European Conf. on Speech Communication and Technology*, pages 1185–1188.

Macho, D., Nadeu, C., Hernando, J., and Padrell, J. (1999). Time and frequency filtering for speech recognition in real noise conditions. *Proceedings of the ESCA Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pages 111–114.

Mak, B. (1997). Combining ANNs to improve phone recognition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 4, pages 3253–3256.

Mari, J.-F., Haton, J.-P., and Kriouile, A. (1997). Automatic word recognition based on second-order hidden Markov models. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 5, pages 22–25.

Martin, R. (1993). An efficient algorithm to estimate the instantaneous SNR of speech signals. *Proc. European Conf. on Speech Communication and Technology*, pages 1093–1096.

McCourt, P., Vaseghi, S., and Harte, N. (1998). Multi-resolution cepstral features for phoneme recognition across speech sub-bands. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1:557–560.

Meinedo, H. and Neto, J. (2000). Combination of acoustic models in continuous speech recognition hybrid systems. *Int. Conf. on Spoken Language Processing*, 2:931–934.

Miller, G. and Nicely, P. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustic Society of America*, 27:338–352.

Ming, J., Jancovic, P., Hanna, P., Stewart, D., and Smith, F. (2000). Robust feature selection using probabilistic union models. *Int. Conf. on Spoken Language Processing*, pages 546–549.

Ming, J. and Smith, F. (1999). Union: A new approach for combining sub-band observations for noisy speech recognition. *Proceedings of the ESCA Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pages 175–178.

Mirghafori, N. (1999). *A Multi-Band Approach to Automatic Speech Recognition.* PhD thesis, ICSI, Berkely, California.

Mirghafori, N. and Morgan, N. (1998a). Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers. *Int. Conf. on Spoken Language Processing*, 3:743–746.

Mirghafori, N. and Morgan, N. (1998b). Transmission and transitions: A study of two common assumptions in multi–band ASR. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2:713–716.

Mirghafori, N. and Morgan, N. (1999). Sooner or later: Exploring asynchrony in multi-band speech recognition. *Proc. European Conf. on Speech Communication and Technology*, 2:595–598.

Mirghafori, N., Morgan, N., and Bourlard, H. (1994). Parallel training of MLP probability estimators for speech recognition: A gender based approach. In *IEEE Workshop on Neural Networks for Signal Processing*, pages 140–147.

Mokbel, C. (1992). *Reconnaissance de la parole dans le bruit: bruitage/debruitage.* PhD thesis, Ecole Nationale Supérieure des Télécommunications, Paris, France.

Mokbel, C. and Chollet, G. (1991). Word recognition in the car: Speech enhancement/spectral transformations. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2:925–928.

Moore, B. (1997). *An Introduction to the Psychology of Hearing.* Academic Press, San Diego, California.

Moreno, P. and Stern, R. (1994). Sources of degradation of speech recognition in telephone environments. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 109–112.

Morgan, N. and Hermansky, H. (1992). RASTA extensions: Robustness to additive and convolutional noise. *ESCA Proceedings*, pages 115–118.

Morris, A., Barker, J., and Bourlard, H. (2001a). From missing data to maybe useful data: Soft data modeling for noise robust ASR. In *Workshop on Innovation in Speech Processing (WISP) 2001*, pages 153–164, Stratford-upon-Avon, UK.

Morris, A., Cooke, M., and Green, P. (1998). Some solutions to the missing features problem in data classification, with application to noise robust ASR. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 737–740.

Morris, A., Hagen, A., and Bourlard, H. (1999). The full combination sub-bands approach to noise robust HMM/ANN-based ASR. *Proc. European Conf. on Speech Communication and Technology*, 2:599–602.

Morris, A., Hagen, A., Glotin, H., and Bourlard, H. (2001b). Multi-stream adaptive evidence combination to noise robust ASR. *Speech Communication*, 34(1-2):25–40.

Myers, C. S. and Rabiner, L. R. (1981). Connected digit recognition using a level-building DTW algorithm. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-29(3):351.

Nadeu, C. and Juang, B. (1994). Filtering of spectral parameters for speech recognition. In *Int. Conf. on Spoken Language Processing*, pages 1927–1930.

Nadeu, C., Macho, D., and Hernando, J. (2001). Time and frequency filtering of filter-bank energies for robust HMM speech recognition. *Speech Communication*, 34(1–2):93–114.

Nadeu, C., Paches-Leal, P., and Juang, B. (1997). Filtering the time sequences of spectral parameters for speech recognition. *Speech Communication*, 22(4):315–332.

Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., and MaShari, A. (2000). Audio visual speech recognition. Technical report, Johns Hopkins University-CLSP.

Okawa, S., Bocchieri, E., and Potamianos, A. (1998). Multi-band speech recognition in noisy environment. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2:641–644.

Okawa, S., Nakajima, T., and Shirai, K. (1999). A recombination strategy for multi-band speech recognition based on mutual information criterion. *Proc. European Conf. on Speech Communication and Technology*, 2:603–606.

Papoulis, A. (1991). *Probability, Random Variable and Stochastic Processes.* McGraw-Hill Series in Electrical Engineering, New York, USA.

Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing Series. PTR Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632.

Rao, S. and Pearlman, W. A. (1996). Analysis of linear prediction, coding and spectral estimation from subbands. *IEEE Transactions on Information Theory*, 42(4):1160–1178.

Renals, S. and Hochburg, M. (1995). Decoder technology for connectionist large vocabulary speech recognition. Technical report, Cambridge University Engineering Department, Cambridge, UK.

Richard, M. and Lippmann, R. (1991). Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Computation*, pages 461–483.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge, UK.

Rogova, G. (1994). Combining the results of several neural network classifiers. *Neural Networks*, 7(5):777–781.

Rogozan, A. and Deléglise, P. (1998). Adaptive fusion of acoustic and visual sources for automatic speech recognition. *Speech Communication*, 26(1–2):149–161.

Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning internal representations by error propogation. In Rumelhart, D. and McClelland, J., editors, *Parallel Distributed Processing. Exploration of the Microstructure of Cognition. Vol. 1: Foundations*, pages 318–362. MIT Press.

Saporta, G. (1990). *Probabiliés, analyse des données et statistique.* Editions Technip, Paris, France.

Schless, V. and Class, F. (1997). Adaptive model combination for robust speech recognition in car environmnes. *Proc. European Conf. on Speech Communication and Technology*, pages 1091–1094.

Schukat-Talamzzini, E. G. (1995). *Automatische Spracherkennung: Grundlagen, statistische Modelle und effiziente Algorithmen.* Vieweg, Braunschweig, Germany.

Shannon, C. and Weaver, W. (1949). *The Mathematical Theory of Communication.* University of Illinois Press, Urbana.

Sharma, A., Eliis, D., Kajarekar, S., Jain, P., and Hermansky, H. (2000). Feature extraction using non-linear transformation for robust speech recognition on the aurora database. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2:1117–1120.

Shire, M. L. (2000). *Discriminant Training of Front-End and Acoustic Modeling Stages to Heterogeneous Acoustic Envrionments for Multi-Stream Automatic Speech Recognition.* PhD thesis, University of California, Berkeley, USA.

Shire, M. L. (2001). Multi-stream ASR trained with heterogeneous reverberant environments. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 253–256.

Silipo, R., Greenberg, S., and Arai, T. (1999). Speech intelligibility derived from exceedingly sparse spectral information. *Proc. European Conf. on Speech Communication and Technology*, pages 2687–2690.

Steeneken, H. and Houtgast, T. (1980). A physical method for measuring speech transmission quality. *Journal of the Acoustic Society of America*, 67(1):318–326.

Steeneken, H. and Houtgast, T. (1999). Mutual dependence of the ocatve-band weights in predicting speech intelligibility. *Speech Communication*, 28:109–123.

Teissier, P., Robert-Ribes, J., Schwartz, J., and Guerin-Dugu, A. (1999). Comparing models for audiovisual fusion in a noisy vowel recognition task. *IEEE Trans. Speech Audio Processing*, 7(6):629–642.

Tibrewala, S. and Hermansky, H. (1997). Sub-band based recognition of noisy speech. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1255–1258.

Tomlinson, M. J., Russel, M. J., and Brooke, N. (1996). Integrating audio and visual information to provide highly robust speech recognition. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 821–824.

Tumer, K. and Ghosh, J. (1996). Analysis of decision boundaries in linearly combined neural classifiers. *Pattern Recognition*, 29(2):341–348.

Varga, A., Steeneken, H., Tomlinson, M., and Jones, D. (1992). The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical report, DRA Speech Research Unit, Malvern, England.

Varga, A. P. and Moore, R. K. (1990). Hidden Markov model decomposition of speech and noise. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 845–848.

von Bekesy, G. (1960). *Experiments in Hearing*. McGraw-Hill, New York.

Waibel, A., Sawai, H., and Shikano, K. (1989). Modularity and scaling in large phonemic neural networks. *IEEE Transactions on Acoustics, Speech and Signal Processing*, pages 1888–1898.

Weber, K. (2000). Multiple time scale feature combination towards robust speech recognition. *Konvens, 5. Konferenz zur Verarbeitung natürlicher Sprache*, pages 295–299.

Widrow, B., Jr, J., McCool, J., Kaunitz, J., Williams, C., Hearn, R., Zeidler, J., Jr, E., and Goodlin, R. (1975). Adaptive noise cancelling: Principles and applications. *Proceedings of the IEEE*, 63(12):1692–1716.

Wu, S., Kingsbury, B., Morgan, N., and Greenberg, S. (1998a). Incorporating information from syllable-length time scales into automatic speech recognition. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1:721–724.

Wu, S., Kingsbury, B., Morgan, N., and Greenberg, S. (1998b). Performance improvements through combining phone- and syllable-scale information in automatic speech recognition. *Int. Conf. on Spoken Language Processing*, 2:459–462.

Yang, X., Wang, K., and Shamma, S. A. (1992). Auditory representations of acoustic signals. *IEEE Trans. on Inf. Theory*, 38(2):824–839.