Nonlinear Feature Transformations for Noise Robust Speech Recognition

presentée a la Facultè des sciences et techniques de l'ingènieur

École Polytechnique Fédérale de Lausanne

pour l'obtention du grade de docteur ès sciences

par

SHAJITH IKBAL

Bachelor of Science in Physics,
Madras University, Madras, India
and
Bachelor of Technology in Instrumentation Engineering,
Madras Institute of Technology, Anna University, Madras, India
and
Master of Science (by research) in Computer Science and Engineering,
(Thesis title: Autoassociative Neural Network Models for Speaker Verification)
Indian Institute of Technology (Madras), Chennai, India

Thesis committee members:

Prof. Juan Mosig, EPFL, Switzerland
Prof. Hervé Bourlard, directeur de thèse, IDIAP/EPFL , Switzerland
Prof. Hynek Hermansky, co-directeur de thèse, IDIAP, Switzerland
Prof. Hermann Ney, Aachen University, Germany
Prof. Richard Stern, Carnegie Mellon University, USA
Prof. Pierre Vandergheynst, EPFL, Switzerland

Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
November 2004.

Abstract

Robustness against external noise is an important requirement for automatic speech recognition (ASR) systems, when it comes to deploying them for practical applications. This thesis proposes and evaluates new feature-based approaches for improving the ASR noise robustness. These approaches are based on nonlinear transformations that, when applied to the spectrum or feature, aim to emphasize the part of the speech that is relatively more invariant to noise and/or deemphasize the part that is more sensitive to noise.

Spectral peaks constitute high signal-to-noise ratio part of the speech. Thus an efficient parameterization of the components only from the peak locations is expected to improve the noise robustness. An evaluation of this requires estimation of the peak locations. Two methods proposed in this thesis for the peak estimation task are: 1) frequency-based dynamic programming (DP) algorithm, that uses the spectral slope values of single time frame, and 2) HMM/ANN based algorithm, that uses distinct time-frequency (TF) patterns in the spectrogram (thus imposing temporal constraints during the peak estimation). The learning of the distinct TF patterns in an unsupervised manner makes the HMM/ANN based algorithm sensitive to energy fluctuations in the TF patterns, which is not the case with frequency-based DP algorithm.

For an efficient parameterization of spectral components around the peak locations, parameters describing activity pattern (energy surface) within local TF patterns around the spectral peaks are computed and used as features. These features, referred to as spectro-temporal activity pattern (STAP) features, show improved noise robustness, however they are inferior to the standard features in clean speech. The main reason for this is the complete masking of the non-peak regions in the spectrum, which also carry significant information required for clean speech recognition.

This leads to a development of a new approach that utilizes a soft-masking procedure instead of discarding the non-peak spectral components completely. In this approach, referred to as phase

autocorrelation (PAC) approach, the noise robustness is actually addressed in the autocorrelation domain (time-domain Fourier equivalent of the power spectral domain). It uses phase (i.e., angle) variation of the signal vector over time as a measure of correlation, as opposed to the regular autocorrelation which uses dot product. This alternative measure of autocorrelation is referred to as PAC, and is motivated by the fact that angle gets less disturbed by the additive disturbances than the dot product. Interestingly, the use of PAC has an effect of emphasizing the peaks and smoothing out the valleys, in the spectral domain, without explicitly estimating the peak locations. PAC features exhibit improved noise robustness. However, even the soft masking strategy tends to degrade the clean speech recognition performance.

This points to the fact that externally designed transformations, which do not take a complete account of underlying complexity of the speech signal, may not be able to improve the robustness without hurting the clean speech recognition. A better approach in this case will be to learn the transformation from the speech data itself in a data-driven manner, compromising between improving the noise robustness while keeping the clean performance intact. An existing data-driven approach called TANDEM is analyzed to validate this. In TANDEM approach, a multi-layer perceptron (MLP) used to perform a data-driven transformation of the input features, learns the transformation by getting trained in a supervised, discriminative mode, with phoneme labels as output classes. Such a training makes the MLP to perform a nonlinear discriminant analysis in the input feature space and thus makes it to learn a transformation that projects the input features onto a sub-space of maximum class discriminatory information. This projection is able to suppress the noise related variability, while keeping the speech discriminatory information intact. An experimental evaluation of the TANDEM approach shows that it is effective in improving the noise robustness. Interestingly, TANDEM approach is able to further improves the noise robustness of the STAP and PAC features, and also improve their clean speech recognition performance. The analysis of noise robustness of TANDEM has also lead to another interesting aspect of it namely, using it as an integration tool for adaptively combining multiple feature streams.

The validity of the various noise robust approaches developed in this thesis is shown by evaluating them on OGI Numbers95 database added with noises from Noisex92, and also with Aurora-2 database. A combination of robust features developed in this thesis along with standard features, in a TANDEM framework, result in a system that is reasonably robust in all conditions.

Version abrégée

La robustesse aux perturbations accoustiques externes est une condition importante pour les sytèmes de reconnaissance automatique de la parole (ASR) quand il est question de les déployer dans des applications pratiques. Cette thèse propose et évalue de nouvelles approches basées sur les charactéristiques extraites du signal vocal pour améliorer la robustesse au bruit des ASR. Ces approches sont centrées sur des transformations non linéaires qui, quand elles sont appliquées au spectre ou à la composante extraites du signal vocal, ont pour but de mettre en valeur la partie de parole qui est relativement moins invariante au bruit.

Les pics spectraux constituent une zône de rapport signal sur bruit élevé de la parole. Alors, un paramétrage efficace des composantes appartenants aux endroits des pics permettrait d'améliorer la robustesse au bruit. Deux méthodes proposées dans cette thèse pour la tâche d'estimation des pics sont :1) un algorithme de programmation dynamique (DP) basé sur la fréquence, utilisant les valeurs de dérivées spectrales de la période d'échantillage, et 2) un algorithme basé sur une méthode hybride HMM/ANN, qui utilise des formes de temps-fréquence (TF) distinctes dans le spectrogramme (imposant donc des contraintes temporelles au niveau de l'estimation des pics).

Pour un paramétrage efficace des composantes spectrales au niveau des pics, les paramètres décrivant la forme des activités (zône d'énergie) à l'intérieur des formes locales de TF au niveau des pics sont calculés et utilisés comme caratéristiques du signal. Ces caractéristiques, référées comme étant des caractéristiques de formes d'activités spectro-temporelles (STAP) montrent une amélioration de la robustesse au bruit, cependant ils sont inférieurs aux dans le cas d'un signal de parole non bruité.

Ceci mène au développement d'une nouvelle approche qui utilise une procédure de masquage léger. Dans cette approche intitulée autocorrélation de phase (PAC), la robustesse au bruit est présentée dans le domaine d'autocorrélation (Fourier dans le domaine temporelle équivalent au

domaine de puissance spectrale). Il utilise la variation de phase (ex. un angle) du vecteur au cours du temps comme étant la mesure de corrélation, opposée à l'autocorrélation standard qui utilise le produit scalaire. Cette mesure alternative d'autocorrélation est intitulée PAC et est motivée par le fait que l'angle est moins perturbé par les perturbations auditives. D'ailleurs, l'utilisation du PAC a un effet de mise en valeur des pics. PAC montre une ameélioration de la robustesse au bruit mais est inférieur dans le cas de signaux de parole non bruités.

Ceci nous mène alors au fait que les transformations qui ne prennent pas en compte la complexité du signal vocal peuvent ne pas être à même d'améliorer la robustesse, ceci sans dégrader la reconnaissance du signal vocal non-bruité. Une approche meilleur dans ce cas sera d'apprendre la transformation à partir des données accoustiques elles-mêmes avec une approche centrée sur les données, en améliorant la robustesse au bruit tout en gardant intactes les performances avec des signaux de parole non bruités. Une approche orientée sur les données appelée TANDEM est analysée pour valider cette hypothèse. Dans l'approche TANDEM, un MLP est utilisé pour exécuter une transformation orientée sur les données des caractéristiques d'entrée, il fait l'apprentissage de la transformation en étant entrainé dans un mode supervisé et discriminant, avec des t́iquettes de phonèmes comme étant les classes de sortie. Un tel apprentissage permet au MLP de projeter les caractéristiques d'entrée dans un sous-espace d'information linguistique qui permet de supprimer la variabilité relative du bruit. Une évaluation expérimentale de l'approche TANDEM montre que cette approche est efficace pour l'amélioration de la robustesse au bruit.

D'ailleurs, l'approche TANDEM améliore d'avantage la robustesse des caratéristiques STAP et PAC au bruit, et améliore aussi leurs performances dans le cas de signaux de parole non bruités. L'analyse de la robustesse au bruit de la méthode TANDEM permet de découvrir un autre aspect intéressant de celle-ci : son utilisation comme un outil d'intégration pour combiner adaptativement plusieurs flux de caratéristiques.

La validité des différentes approches robustes aux perturbations accoustiques développées dans cette thèse est montrée en les évaluant sur la base de données OGI Number95 en y ajoutant des perturabations accoustiques des base de donnée Noisex92 et Aurora-2.

Une combinaison des composantes robustes extraites du signal développées dans cette thèse avec des charactéristiques standards, dans un schéma TANDEM, résulte à un système qui est raisonnablement robuste dans toutes les conditions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Automatic Speech Recognition (ASR) provides a powerful means of communication between humans and computers. ASR systems have capability to revolutionize the way the world work, if they can achieve human like recognition performance. However, unfortunately, human like recognition performance is still far from the reach of the current systems. The main reason for this is the inability of the current algorithms to cope with various kinds of unwanted variabilities observed in the speech signal. The unwanted variabilities in speech arise due to several factors including the differences in environmental conditions in which the signal is generated, differences in the speech production mechanisms of the speakers, and differences between the characteristics of the transmission channels. This leads to a non-robust recognition performance of the current ASR systems when exposed to different conditions, limiting its scope.

1.1 Objective of the thesis

This thesis aims to improve the robustness of the ASR systems by handling the effects of a particular class of variability source, namely the differences in environmental conditions. The differences in environmental conditions could arise because of interfering noise, interfering speech, and environmental acoustics. Among these, the current thesis addresses the problem of interfering noise. The general approach followed in this thesis work to achieve noise robustness is predominantly motivated by the outcomes of the perceptual studies conducted with human beings that show evidences for a relative masking of the noise sensitive parts of the signal and an emphasis of the noise invariant parts of the signal while recognizing speech in noisy conditions (Moore, 1997; Allen, 1994; Fletcher, 1953).

1.2 The problem of robustness

1.2.1 Automatic speech recognition

Speech signal carries various information: linguistic information corresponding to the intended message, speaker information due to the differences between the vocal tract systems of the speakers, speaker's accent and speaking style information, information about the state of the speaker such as the speaker's stress level, information about the environment in which the signal is generated, and information about the channel through which the signal is transmitted. In ASR the aim is to infer the message conveyed from the speech signal. Hence the information of importance is the linguistic information. All other information stands in as irrelevant information introducing unwanted variability in the signal.

As shown in Figures 1.1 and 1.2, automatic speech recognition systems typically process the speech signal to extract *feature vectors* representing the linguistic information, and do a *statistical modeling* of these feature vectors to decode the message conveyed. Statistical modeling is mainly to handle the unwanted variabilities present in the speech signal. Although the feature extraction schemes ideally aim to extract feature vectors representing only linguistic information from the speech signal, usually it is not possible to discard the other information. This is because it is not clear from the signal which part of it corresponds to the linguistic information. The presence of other information introduces unwanted variability at the feature level also.

Statistical modeling step finds out an estimation of the class-dependent probability distributions of the feature vectors during a stage called *training* and during *recognition* it finds out how well the new features fits in those distributions. The estimation of the probability distribution is performed using a data set called training data, which usually has transcription of the speech in it to facilitate a supervised learning.

1.2.2 Issue of robustness

Ideally, to have a good estimation of the class-dependent probability distributions of the features, training data should include speech signals generated from all possible conditions. With that it will be possible to model all possible variabilities observed in the feature vectors. However, it is impractical to include all possible conditions during the training. This results in mismatch between the training and recognition data, i.e., training data may not represent the all possible speech data that the system may come across during recognition. As a result, the knowledge about the probability distributions of the feature vectors gained out of the training data do not necessarily apply to the recognition data. A typical example of this situation is an interaction with the ASR system through a mobile phone, where the environment in which speaker generates the speech differs from time to time.

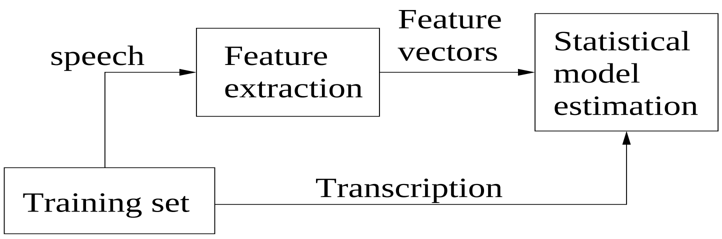

Figure 1.1. ASR systems: Block diagram of the training stage.

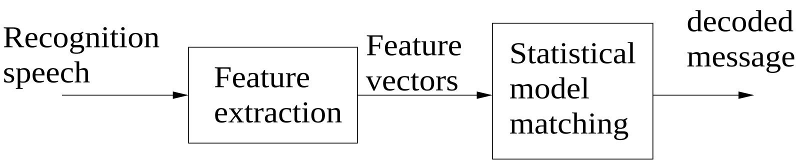

Figure 1.2. ASR systems: Block diagram of the recognition stage.

The practical limit on the coverage of all possible conditions by training data can be overcome if the statistical model can generalize what ever it has learned from the training data to unseen data. However, except for certain types of variabilities, the currently existing statistical techniques do not generalize to all possible variabilities observed in the speech signal. This leads to a poor performance of the ASR systems when exposed to different conditions. A large effort has been and is being spent by the speech recognition research community trying to solve this problem.

1.2.3 Environmental mismatch

Among various factors affecting the speech signal, environmental factors are of major concern when it comes to employing the ASR systems in practical scenarios. This is because of the huge variabilities introduced by those factors in the speech signal while generated in practical conditions. Their importance improves furthermore with the current emergence of mobile telephones, where the environmental condition in which the signal is generated changes drastically from time to time. Figure 1.3 shows an illustration of the typical environmental factors that influence the signal. These factors include among others the interfering noise, cross-talking speakers, room acoustics, room temperature, and wind speed. All these factors affect the signal as it propagates from the speaker to the receiver. Sometimes a few of these factors have influence on the speech generation itself. For example, in high noise conditions the speaker generating the speech signal tends to speak with a louder voice, and as a result, the whole characteristics of the speech signal changes. In literature, this is referred to as Lombard effect. However, we do not consider this effect for the work of this thesis.

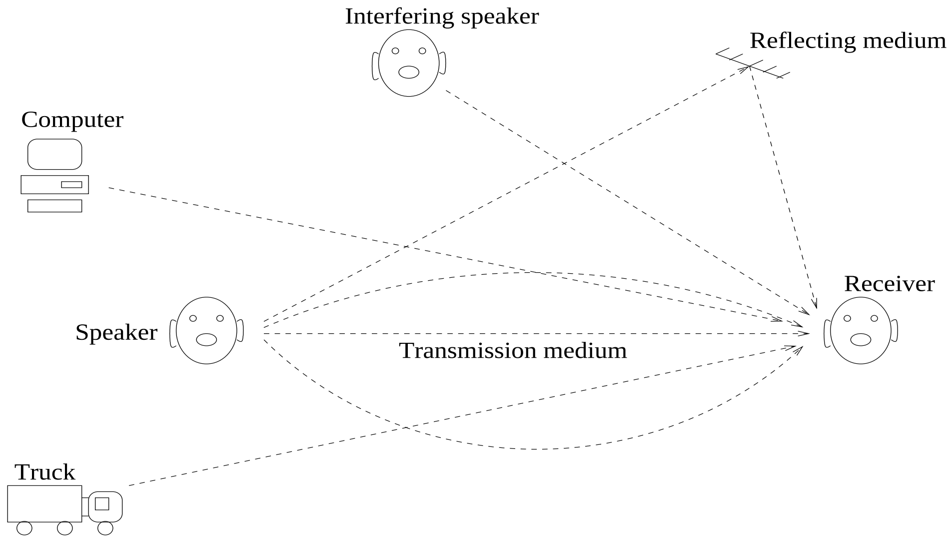

Figure 1.3. Illustration of various environmental factors affecting the speech signal.

There are several manners in which the different environmental factors can affect the speech signal. For example, in the presence of noise, the resultant signal is approximately an addition of the original speech signal and the interfering noise (Junqua and Haton, 1996). Depending upon whether the interfering noise is loud or weak the effect it has on the resultant signal could be severe or mild. The room acoustics affects the signal in such a way that the resultant signal is a convolution of the original signal and the impulse response characteristics of the room. The channel of propagation also has a convolutional effect on the speech signal (Junqua and Haton, 1996). However, convolutional effect of the channel is different from that of the room acoustics by the fact that impulse response of the channel has very short time constant. The signal that finally reaches the receiver is a mix of everything.

The wide variety of environmental factors and the variety in the manners in which they affect the speech signal lead to a huge variability in the speech signal. Current ASR systems, not able to cope with such variability, show a significant fall in the recognition performance when exposed to environmental mismatch. The fall in performance is a function both the environmental factor and the extent to which it affects the signal. This limits the ASR systems from being used in practice.

1.3 Scope of this thesis

The current thesis tries to address the problem of the effect of a particular class of environmental factor on ASR task, namely the interfering noise. A few nonlinear techniques to improve the robustness of the ASR system against these factors have been proposed, investigated, and evaluated. Previous methods proposed in the literature to handle the external noise are of two classes: *model based* and *feature based*. Model based methods assume the feature vectors to be sensitive to the external noise and tries to achieve robustness at the statistical modeling level, where as feature based methods try to reduce the sensitivity of the feature vectors to the external noise. The noise robust methods proposed in this thesis comes under the category of feature based methods.

1.4 Motivation for this thesis

Until the performance of automatic speech recognition system surpasses human performance in accuracy and robustness, we stand to gain by understanding the basic principles behind human speech recognition.

This is a quote by Jont. B. Allen (Allen, 1994) in his paper titled "How do humans process and recognize speech?". One of the main understandings obtained from the experiments on human speech recognition is that the humans work with partial information across frequency, probably in the form of speech features that are local in frequency, for example, the formants (Allen, 1994). However, the current automatic speech recognition systems use features extracted from the entire pattern of information across the frequency (Rabiner and Juang, 1993). This could be the main reason why the ASR systems are inferior to the humans in terms of the noise robustness. In the presence of noise not all the information across frequency are disturbed equally and hence processing them differently and separately gives good scope for improving the noise robustness. This is supported by the outcomes of perceptual studies conducted with human beings that show evidence for masking of the confusing aspects of the signal by human auditory system while recognition of speech in noisy conditions (Moore, 1997).

This constitute the major motivation for the current thesis work. The strategy we follow to improve the noise robustness is to give higher emphasis to the relatively noise invariant parts of the speech and lower emphasis to the relatively more noise sensitive parts of the speech, through appropriately designed nonlinear transformations. As we know, the regions around the spectral peaks get least disturbed in the presence of noise because of the high SNR ratio in those parts. On the other hand spectral valley regions have relatively low SNR and as a result are more sensitive to noise. Thus, an emphasis of spectral peaks and masking of the valleys is expected to improve the noise robustness. Also, identifying the general noise invariant parts of the speech in a data-driven manner and emphasizing them is expected to improve the noise robustness.

1.5 Evolution of the thesis work

The initial part of the work tries to utilize the knowledge that part of the speech corresponding to the peaks in the spectral domain is relatively more noise invariant and part of the speech corresponding to spectral valleys is relatively more sensitive to noise. A few feature extraction procedures, that impose nonlinear transformations that emphasis the spectral peaks and deemphasis or mask the spectral valleys are developed. The later part of the work tries to identify the noise invariant part of the speech in a data-driven manner to emphasis them.

Starting point - Previous work: Various studies conducted on human perception system claim spectral peaks as important and robust cues for speech recognition (Moore, 1997). The spectral peaks are also called formants. A few experimental studies done in the context of formant based ASR reported in previous literature also make this claim based on the observations made (McCandless, 1974; Welling and Ney, 1998; Misra *et al.*, 2004). A relatively recent work in this direction is the use of a new acoustic model called HMM2 for the formant-like frequency extraction (Weber *et al.*, 2003a). Formant-like frequencies estimated by applying the HMM2 to the spectrum has been shown to improve the speech recognition performance when used as an additional evidence along with the traditional features. However, in this thesis work, the spectral peak locations are used, not as an additional feature to the ASR system, but as an additional information to the feature extraction, in order to emphasis and deemphasis different parts of the speech. The first task to be done for this is the spectral peak location estimation.

Spectral peak location estimation: Two methods for spectral peak location estimation, namely 1) frequency-based dynamic programming algorithm, and 2) HMM/ANN based algorithm, have been developed. The development of these methods were motivated mainly by the HMM2 based peak location estimation (Weber *et al.*, 2003a). However, they differ from the HMM2 based method by the fact that the number of peak locations estimated is not constrained apriori. This is an important requirement as the aim of peak location estimation in our case is to use them for emphasis or deemphasis of different regions of the spectrum. The frequency-based dynamic programming algorithm utilize spectral slope values of a single time frame in order to estimate the spectral peaks, whereas HMM/ANN based algorithm use distinct time-frequency patterns in the spectrogram for the same task. The use of temporal context in case of HMM/ANN based algorithm is expected to impose temporal constraints during the peak location estimation. Assuming a reliable estimation of the spectral peak locations, the next step is to design nonlinear transformations that emphasis the spectral peaks and deemphasis the spectral valleys. This leads to the next work namely the spectro-temporal activity pattern (STAP) features.

Spectro-temporal activity pattern (STAP) features: STAP approach uses a simple technique of masking the non-peak regions to zeros, and thus utilizing information only from the regions around the spectral peaks for feature computation. For efficiently using the information from regions around the spectral peaks, STAP approach has drawn motivation from outcomes of physiological studies conducted on mammalian auditory system that show evidences of cortical neurons being sensitive to certain local time-frequency patterns in the incoming signal (Depireux *et al.*, 2001). Accordingly, in STAP approach parameters extracted from local time-frequency patterns around the spectral peaks (describing their activity, i,e., energy surface) are used as features. Experimental studies conducted using STAP features show that they are infact robust in noisy conditions. However, STAP feature has a major drawback that their clean speech recognition performance is significantly inferior, when compared to the standard features, which makes them difficult to use as a stand-alone feature in speech recognition systems. The main reason for this inferior performance in clean speech is the complete masking of the non-peak regions in the spectrum. This leads to the next step in the work namely phase autocorrelation approach where a soft-masking strategy is followed in order to improve the noise robustness.

Phase autocorrelation (PAC) features: PAC approach addresses the problem of noise robustness at the autocorrelation domain, as opposed to most of the other approaches which work at the spectral domain. Autocorrelation is a time-domain Fourier equivalent of the spectrum. The main motivation behind the development of the PAC features is the fact that the angle between the time delayed signal vectors gets less disturbed when compared to their dot product, in the presence of the noise. Regular autocorrelation computes correlation coefficients as dot product between the time-delayed signal vectors, whereas PAC computes correlation as angle between the vectors. Interestingly, the use of angle has an effect of emphasizing the peaks and smoothing out the valleys in the spectral domain. As opposed to the STAP approach, such emphasis and smoothing serve as soft-masking. Additionally, the emphasis and smoothing are performed without explicit estimation of the peak locations. The experimental evaluation of the PAC features illustrate their noise robustness. However, PAC features are also inferior to the regular features in clean speech, since soft-masking also hurts clean speech recognition performance. This leads to realization of the fact that designing transformations externally based on the limited knowledge that we have perceived from the speech signals may not be the right solution, as the underlying complexity of the speech signals usually do not allow to improve one factor without hurting the other. This is usually observed in most of the noise robust techniques developed in the past, where improving the noise robustness usually hurt the clean speech recognition performance. This leads to the next step in the thesis work namely data-driven approaches for noise robustness, where the transformation required for improving the noise robustness is learned from the data itself, compromising between improving the noise robustness and keeping the clean speech recognition performance intact.

Noise robustness analysis of TANDEM approach: Tandem approach has been proposed recently as a combination of two approaches namely the HMM/GMM and HMM/ANN (Hermansky *et al.*, 2000). It uses the transformed outputs of a discriminatingly trained MLP as a feature input to the HMM/GMM system. The MLP in this case acts as a data-driven feature extractor. In the current work we analyze and explore the prospects of improving noise robustness using TANDEM approach. The MLP actually performs a nonlinear discriminant analysis (NLDA) to project the input feature space onto a nonlinear sub-space of maximum possible sound class discriminatory information. Such a projection is expected to keep only the information along that space and all other information are either reduced or removed completely. Thus the transformation by the MLP is expected to improve the noise robustness if the noise related information in the feature space is not along the subspace of class discriminatory information. A simple analysis and experimental results (Ikbal *et al.*, 2004a) show that this is indeed the case.

Evidence combination in TANDEM approach: The final step in the work for the thesis is the consolidation of all the work so far. STAP, PAC, and TANDEM feature extraction algorithms perform different processing of the speech signal to come up with their own noise robust representations. Their independent processing give scope for the presence of some complementary information between these features. Thus an adaptive combination of these features may yield an improvement in the overall recognition performance. The TANDEM approach provides a nice framework for the combination of the features. The features can be combined in TANDEM either at the input of the MLP or at the output of the MLP. For combination at the output of the MLP, a method based on entropy similar to the one suggested in (Okawa *et al.*, 1998; Misra *et al.*, 2003) is used. An experimental evaluation of the combination show an improvement in the overall robustness of the resulting system (Ikbal *et al.*, 2004a).

Experiments on OGI Numbers95 and Aurora database: Experimental evaluation of noise robust methods developed in this thesis have been performed on two types of databases namely, 1) OGI Numbers95 database corrupted with noises from Noisex92 database and 2) Aurora database. Results on OGI Numbers95 are used throughout the thesis for illustration and the results on Aurora are given separately at the end of the thesis.

1.6 Contributions of the thesis

The main contributions[1] of this thesis are the development, analysis, and evaluation of a few non-linear transformations that when applied to the spectrum or feature improve their noise robustness. The central idea behind the development of such transformations is the fact that an improvement in noise robustness can be achieved by an emphasis of the part of the speech that is relatively noise invariant and/or deemphasis or masking the part that is more sensitive to the noise. This formulation, first of all, requires division of speech into two components, one more robust to the noise and the other more sensitive to the noise. Such division can be done either externally (based on the knowledge about the speech) or in a data-driven manner. In this thesis both the possibilities are explored, as explained below:

1. For the external division, the knowledge used is the fact that spectral peaks have relatively high signal-to-noise ratio (SNR). Accordingly, the part of the speech corresponding to peaks in the spectral domain is more robust to noise. Two different strategies followed for the enhancement of the spectral peaks and deemphasis of the spectral valleys, has lead to development of two different approaches for noise robust speech recognition explained as follows:

   (a) **Spectro-temporal activity pattern (STAP) features:** In this approach, non-peak regions in the spectrum are completely masked to zeros and parameters describing the activity (energy surface) within local time-frequency patterns around the peak location are used as features. The computation of STAP feature requires an estimation of the spectral peak locations. Two algorithm developed in this thesis for peak location estimation are:

      i. **Frequency-based dynamic programming algorithm**, that utilize the spectral slope values of the single time frame, for estimating the spectral peak locations.

      ii. **HMM/ANN based algorithm:** HMM/ANN used along the frequency axis, utilize distinct time-frequency patterns in the spectrogram to locate the spectral peaks. The use of temporal context impose temporal constraints during the peak location estimation.

   Both these algorithms differ from the previous peak estimation algorithm by the fact that the number of peak locations estimated is not fixed apriori.

[1]The exact contributions of this thesis are highlighted in this section by the bold-face letters.

   (b) **Phase autocorrelation (PAC) features:** This follows a soft-masking approach, as opposed to the complete masking of the non-peak regions as done in the STAP features. Additionally, the explicit peak location estimation is also avoided, thereby saving the features from being sensitive to the peak location estimation algorithms. An implicit enhancement of the peaks and smoothing of the valleys improve the noise robustness. Interestingly, these aspects of the PAC features are achieved by an use of an alternative measure to the autocorrelation called phase autocorrelation (PAC), that uses phase (i.e., angle) variation of the signal vectors over time.

2. For the data-driven approach, feature sub-space of maximum possible sound class discriminatory information, learned from the training data, using a recently proposed TANDEM approach, are assumed to constitute the noise invariant part of the speech. Two different manners in which TANDEM approach is used in this thesis are as follows:

   (a) **Noise robustness analysis of TANDEM approach:** The MLP used in TANDEM projects the feature space onto a sub-space of maximum possible sound discrimination information. Such projection leads to a reduction of noise related variability.

   (b) **Evidence combination in TANDEM approach:** Transformation performed by the MLP, that projects the input space onto a sub-space of maximum sound discrimination, acts as a nice integration tool to combine the features when the feature streams are fed simultaneously to the input of the MLP. In addition, the outputs of the MLPs that process the feature streams independently can also be combined through posterior combination schemes.

1.7 Organization of the thesis

The motivation for the current thesis work and the framework in which the work has been developed were discussed in the current chapter. The evolution of this thesis work and the contributions of this thesis work were also discussed.

Chapter 2 will give a detailed coverage of the prominent noise robust methods that appeared in the previous literature, discussing their advantage and disadvantages. For an easy understanding of these methods and also to have a smooth following of the later chapters of the thesis, a brief introduction to the state-of-the-art speech recognition system is given in the starting of the chapter.

In Chapter 3, two methods for spectral peak location estimation namely, 1) frequency-based dynamic programming algorithm, and 2) HMM/ANN based algorithm, are described. A brief explanation of a previous work, using HMM2 for the peak estimation is given in the starting of this chapter.

In Chapter 4, the peak locations estimated are further used to develop a new noise robust feature representation called STAP features. STAP approach uses parameters extracted from local time-frequency patterns, around the spectral peaks, as features. The motivation for this from previous physiological studies on mammalian auditory cortex is presented.

In Chapter 5, further developing from some of the interesting points learned from the STAP features, we introduce a new class of noise robust features called PAC features. In PAC features, an alternative measure of the regular autocorrelation, where an angle variation of the signal vector over time is used as a measure of correlation. In this chapter, various advantages and disadvantages of using angle as a measure of correlation are discussed. The effects of using angle in the spectral domain is discussed. Noise robustness of PAC features is discussed in detail along with validation through experimental results.

Chapter 6 presents an analysis and experimental validation of an existing data-driven approach for speech recognition, called TANDEM approach, for the case of noise robustness. The MLP used in TANDEM projects the input feature onto a space where speech discriminatory information are better emphasized. An explanation of how such a transformation helps in improving the robustness is given in this chapter along with the experimental validations. The suitability of employing STAP, PAC, and state-of-the-art features in TANDEM approach is discussed.

In Chapter 7, the complementary information between various features developed in this thesis are utilized to improve the overall recognition performance of the speech recognition system. Feature combination schemes in the TANDEM framework is presented. Feature combination at the MLP input and entropy based combination of the MLP output are discussed. Experimental evaluations showing the effectiveness of such a procedure is presented.

Throughout the thesis, experimental evaluation of the noise robust techniques developed are performed on Numbers95 database, corrupted with noises from Noisex92. In Chapter 8, critical experiments of the thesis are repeated on a database that is widely used by robust speech recognition community called Aurora. The results presented were discussed in comparison with results on Numbers95 database reported in previous chapters.

Chapter 9, summarizes and give a conclusion of the thesis work, mentioning the potential future directions.

Chapter 2

Robust Speech Recognition: A Review

Over several years of speech recognition research, several algorithms to improve the noise robustness have been developed by the researchers. While many of them work quite well for specific situations, in general, they do not generalize to all conditions. This chapter give a comprehensive review of the prominent noise robust techniques developed in the past. For a smooth following of the explanation of these techniques this chapter starts with a brief introduction to state-of-the-art ASR systems.

2.1 State-of-the-art ASR systems

Prominent approaches for ASR are based on pattern matching of the statistical representations of the speech signals. As illustrated in Figures 1.1 and 1.2, ASR involves sequence of operations: 1) feature extraction, and 2) statistical modeling. Feature extraction computes a sequence of vectors representing linguistic information in the speech signal. Statistical modeling estimates likelihood of match between that vector sequence and a set of reference probability density functions, to facilitate message decoding. As mentioned in section 1.2.1, the reference density functions are learned from a set of speech data called training data.

The existence of feature extraction as a separate block may be questioned, as the statistical modeling can also be performed directly on the speech signal. However, the existing statistical modeling techniques are not able to cope with all kinds of variabilities observed in the direct speech signal. Feature extraction often helps to discard a few of such unwanted variabilities by transforming the signal to another form, with the help of some external knowledge. It also helps to reduce the dimensionality of the signal vectors, thereby saving the statistical modeling step from the curse of dimensionality problem.

Additionally, because of the infinitely large number of possible word sequences, there are infinitely large number of possible distinct representations of the whole speech signal. This makes it highly impossible to perform a statistical modeling of the whole vector sequence. A divide-and-conquer strategy is followed to simplify this problem, where the word sequences are divided into smaller segments, with the total number of distinct segments being restricted to a finite number. Typically such a segmentation is done at phonetic level. Later on, powerful dynamic programming algorithms are then employed to recognize the whole sequence (Ney, 1984; Bourlard *et al.*, 1985),

The feature extraction and statistical modeling blocks of the state-of-the-art ASR systems are explained with more details in the following subsections:

2.1.1 Feature extraction

The ideal aim of feature extraction in ASR systems is to extract representations from the speech signal that has only linguistic information. However, this is hard to achieve. Various steps involved in typical feature extraction, as shown in figure 2.1, are explained in detail below:

Digitization of the signals: The speech signals generated by humans are continuous-time signals. For the processing of these signals by the machines, which can do only a digital processing, the signals are first digitized by an analog-to-digital (A/D) converter. A/D converter outputs the digital version of continuous-time signal by sampling it at equidistant points in time and then quantizing the amplitudes. Telephone speech is the most common speech used in ASR systems, whose bandwidth is typically from 200 Hz to 3400 Hz. According to Nyquist's sampling theorem, minimum sampling frequency for A/D conversion of a signal should at least be twice the maximum bandwidth of the signal, to avoid aliasing of the signal (an effect that avoids the perfect reconstruction of the continuous-time signals from the digitized signal) (Nyquist, 1928; Shannon and Weaver, 1949). Hence, the typical sampling frequency used for sampling of the speech signals is 8000 Hz.

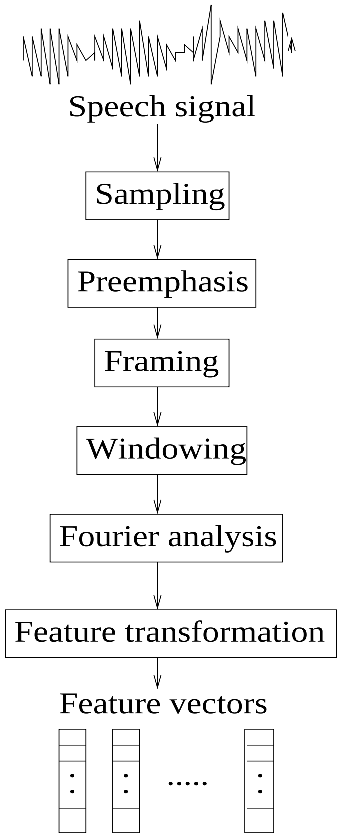

Figure 2.1. Block diagram of feature extraction.

Signal preemphasis: Signal preemphasis is originally motivated by the production model for voiced speech, according to which there is a spectral roll-off of −6dB/octave due to glottal closure and radiation from the lips. This is typically compensated by a pre-emphasis filter of form $1 - az^{-1}$, which flattens the spectrum of the voiced speech (O'Shaughnessy, 1987). Typical values used for a in the filter equation are in the range from 0.95 to 1.0.

Short-term analysis of the signals (framing): Most of the state-of-the-art features used for speech recognition are based on Fourier analysis of the signals. Fourier analysis requires the characteristics of the signal taken for analysis to be stationary throughout. But speech signals are, in general, nonstationary. However, from the knowledge about the human speech production system, inertia of the articulators do not allow the characteristics of the speech signal to change rapidly over time. In other words, the characteristics of the signal can be approximated to be stationary over a short period of time segments, typically of lengths 5 to 30 msec (Rabiner and

Schafer, 1978). Hence, for further processing, the speech signal is divided into a sequence of short signals called frames, by performing a sequence of shifting and windowing operation on the original signal. The typical length of the window used is 20-30 msec. Typical window shift used to obtain the frames is around 10 msec.

Windowing: Windowing in the time domain represents a convolution of the frequency domain Fourier equivalents of the speech and the window function (Oppenheim and Schafer, 1975). Such operation alters the characteristics of the signal if the frequency domain equivalent of the window function is not a spike function. A window function that is found to be more appropriate for the speech feature extraction is the Hamming window (Rabiner and Schafer, 1978), given by equation:

$$\begin{aligned} w[n] &= 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1 \\ & \quad 0, \qquad\qquad\qquad\qquad\quad \text{otherwise} \end{aligned} \tag{2.1}$$

Spectral analysis: Typical features used for speech recognition are based on the power spectral representation of the speech signal (Rabiner and Schafer, 1978). The power spectrum is computed by finding out magnitudes of the complex-valued Fourier coefficients obtained from the Discrete Fourier Transform (DFT) of the speech frames. If N represents the frame length and $\mathbf{s} = \{s[0], s[1], ..., s[N-1]\}$ represents the speech frame, then the DFT coefficients $S[k]$ can be computed by equation (Oppenheim and Schafer, 1975),

$$S[k] = \sum_{n=0}^{N-1} s[n] \exp\left(j\frac{2\pi}{N}kn\right), \quad 0 \le k \le N-1 \tag{2.2}$$

Feature transformation: Usually some external knowledge about the human perception system or human speech production system is utilized to transform the power spectrum to feature vectors. During such transformation, the main aim is to emphasis the linguistic information and suppress the unwanted variabilities present in the power spectrum. A few examples of the features extracted from power spectrum, that are shown to be successful for ASR, mel-frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980), linear prediction cepstral coefficients (LPCC) (McCandless, 1974; Rabiner and Juang, 1993), and perceptual linear prediction cepstral coefficients (PLP-CC) (Hermansky, 1990).

Instead of incorporating external knowledge during the feature extraction, the algorithms can also be made to learn transformations that would discard unwanted variabilities in the power spectrum. Such feature extraction schemes are called data-driven feature extraction. Examples of such data-driven feature extraction procedures are are principle component analysis (PCA) (Bourlard and Kamp, 1988), and its nonlinear equivalent using artificial neural networks (ANN) (Ikbal *et al.*, 1999; Ikbal, 1999a; Bishop, 1995), linear discriminant analysis (LDA) (Duda and Hart, 1973; Haeb-Umbach and Ney, 1992), its nonlinear equivalent called TANDEM approach (Hermansky *et al.*, 2000; Ellis *et al.*, 2001). The MFCC features, which are quite successful in the state-of-the-art speech recognition systems (and also will be used throughout this thesis as reference feature), are explained in more detail below. A few other features, specifically the features extracted in a data-driven manner such as the TANDEM approach based features, will be discussed later in the thesis at appropriate places.

The development of MFCC has its root in the outcomes of studies (Davis and Mermelstein, 1980) on human auditory perception. It tries to mimic the human auditory periphery by utilizing the knowledge about the human perception system in the feature transformation. The computation of MFCC is as follows: The power spectral values are integrated within overlapping mel-scaled critical band windows to obtain what is called mel-scaled critical bank spectrum. The critical band spectral amplitudes are then compressed by a logarithmic function. The resultant values are then transformed through an inverse discrete cosine transformation (DCT) to obtain the MFCC coefficients. Equation of DCT used to obtain cepstral representation from any spectral representation is:

$$c[l] = \sqrt{\frac{2}{\pi}} \sum_{k=0}^{N-1} S[k] \cos\left(\frac{\pi l}{N}(k+0.5)\right), \quad l = 0, 1, ..., N-1 \tag{2.3}$$

The higher order cepstral coefficients are usually dropped, assuming the nonlinguistic sources of variabilities would then be reduced. An equivalent of this operation in spectral domain is smoothing.

State-of-the-art speech recognition systems typically incorporate the temporal dynamics of the speech signal in the feature representation by including the first and second derivatives of the static feature vectors (Furui, 1986). The first derivative of the static feature coefficients is referred to as the delta feature and the second derivative is referred to as the acceleration feature. The delta is computed over a short time window using equation:

$$\triangle \mathbf{x}_t = \frac{\sum_{d=1}^{D} d(\mathbf{x}_{t+d} - \mathbf{x}_{t-d})}{2\sum_{d=1}^{D} d^2} \tag{2.4}$$

where D denotes the time window length over which the delta is computed. The same equation is used to compute the acceleration feature, $\triangle\triangle\mathbf{x}_t$, by replacing the static features with delta features.

These sequence of operations in the feature extraction block finally outputs feature vectors every 10 msec, as shown in figure 2.1. These feature vectors then goes as input to the next stage, the statistical modeling.

2.1.2 Statistical modeling

If $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{T-1}\}$ represents the set of feature vectors extracted from the speech signal, statistical modeling techniques formulate the speech recognition problem as a maximum a posteriori (MAP) problem (Bourlard and Morgan, 1993; Rabiner and Juang, 1993) as follows:

$$W^* = \underset{W \in \mathcal{L}}{\operatorname{argmax}} \; P(W|\mathbf{X}) \tag{2.5}$$

The most likely word sequence W^* from the set of all possible word sequences $\mathcal{L}$, given $\mathbf{X}$, is chosen as the recognition result. The MAP formulation of speech recognition is hard to deal with directly. It is usually reformulated into a problem based on likelihood estimation using Bayes equation[1], as follows:

$$W^* = \underset{W \in \mathcal{L}}{\operatorname{argmax}} \; \frac{P(\mathbf{X}, W)}{P(\mathbf{X})} \simeq \underset{W \in \mathcal{L}}{\operatorname{argmax}} \; P(\mathbf{X}|W)P(W) \tag{2.6}$$

$P(\mathbf{X})$ has been dropped from the above equation as it serves just as a scaling factor. In the above equation, $P(\mathbf{X}|W)$, the conditional probability of $\mathbf{X}$ given W, is usually referred to as *acoustic model*, and $P(W)$, the prior probability of word sequence W, is referred to as *language model*. In practice, both the acoustic and language models are assumed to fit in some parametric form, say $P_\Theta(.)$ and $P_\Gamma(.)$, with parameter sets Θ and Γ respectively. Then $P_\Theta(\mathbf{X}|W, \Theta)$ and $P_\Gamma(W|\Gamma)$ are used

[1] $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

in (2.6) as estimates for $P(\mathbf{X}|W)$ and $P(W)$. The values of model parameters Θ and Γ are estimated from the training database containing a large collection of utterances, with known transcriptions. If $\mathcal{E}$ represents the set of all the training utterances along with the corresponding transcriptions, ideally the parameters can be estimated according to:

$$\{\Theta^*, \Gamma^*\} = \underset{\Theta \text{ and } \Gamma}{\operatorname{argmax}} \left[\prod_{[\mathbf{X},W] \in \mathcal{E}} P_\Theta(\mathbf{X}|W,\Theta) P_\Gamma(W|\Gamma) \right] \tag{2.7}$$

But practical constraints do not allow the joint estimation of Θ and Γ. They are usually estimated independently of each other from different training sets, say $\mathcal{E}_a$ and $\mathcal{E}_l$ respectively, yielding:

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \left[\prod_{\mathbf{X} \in \mathcal{E}_a} P_\Theta(\mathbf{X}|W,\Theta) \right] \tag{2.8}$$

$$\Gamma^* = \underset{\Gamma}{\operatorname{argmax}} \left[\prod_{W \in \mathcal{E}_l} P_\Gamma(W|\Gamma) \right] \tag{2.9}$$

Equation (2.8) is referred to as maximum likelihood (ML) training. A popular ML training algorithm is expectation-maximization (EM) algorithm (Baum *et al.*, 1970; Dempster *et al.*, 1977), where a few hidden variables are added to the existing parameter set in order to simply the otherwise usually intractable training problem. EM is an iterative procedure where, in each iteration, the new values for the parameter set, Θ^{new}, are found from the old values, Θ^{old}, so that the overall likelihood of the training data is increased:

$$\prod_{\mathbf{X} \in \mathcal{E}_a} P_\Theta(\mathbf{X}|W,\Theta^{\text{new}}) \geq \prod_{\mathbf{X} \in \mathcal{E}_a} P_\Theta(\mathbf{X}|W,\Theta^{\text{old}}) \tag{2.10}$$

Every iteration of EM has two steps: E and M steps. In E step, estimates of posterior distribution of the hidden variables are found from the old values of the parameter set. In M step, those estimates are used to find new values for the parameters. These steps, when repeated, increases the overall likelihood of training data and there are proofs that show the guaranteed convergence of this procedure (Baum *et al.*, 1970; Dempster *et al.*, 1977).

State-of-the-art ASR systems use hidden Markov models (HMM) (Bourlard and Morgan, 1993;

Rabiner and Juang, 1993) for acoustic modeling and bigram/trigram probabilities for language modeling (Shikano, 1987). As language modeling does not fall within the scope of this thesis it will not be discussed further. HMM used for acoustic modeling is explained in more detail below:

Hidden Markov model (HMM): The most successful approach developed so far for the acoustic modeling task of the ASR is the hidden Markov model (HMM). HMM is basically a stochastic finite state automaton, i.e., a finite state automaton with stochastic output process associated to each state. HMM models the speech by assuming the feature vector sequence $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{T-1}\}$ to be a piece-wise stationary process that has been generated by a sequence of HMM states, denoted by $Q = \{q_0, q_1, ..., q_{T-1}\}$, that transit from one to another over time. The stochastic output process associated with each state is assumed to govern the generation of feature vectors by the states. If $\mathcal{C}$ represents the set of all possible state sequences, the acoustic model in (2.8) can be rewritten as:

$$P(\mathbf{X}|W) = \sum_{Q \in \mathcal{C}} P(\mathbf{X}, Q|W) \tag{2.11}$$

In the above equation, Θ as appear in (2.8) is dropped for simplicity reasons. To make the model simple and computationally tractable a few simplifying assumptions are made while applying HMMs to the acoustic modeling problem. They are:

1. First-order hidden Markov model assumption[2], i.e.,

$$P(q_t|q_0, q_1, ..., q_{t-1}, q_{t+1}, ..., q_{T-1}, \mathbf{X}) = p(q_t|q_{t-1}) \tag{2.12}$$

   where $p(q_t|q_{t-1})$ is referred to as the state transition probability.

2. Feature independence (i.i.d.) assumption, i.e.,

$$P(\mathbf{x}_t|\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{t-1}, \mathbf{x}_{t+1}, ..., \mathbf{x}_{T-1}, Q) = p(\mathbf{x}_t|q_t) \tag{2.13}$$

   where $p(\mathbf{x}_t|q_t)$ is referred to as the emission probability, i.e., the probability of the state q_t emitting the feature vector $\mathbf{x}_t$.

[2] Here and throughout this thesis $P(.)$ is used to denote the global probability and $p(.)$ is used to denote the local probabilities.

With these assumptions, (2.11) becomes,

$$P(\mathbf{X}|W) = \sum_{Q \in \mathcal{C}} p(q_0) p(\mathbf{x}_0|q_0) \prod_{t=1}^{T-1} p(q_t|q_{t-1}) p(\mathbf{x}_t|q_t) \tag{2.14}$$

The above equation gives an exact formula for the computation of the likelihood. However, some times it is also approximated as the likelihood of the best state sequence as follows:

$$P(\mathbf{X}|W) = \max_{Q \in \mathcal{C}} p(q_0) p(\mathbf{x}_0|q_0) \prod_{t=1}^{T-1} p(q_t|q_{t-1}) p(\mathbf{x}_t|q_t) \tag{2.15}$$

This is called Viterbi approximation.

An illustration of HMM serving as an acoustic model with the above assumptions is given in Figure 2.2. If the number of states in HMM is M, the complete set of parameters that describe the complete HMM are the following:

1. transition probabilities, $p(q_t = \text{state } j | q_{t-1} = \text{state } i)$, denoted by $a_{ij}, 0 \leq i, j \leq M-1$ satisfying constraint $\sum_{j=0}^{M-1} a_{ij} = 1$, and

2. state emission density functions, $p(\mathbf{x}_t | q_t = \text{state } i)$, denoted by $p_i(.), 0 \leq i \leq M-1$.

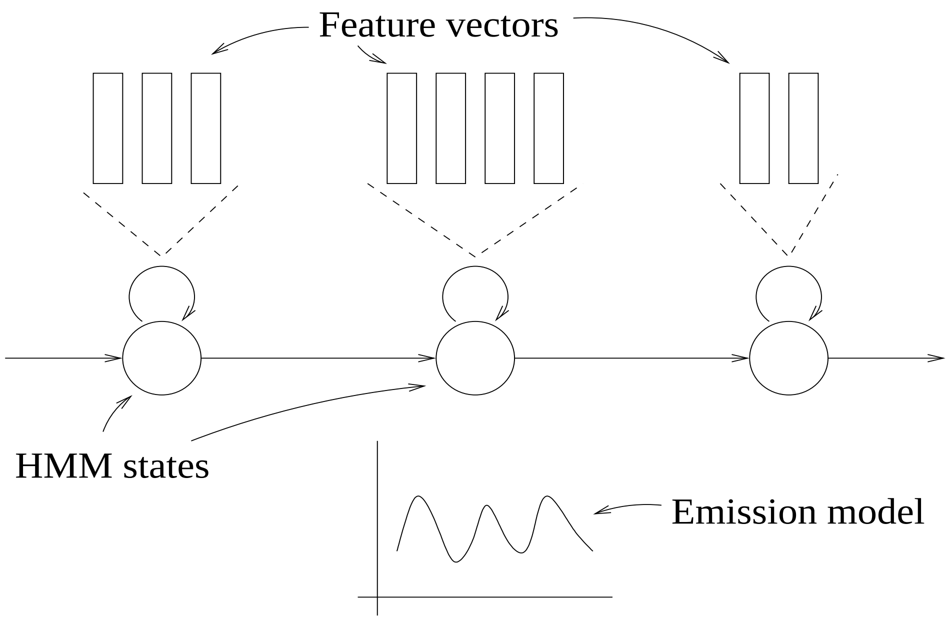

Figure 2.2. Illustration of Hidden Markov Model.

An explanation of EM algorithm to train the HMM parameters is given in (Ikbal *et al.*, 2001; Bilmes, 1998). Popular tools used for emission density modeling are Gaussian mixture model and multi-layer perceptron, explained as follows:

- Gaussian mixture model (GMM) is a weighted mixture of several Gaussians. It is characterized by the weighting factors, mean vectors, and covariance matrices of all the constituent Gaussians. The expression for density function $p(\mathbf{x})$ for GMM is given by,

$$p(\mathbf{x}) = \sum_{k=0}^{K-1} c_k G_k(\mathbf{x}) \tag{2.16}$$

  where K denotes the number of Gaussians in the GMM, and c_k denotes the weighting factor for k^{th} Gaussian, $G_k(.)$. If μ_k and $\mathbf{\Sigma}_k$ denote respectively the mean vector and covariance matrix of the k^{th} Gaussian, and if D denote the feature vector dimension, the expression for $G_k(\mathbf{x})$ is given by,

$$G_k(\mathbf{x}) = \frac{1}{2\pi^{\frac{D}{2}}|\mathbf{\Sigma}_k|^{\frac{1}{2}}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\mu_k)^T \mathbf{\Sigma_k}^{-1}(\mathbf{x}-\mu_k)} \tag{2.17}$$

  Thus the whole parameter set corresponding to the emission density function when using GMM are,

$$[c_k \; \mu_k \; \mathbf{\Sigma}_k] \, , 0 \le k \le K-1$$

- multi-Layer perceptron (MLP), a special case of artificial Neural network (ANN) (Bishop, 1995; Haykin, 1994), has a series of layers of artificial neurons. Neurons in each layer is fully connected with the neurons of the following layer. The first layer is called input layer, the last layer is called output layer, and all the inbetween layers are called hidden layers. Every neuron of all the layers perform nonlinear operation: a weighted sum of the outputs of all the neurons from which it receives the input, followed by a sigmoid or a softmax operation. The connection weighting factors between the neurons are called the weights. The vector applied at the input layer propagates through the hidden layers one by one until it finally reaches the the output layer. MLPs can be used for classification or function mapping

(Bishop, 1995; Haykin, 1994). Arbitrary complex decision regions can be formed by the MLPs while using them in classification mode (Lippmann, 1997). Any continuous mapping between the input and output space can be represented by the MLP while using them in function mapping mode (Funahashi, 1994). The connection weights of the MLP can be trained using Error Back Propagation training (Rumelhart *et al.*, 1988).

For emission density modeling, the MLP is used in pattern classification mode, where the number of output classes, i.e., the number of output neurons, corresponds to the number of HMM states. In pattern classification mode, the outputs of MLP estimate the posterior probability of the input vector $\mathbf{x}$, $p(q_k|\mathbf{x})$ (Bourlard and Morgan, 1993). Using Bayes rule, likelihood as used in (2.14) and (2.15) can be calculated as:

$$p(\mathbf{x}|q_k) = \frac{p(q_k|\mathbf{x})p(\mathbf{x})}{p(q_k)} \tag{2.18}$$

where $p(q_k)$ denote the apriori class probability and can be estimated from the training set. However, since $p(\mathbf{x})$ is a constant for all the classes and thus appearing just as a scaling factor, a scaled likelihood, as calculated by the equation below, is used in (2.14) and (2.15) instead of the likelihoods $p(\mathbf{x}|q_k)$.

$$\frac{p(\mathbf{x}|q_k)}{p(\mathbf{x})} = \frac{p(q_k|\mathbf{x})}{p(q_k)} \tag{2.19}$$

2.2 Noise robust speech recognition

Solution to the problem of sensitivity of the speech recognition systems to external noise can be approached in two different ways. Accordingly, the noise robust techniques are grouped into classes:

1. Model based approaches, and
2. Feature based approaches.

Model based approaches assume the feature vectors to be sensitive to the external noise and try to handle this sensitivity at the statistical modeling level. Whereas, feature based approaches try to make the feature vectors insensitive to the external noise. Several successful techniques developed under both the approaches have been reported in the literature. A few prominent techniques are explained in the next two sections. However, before going into the details of various noise robust methods, it will be useful to have a look at the various types of noises and understand the manner in which they affect the speech signal, which is discussed in the next subsection.

Effect of noise on speech signal

As mentioned in chapter 1, the noise mainly affects the speech signal as it propagated from the speaker to the receiver. The noise could be correlated with the speech or uncorrelated. Correlated noise results from the reflection or reverberation. Externally generated noise are usually uncorrelated. The uncorrelated noise could be stationary or nonstationary, wide-band or narrow-band (colored), and could last for only a short-time or a long-time (continuous). Nonstationary noises have their statistical characteristics changing over time. Wide-band has distribution of its energy over the entire frequency range. A few examples of the nonstationary short-time noises are door slamming, and car passing. Noises from factory environment and competing speaker are examples of the nonstationary continuous noise. Fan and air-conditioning noise are stationary and continuous. Siren is an example of a colored noise.

The type of noises considered in this thesis work are only the externally generated noises that are uncorrelated with the speech signal. As explained in section 1.2.3, the resultant signal of the two sound sources is approximately an addition of the individual signals. Suppose $s[n]$ denote the speech signal generated by the speaker, and $r[n]$ denote the resultant signal of all the noise sources in the surrounding environment. Then the resultant signal that reaches the receiver is $s[n] + r[n]$. Suppose $c[n]$ represents the impulse response of the transmission channel between the speaker and the receiver. For the sake of simplicity, the transmission channel is assumed to be time-invariant and uniform throughout. Then, including the effect of the transmission channel, the resultant signal that reaches the receiver is given as follows (Stern *et al.*, 1996):

$$\tilde{s}[n] = (s[n] + r[n]) \otimes c[n] \tag{2.20}$$

where $\otimes$ denote the convolution operation. However, it will not a make difference in the above equation if the noise is considered to affect the speech signal after it passes through the transmission channel, as the noise is usually not characterized as being from a single sound source. Hence the above equation can be written as:

$$\tilde{s}[n] = s[n] \otimes c[n] + \tilde{r}[n] \tag{2.21}$$

where $\tilde{r}[n]$ denote the effective noise. Most of the noise robust techniques, reported in the literature, assume the effect of noise on speech is according to the above equation, and try to handle it during recognition, especially for the situation when the system is trained only on the clean speech.

Suppose the power spectral representation of the telephone speech is $S[k]$ and the effective noise is $R[k]$. If the noise is uncorrelated with the speech, then the power spectral representation of the noisy speech is given by:

$$\tilde{S}[k] = S[k] + R[k] \tag{2.22}$$

At this point, it is important to define a term called signal-to-noise ratio (SNR) that gives a measure of the extent to which the speech signal is affected by the noise. SNR is basically a ratio between the amplitudes of the signal component and the noise component, usually specified in decibel (dB), which is 20 times the logarithm of the ratio, given by the equation below:

$$\text{SNR} = 20 \log \left(\frac{\text{signal amplitude}}{\text{noise amplitude}} \right) \tag{2.23}$$

2.3 Model based approaches

The variability due to external noise is accounted for in the model based approaches either by adapting the statistical model to match the new acoustic environment (through the estimation of the noise distribution or through the estimation of the perturbations in the speech distributions caused by the noise) or by making the statistical model to discard the information from the unreliable part of the feature. The model based approaches, especially adaptation based techniques, are computationally expensive. Some specific techniques need an impractical requirement of more amount of speech data during the recognition. A few model based approaches are explained briefly in the following subsections.

2.3.1 Multicondition training

A simple and direct model based method for achieving noise robustness is the inclusion of all possible testing noise conditions in the training set (Furui, 1992). By this way the statistical modeling will be able to model the all possible variabilities observed in the feature vectors due to external noise. This in fact has been shown experimentally to yield good improvements in the noisy speech recognition performance. However, this method is completely unrealistic in the sense that it is noise possible to include all possible testing noise conditions during the training. A little variant of this approach is to include a set of representative noise conditions in the training set and make the statistical models to generalize for the unseen noise conditions. This has been observed to result in improved robustness, though the relative degradations are more than the case when directly trained on the appropriate noise conditions.

2.3.2 Signal decomposition

The idea in signal decomposition is to recognize the concurrent signals simultaneously using a set of HMMs, one each for the components into which the signal is to be decomposed (Varga and Moore, 1990). Recognition is carried out by searching through the combined state space of the constituent models. For example, if the signal considered is speech added with noise from a single source, the search will be through a three dimensional space. If $\mathbf{r}_t$ represents the noise component added to the speech component $\mathbf{x}_t$ to obtain the resultant representation $\tilde{\mathbf{x}}_t$, and if q_t and p_t represent the states of speech and noise HMMs respectively, then the likelihood to be used in three the dimensional search is given by:

$$P(\tilde{\mathbf{x}}_t|q_t, p_t) = \int_{S_{\mathbf{x}_t,\mathbf{r}_t}} P(\mathbf{x}_t, \mathbf{r}_t|q_t, p_t) \tag{2.24}$$

where $S_{\mathbf{x}_t,\mathbf{r}_t}$ represents the set of all possible pairs $\{\mathbf{x}_t, \mathbf{r}_t\}$ such that $\tilde{\mathbf{x}}_t = \mathbf{x}_t + \mathbf{r}_t$.

As like multicondition training, signal decomposition also ideally needs an impractical requirement of inclusion of all possible noise conditions during the training. Additionally, the computational cost also increases exponentially with the increase in number of noise components to be recognized in the signal. Even the presence of a single noise component requires search through three dimensional space, which is computationally more expensive. Also, it is not known apriori how many signal components are present in the signal.

2.3.3 Parallel model combination

A variation of signal decomposition is parallel model combination (PMC) (Gales and Young, 1996) where, instead of using independent speech and noise models to perform recognition of observation sequence in a three or more dimensional space, the speech and noise models are first combined to form a noisy speech model and then the recognition is performed using that model. In this case, in order to device an appropriate model combination algorithm, it is important to understand quite well about how the component signals are combined. Additionally, the domain in which the model combination should be performed is also important. For example, if the speech is independent of the noise they are additive in the power spectral domain. However, if the feature vectors used are in the cepstral domain, then first the model parameters should be transformed from the cepstral domain to the spectral domain and then after combining the models, the combined model should be brought back to the cepstral domain (Gales, 1996a).

2.3.4 Maximum likelihood linear regression (MLLR)

MLLR, originally developed to handle the speaker related variabilities in speech (Legetter and Woodland, 1995), can also be used to handle noise related variabilities. In MLLR based method, the GMM parameters, such as the mean vectors and possibly the covariance matrices, are adapted to new environmental condition using data from the new environment. The linear transformation to adapt the mean vector, μ, of one of the constituent Gaussians of a state is given by equation:

$$\tilde{\mu} = \mathbf{A} \, [\, w \;\; \mu^{\mathrm{T}} \,]^{\mathrm{T}}$$

where $\mathbf{A}$ represents the regression matrix and w the offset, which is either 1 or 0, depending upon whether the offset is present or not. The matrix $\mathbf{A}$ is estimated using speech data from the new acoustic environment in a maximum likelihood fashion. This requirement, of test data to adapt the model parameters, stands in as a disadvantage as usually the estimation of parameters require a large amount of data during recognition. The effective number of parameters in the regression matrix can be reduced by tying the parameters, which in turn require lesser amount of data from the new environmental condition. However, the requirement of even a reasonable amount of data during recognition is impractical, except for a few specific applications.

Additionally, the transformation required to adapt the model parameters to new acoustic condition may not be linear. In such a case, nonlinear transformation can be realized by a mixture of linear regression classes, which in turn require a large amount of adaptation data.

2.3.5 Multi-band and multi-stream processing

In multi-band processing (Bourlard *et al.*, 1996; Bourlard and Dupont, 1996a; Sharma, 1999), the spectral representation of the speech signal is split into several frequency bands and each band is processed separately to first extract subband feature vectors. These features are then processed with different HMMs for every subband to extract subband likelihoods or probabilities. These probabilities are then combined to yield an effective likelihood that can be used for recognition. The contribution from different subbands for computing the combined probability is varied based upon the reliability of the subband. If a particular subband is corrupted its contribution to the combined probability estimate is decreased to decrease the effect of noise. This method is effective for the cases where speech is affected by colored noise. However, independent processing assumes each spectral band to be independent, which in turn causes the performance to degrade for clean speech. This can be avoided by a recently proposed full combination method (Hagen *et al.*, 1998; Hagen, 2001), where all possible combinations of subbands including the full-band, which does not treat different bands independently, are considered while computing the final likelihood to use in decoding.

Multi-stream processing (Bourlard *et al.*, 1996b; Hagen, 2001; Sharma, 1999) combines evidences from several streams of feature vectors extracted from the same speech signal. Different processing techniques used for extracting different features may emphasis different aspects of the signals which are complementary in nature. Thus an adaptive combination of the stream may yield a recognition performance better than that of the individual streams.

Multi-stream feature combination in ASR systems has its roots from the psychoacoustic studies on human beings (Fletcher, 1953; Allen, 1994), which show evidences for multiple representations of the speech signal at different stages of human auditory processing and an integration of them in order to get a robust final representation. Different processing steps done in different feature extraction schemes for ASR systems may provide different kinds of evidences under different environmental conditions. Hence an appropriate combination of such features, making use of their complementary information, is expected improve the overall recognition performance. The combination can be done at various stages of the ASR system as follows:

1. Feature combination: The feature vectors are combined before the statistical modeling. One simple method to combine the features is to concatenate the features to get a single feature vector of larger dimension, as done in the case of static and dynamic features.

2. Posterior combination: The probability outputs of the different acoustic models, processing different feature streams, are combined in order to obtain the combined probability (Morris *et al.*, 2001; Hagen, 2001).

2.3.6 Missing data approach

This method relies on the fact that some of the spectro-temporal regions in the spectrograms will be dominated by the external noise. Thus, during recognition, by treating these regions as missing or unreliable the overall robustness can be improved. The recognition is only based on the regions that are tagged as reliable (Cooke *et al.*, 1997; Raj *et al.*, 1998; Cooke *et al.*, 2001; Raj *et al.*, 2001). If the components in $\mathbf{x}$ belonging to the reliable part are denoted by $\mathbf{x}_r$ and those belonging to unreliable parts are denoted by $\mathbf{x}_u$, then, during the recognition, the missing data is dealt in one of the following two ways.

1. Marginalization, i.e., the local emission probability is approximately computed as just the emission probability of the reliable part, $\mathbf{x}_r$, as follows:

$$p(\mathbf{x}_r|q_t) = \int_{\mathbf{x}_u} p(\mathbf{x}_r, \mathbf{x}_u|q_t)d\mathbf{x}_u \tag{2.25}$$

2. Data imputation, where values corresponding to the unreliable regions are estimated to produce an estimate of the complete observation vector $\hat{\mathbf{x}}$, which is further used for computing the local emission probability as $p(\mathbf{x}|q_t) = p(\hat{\mathbf{x}}|q_t)$.

The practical implementation of the missing data approach requires a robust algorithm to identify the reliable regions in the spectrogram. In the related work, reported in literature, simple noise estimation techniques are used as basis for the identification task.

2.4 Feature based approaches

Feature based methods avoid the computationally intensive model based methods by generating feature representations that are invariant to the noise. A review of prominent feature-techniques can be found in (Stern *et al.*, 1996, 1997). These methods often involve the use of external knowledge about the effect of the noise on the features, in order to device an appropriate algorithm. Such knowledge is basically used to design transformations that would supposedly remove the noise prone aspects of the features.

2.4.1 The use of psychoacoustic and neurophysical knowledge

Early feature based methods involve incorporation of various psychoacoustic and neurophysical knowledge, obtained from human auditory system, into the feature extraction algorithm. As human auditory system is the best speech preprocessing system to date, imitating a few functionalities of it in the feature extraction algorithm is expected to improve the noise robustness of the ASR systems. Mel-frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980) and perceptual linear prediction (PLP) (Hermansky, 1990) features are widely used features falling in this category.

The MFCC, as explained in section 2.1.1, uses the mel-warped frequency axis and approximates the power law of hearing by taking the logarithm of the critical band power spectrum. With these operations the feature vectors have been shown to improve the recognition performance.

One popularly used feature during the early stages of the speech recognition is linear prediction (LP) cepstrum (Makhoul, 1975; Rabiner and Juang, 1993). LP cepstrum is computed through LP analysis of the speech signal. LP assumes the speech production system to be an all-pole model. The all-pole model parameters are first estimated from the samples to compute LP spectrum to finally compute LP cepstrum using (2.3). An improvement over the simple liner prediction (LP) analysis, utilizing the auditory peripheral knowledge, is the perceptual linear prediction (PLP) (Hermansky, 1990). Before doing the LP analysis an estimate of the auditory spectrum is obtained from the power spectrum by applying several transformations that are assumed to happen at the human auditory periphery front-end. The series of transformations include critical band integration (on bark-scale), equal-loudness preemphasis, cubic-root compression (to account for power law of hearing). The auditory spectrum obtained is then used to predict the LP coefficients, and then the equivalent PLP spectrum, and finally the PLP cepstrum. As like MFCC, PLP cepstrum are also expected to have decreased unwanted variabilities as a result of the incorporation of various auditory like transformations (Hermansky, 1990).

2.4.2 Speech enhancement

A different class of feature based noise robust methods try to enhance the speech-specific aspects of the spectrum by suppressing the noise-specific aspects. An early method falling in this category is the spectral subtraction (Boll, 1979). It gets an estimate of the enhanced spectrum $\hat{S}[k]$ from the original spectrum $S[k]$ using an estimate of the noise spectrum $R[k]$ as follows:

$$\hat{S}[k] = S[k] - R[k] \tag{2.26}$$

The success of this method relies on the reliable estimation of the noise power spectrum. The noise power spectrum is usually estimated from the non-speech intervals of the signal. Thus a reliable speech versus noise detector is required. Especially in the case of low SNR, due to spectral similarities between the unvoiced speech sounds and the noise, the noise estimation could become a difficult task. Furthermore, this technique is suitable only for the cases where the noise characteristics are stationary. In case of non-stationary noise, it may result in the removal of significant speech information. The subtraction of the noise power can result in negative values if the noise estimate exceeds the actual noise magnitude. This can be taken into care by setting a threshold for the power values, which introduces residual noise (also called musical noise) in the signal domain.

An improvement over the spectral subtraction is nonlinear spectral subtraction (NSS) (Lockwood and Boudy, 1992), which combines spectral subtraction with noise masking. NSS has been demonstrated to improve the speech recognition performance in car noise conditions (Lockwood *et al.*, 1992a). Another variant of spectral subtraction is continuous spectral subtraction (Nolazco-Flores and Young, 1994) which involves a continuous calculation of smoothed estimate of the long term spectrum for noise removal.

A relatively new technique that has been shown to be quite successful for recognition of speech corrupted by slow varying noise is relative spectra (RASTA) processing (Hermansky and Morgan, 1994). It tries to suppress those noise components whose temporal properties are quite different from that of the speech, in the spectral domain. The temporal properties of different frequency bands in the spectrum are modeled by the modulation spectrum. The lower bound of the modulation spectral bandwidth of the clean speech gives a measure of the lowest possible rate at which the signal components of speech can be generated, while higher bound gives the highest possible rate. Thus the modulation spectral components beyond the bandwidth of the clean speech can be assumed to be from the noise source. A band-pass filter, whose bandwidth is equal to the modulation spectral bandwidth of the clean speech is applied to each frequency band of the spectrum, to get suppress the noise components. The transfer function of the filter is:

$$H(z) = 0.1z^4 \frac{2 + z^{-1} + z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \tag{2.27}$$

The above filter is best suited for channel effects in the logarithmic spectral domain. To handle both noise and channel effects simultaneously, RASTA filtering is more effective when applied to an equivalent spectrum, $\hat{S}[k]$, computed according to equation:

$$\hat{S}[k] = \log(1 + JS[k]) \tag{2.28}$$

where J is the scaling factor to be found empirically. This procedure is called constant-J-RASTA (CJ-RASTA) processing. PLP cepstrum obtained from CJ-RASTA filtered spectrum is called CJ-RASTA-PLP.

An explicit modeling of the temporal dynamics of the speech has been shown to be useful in improving the noise robustness (MCMS feature, (Tyagi *et al.*, 2003)). Likewise, recently introduced features such as TRAP (Hermansky, 2003) and FDLP (Athineos and Ellis, 2003), modeling the temporal trajectories, also provide a good scope for improving the noise robustness.

2.4.3 Noise masking

Noise masking is a psychological phenomenon observed in humans where the perceptibility of the signal is reduced in the presence of noise, to decrease the effect of the noise (Moore, 1997). As a result of the masking, acoustic stimuli lower than certain threshold, fixed adaptively based on the noise level, cannot be perceived. Based on our knowledge of perception, this involves reduction of contribution of the lower energy regions of the spectrum during recognition process. Employing this idea in the ASR system, a simple noise flooring (Klatt, 1976), and its extension in the HMM framework (Varga and Ponting, 1989) were shown to provide improved noise robustness. Noise masking in logarithmic spectral domain and cepstral domain has also been tried (Mellor and Varga, 1993).

Spectral root homomorphic deconvolution schemes introduced in (Lim, 1979) perform a root operation instead of logarithmic operation on the spectral values, before transforming them to the cepstral domain. An appropriate root value for the root operation, in effect, relatively emphasizes and deemphasizes the peaks and valleys respectively. In (Alexandre *et al.*, 1993; Lockwood and Alexandre, 1994) the use of root-MFCC features derived using such technique were shown to improve the noise robustness.

2.5 Databases and experimental setup

The speech database used for experimental evaluation of noise robust techniques developed in this thesis is OGI Numbers95 database (Cole *et al.*, 1995). For noisy speech recognition experiments, different types of noises are added with clean speech database, as explained in subsection 2.5.2. A few critical experiments of the thesis are also repeated on an alternative database, called Aurora-2 (Hirsch and Pearce, 2000), which is used widely by the noise robust speech recognition community. The description of Aurora-2 database is given in 8. The description of the OGI Numbers95 database and noise databases are as follows:

2.5.1 OGI Numbers95 database

The OGI Numbers95 database (Cole *et al.*, 1995) consists of naturally spoken connected digits, pronounced by American English speakers. The utterances were recorded over the telephone and are hand-labeled with phonetic transcriptions by trained phoneticians. It has a lexicon of 30 words, and 27 different phonemes.

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 3233 utterances comprising approximately three hours of speech. 20% of the training utterances are used as cross-validation data. The test consists of 1206 utterances.

2.5.2 Noise data

For noisy speech experiments, different noises are added with the clean speech utterances from OGI Numbers95 database. The noise types considered are factory and lynx (helicopter) noise from Noisex92 database (Varga *et al.*, 1992) and car noise from a database supplied by Daimler Chrisler Inc. (reported in this thesis as 'car'). These noises are added with clean speech at various SNR levels as follows: 18dB, 12dB, 6dB, and 0dB.

2.5.3 Experimental Setup

The main speech recognition system used for the experiments is HMM/GMM based system (Rabiner and Juang, 1993). It consists of 80 triphones, 3 left-to-right states per triphone, and 12 mixture HMM to estimate the emission probability within each state. Training is performed using HMM toolkit (HTK) (Young *et al.*, 1992). In some of the experiments HMM/ANN based system is used for experiments. It consists of a discriminatingly trained MLP, that typically takes 9 contextual input, and has 27 output units corresponding to the number of context-independent phonemes.

2.6 Conclusion

In this chapter, we have given a brief introduction to the state-of-the-art ASR systems. and gave a comprehensive coverage of the prominent noise robust techniques developed in the past.

The work of the current thesis is explained starting from the next chapter. As explained before the noise robustness techniques developed in this thesis are based on a central idea that emphasizing the part of the speech that is more invariant to noise and/or deemphasizing the part that is more sensitive to noise would result in improved noise robustness.

Chapter 3

Spectral Peak Location Estimation

The noise robust techniques developed in this thesis are based on transformations that give higher emphasis to the relatively noise invariant parts of the speech and deemphasis or mask the noise sensitive parts of the speech. It is well known that, in the spectral domain, peaks correspond to part of the speech signal with relatively high SNR and valleys correspond to part with relatively low SNR. Thus a simple emphasis of spectral components corresponding to peaks and/or masking of spectral components corresponding to valleys is expected to yield an improved noise robustness.

3.1 Introduction

The first problem that is needed to be solved in such a formulation for noise robustness is estimation of the locations of the spectral peaks and valleys. Although this appears to be a simple problem while looking at the typical spectra of the speech signal, in actual, it is a hard problem to solve. The reasons for this include various disturbing factors such as the differences in the number of peaks and valleys across different phonemes, variability in their relative energy levels across phonemes and in the presence of noise, variability in their frequency locations across speakers, the presence of pitch information, and the existence of spurious spectral peaks in the presence of noise. On the other hand, upon the successful estimation of the spectral peak locations, in addition to using them for improving the noise robustness, they can as well be used as an additional source of information (features) for speech recognition. The regions around the spectral peaks are known commonly in speech recognition literature as the formants. They represent the resonances of the vocal tract cavity and hence are the immediate source of articulatory information.

There have been several methods proposed in the literature to estimate the spectral peaks. Among them, a relatively old method that received large attention in the speech recognition community is linear prediction (LP) (McCandless, 1974; Kopec, 1986; Makhoul, 1975). In LP, the speech signal is assumed to have been generated by an all-pole model. The poles of the model give a measure of the spectral peak locations. A recursive algorithm, called Durbin algorithm (Rabiner and Juang, 1993), estimates the poles such that the spectral peaks of the model fits optimally with the spectral peaks of the signal. Although LP is quite successful and used widely, there are a few disadvantages. The number of poles for the model, fixed apriori, restricts the number of peaks to be identified. When the actual number of peaks differ from the number of peaks to be estimated, the algorithm may lead to an erroneous estimation. In the presence of a spurious spectral peak, which is usually the case in the presence of the noise, the algorithm will try to take that also into account, leading again to erroneous estimation of the peak locations.

In a recent work, an algorithm based on parallel digital resonator model and dynamic programming (Welling and Ney, 1998), has been shown to yield robust estimation of the peak locations (formant frequencies). Imposing temporal constraints on the peak locations estimated, leading to an estimation of threaded spectral peaks, has been shown to be useful in improving the noise robustness (Strope and Alwan, 1998).

Recently, a new acoustic model, called HMM2, has been used for spectral peak location estimation (Weber *et al.*, 2002, 2003a). HMM2 is basically an alternative form of regular HMM obtained by replacing the emission modeling GMMs or MLP with a set of state-dependent HMMs, called internal HMMs or frequency HMM2 (Bengio *et al.*, 2000; Weber, 2003). For the spectral peak location estimation task, the HMM2 has been applied directly to the spectrum. A fixed number of peak locations estimated, referred as formant-like features, when used along with the traditional features, has been shown to yield an improvement in the speech recognition performance.

Our approach

In this chapter, we explain two new approaches for peak location estimation. The first approach is a simple frequency-based dynamic programming (DP) algorithm, acting as a filter, taking spectral slope values of single time frames as the input and yielding estimated peak locations as output (Ikbal *et al.*, 2004b). The second approach is an extension of the frequency-based DP algorithm, using frequency based HMM/ANN, that makes use of distinct time-frequency patterns in the spectrogram to estimate peak locations. Such use of time-frequency patterns imposes temporal constraints during the peak location estimation, thereby yielding a smoother estimate of the peak locations over time (Ikbal *et al.*, 2004d). Both the approaches are basically motivated by a previous work, where a HMM employed along frequency axis, in a general framework called HMM2, has been used for estimating a fixed number of spectral peaks. First the HMM2 based peak location estimation is explained in the next section.

3.2 HMM2

HMM2 has originally been introduced as an alternative acoustic model to the regular HMM (Bengio *et al.*, 2000; Weber *et al.*, 2000; Weber, 2003). However, later it has also been shown to be useful in spectral peak location estimation (Weber *et al.*, 2002, 2003a). Although, estimation of spectral peak locations using HMM2 is topic of relevance to the context of this chapter, the acoustic modeling aspects of HMM2 is also explained in the next subsection, for a smooth following of the further explanations.

3.2.1 Acoustic modeling by HMM2

HMM2 is obtained from the regular HMM by replacing the state-dependent emission modeling GMMs with a set of state-dependent HMMs called frequency HMMs[1] (Bengio *et al.*, 2000; Weber *et al.*, 2000). An illustration of HMM2 is given in 3.1. These frequency HMMs treat the feature vectors as fixed-length sequences and estimate the emission probability by calculating the likelihood of those feature vectors being generated by them. For this purpose, each feature vector is converted into a sequence of smaller vectors called frequency vectors, as illustrated in the figure 3.1. The states of the frequency HMM, called frequency states, are assumed to have emitted those frequency vectors. The emission of the frequency vectors by frequency states is governed by

[1]In the previous literature the frequency HMMs were also called the internal HMMs. But through out this thesis the name frequency HMM is used as it suits the present context (of spectral peak estimation) well. Additionally, in order to distinguish from the frequency HMMs, the temporal HMM state sequence is referred to as the temporal HMM.

a lower dimensional emission probability model (lower dimensional GMM) assigned to each frequency state. The complete parameter set of the HMM2 includes, the transition probabilities of all the temporal states, transition probabilities of all the frequency states of every temporal state, and the parameters of GMMs assigned to every frequency state.

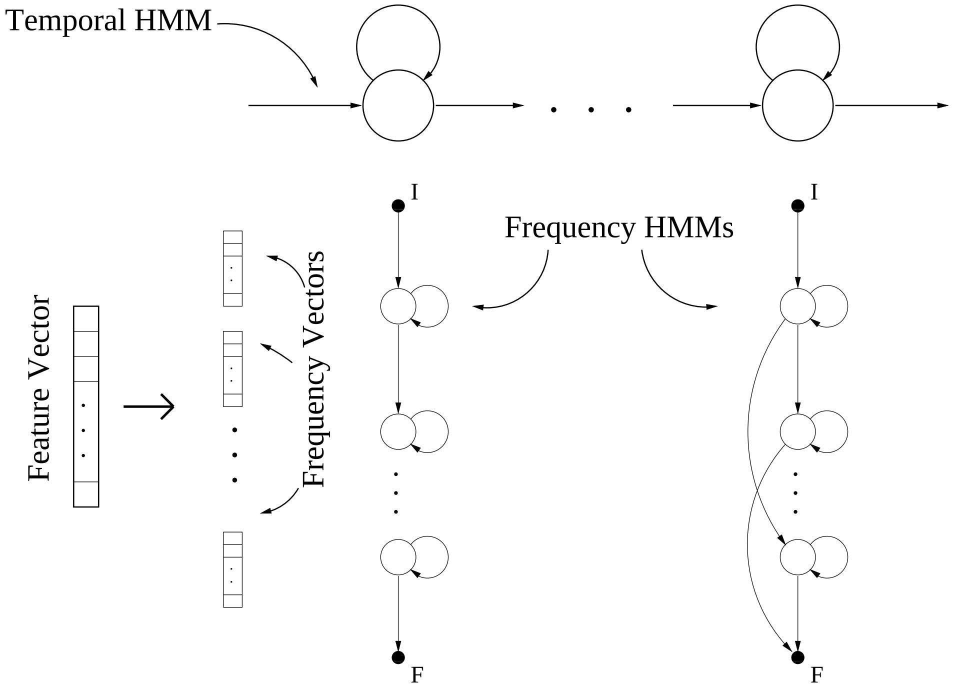

Figure 3.1. Illustration of the HMM2.

If $\{\mathbf{x}_{t,0}, \mathbf{x}_{t,1}, ..., \mathbf{x}_{t,s}, ..., \mathbf{x}_{t,S-1}\}$ denote the frequency vector sequence derived from $\mathbf{x}_t$, then the likelihood of a sample frequency state sequence $R = \{r_0, r_1, ..., r_s, ..., r_{S-1}\}$ of the frequency HMM belonging to the temporal state q_t, generating the vector sequence is:

$$p(\mathbf{x}_t, R|q_t) = p(r_0|q_t)p(\mathbf{x}_{t,0}|r_0, q_t) \prod_{s=1}^{S-1} p(r_s|r_{s-1}, q_t)p(\mathbf{x}_{t,s}|r_s, q_t) \tag{3.1}$$

Based on the topology of the frequency HMM for temporal state q_t, if $\mathcal{D}_t$ represents the set of all possible frequency state sequences, then the emission probability calculated using the frequency

HMM is:

$$p(\mathbf{x}_t|q_t) = \sum_{R \in \mathcal{D}_t} p(\mathbf{x}_t, R|q_t) \tag{3.2}$$

Alternatively, the emission probability can also be computed using the Viterbi approximation as,

$$p(\mathbf{x}_t|q_t) = \max_{R \in \mathcal{D}_t} p(\mathbf{x}_t, R|q_t) \tag{3.3}$$

A full derivation of an EM algorithm to estimate the parameters of the HMM2 from the training data set is given in (Bengio *et al.*, 2000). An explanation of training and recognition based on HMM2, from the implementation viewpoint is given in (Ikbal *et al.*, 2001).

HMM2 was introduced as an alternative to the regular HMM for acoustic modeling of the speech, with expectations of possessing several potential advantages. The regular HMM has capability to handle the temporal variabilities observed in the feature vector sequences, arising out of the inherent difference in the rates at which speakers generate the speech. This is possible by an appropriate alignment of the temporal states across the vector sequence. In addition to possessing the ability to time warp, the HMM2 can handle the variabilities observed in the frequency axis also, when applied to the spectrum, by an appropriate alignment of the frequency states along the frequency vectors. Variabilities along frequency axis of the power spectrum typically arise because of the speaker differences. The differences in the speech production mechanisms of the speakers result in nonlinear warping of the spectral frequency axis. HMM2 can be used to handle this if utilized accordingly (Ikbal *et al.*, 2002).

In spite of the potential advantages, HMM2 has not been shown to be successful for the speech recognition task, so far. A systematic analysis of the modeling difficulties that arise while using the HMM2 for the acoustic modeling task is given in (Weber *et al.*, 2001). However, HMM2 has been shown to be useful for an alternative task, namely the feature extraction (Weber *et al.*, 2002). Specifically, it has been shown to be successful for locating the spectral peaks in the spectrum, which is explained in the next subsection.

3.2.2 Feature extraction using HMM2

The capability of the HMM2 to warp its frequency HMM states appropriately along the frequency vectors gives it an additional ability to extract useful information from the spectrum, which can further be used as features for the speech recognition task (Weber *et al.*, 2002, 2003a). Assume that the states of a frequency HMM have been trained one each over certain specific regions of the spectrum, for example the peaks and valleys, to learn their statistics. A Viterbi alignment of this frequency HMM against a test spectrum would yield an alignment of its states across the frequency axis so that an optimal match happen between the statistics learned by the frequency states and the regions to which they are aligned. The alignment basically yields an information about the optimal segmentation of the frequency axis into different regions, the peaks and valleys. The usefulness of such peak location estimates has been evaluated by using them as additional feature along with the regular features. Such an use has been shown to yield an improved speech recognition performance accuracy (Weber *et al.*, 2002, 2003a).

An important point to note here is the fact that the number of peak locations estimated using this HMM2 based method is restricted apriori by a fixed topology used for frequency HMM. However, the two new approaches explained in this chapter (frequency-based DP algorithm and the HMM/ANN based algorithm) are tailor-made for use in the noise robust techniques developed in the later chapters of this thesis, and differs from the HMM2 based peak location estimation in the following manner: The number of peak locations estimated by these algorithms are not restricted to be fixed in number. The reason for this is explained in the next section. Later on, the frequency-based DP algorithm and the HMM/ANN based algorithm are explained.

3.3 Fixed vs variable number of spectral peak locations

One of the important factors to be considered while designing the peak location identification algorithm is the number of peak locations to be identified. Most of the algorithms for peak location estimation, fix apriori, the number of peaks to be identified. This some times lead to erroneous estimation, because of the difference in the number of peak locations for different phonemes. However, the main purpose of the peak identification task in the current thesis is to use the peak location information to improve the noise robustness. As explained before, the strategy followed for this is to enhance the parts of the speech corresponding to spectral peaks and deemphasis the spectral valleys. In such a formulation, a wrong identification of the peak location would lead to enhancement of the non-peak locations and hence would lead to undesirable results. Thus identifying as many available peaks locations in the spectrum is preferable.

In the next section the frequency-based dynamic programming algorithm is explained.

3.4 Frequency-based dynamic programming (DP) algorithm

Figure 3.2 shows a fully-connected two state sequence that is assumed to have emitted the spectrum. The emission of spectral energy value at any point in the frequency axis is assumed to be governed by the slope of the spectral energy at that point with respect to frequency axis. The likelihood of the first state emitting spectral energy value is equal to the slope of the spectrum along the frequency axis. Likewise, the likelihood of the second state emitting the spectral energy value is equal to the negative value of the slope. In such a case, a Viterbi alignment of these two states along the frequency axis of the spectrum would yield high score only when the first state is aligned to the positive sloped regions of the spectrum and the second state is aligned to the negative sloped regions. Thus the points of transitions in the Viterbi alignment from the first state to the second state would constitute a spectral peak.

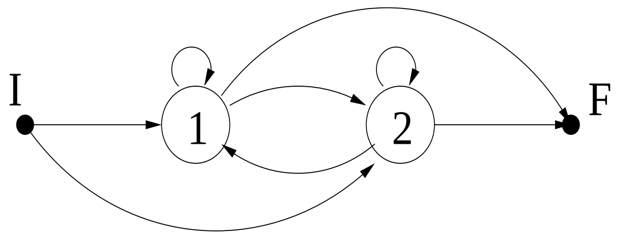

Figure 3.2. Frequency-based dynamic programming (DP) algorithm to locate the spectral peaks.

This algorithm is simple, however not enough, as the spectrum also has pitch information which introduces a series of small pitch related peaks and dips throughout the frequency range. Furthermore, in the presence of noise, spurious peaks start to appear in the spectrum. The algorithm in this case would identify all such peaks as spectral peaks. Additionally, a genuine peak may also be missed if there is a pitch dip exactly at the peak location. These problems can be handled to some extent by smoothing out the spectrum by low-pass filtering or by cepstral smoothing, where the spectrum is reconstructed from its lower order cepstral coefficients. Interestingly, mel-warped critical bank spectrum usually does not have the pitch related peaks as the critical band filtering acts as a low-pass filter suppressing the pitch related information. Throughout this thesis mel-frequency critical band spectrum is used as the spectral representation. Another way of avoiding spurious and pitch related peaks is given in given next subsection.

3.4.1 Minimum duration constraint

Apart from spectral smoothing, the pitch related peaks and spurious peaks can also be handled by imposing minimum duration constraints for states of the frequency-based DP algorithm, as illustrated in Figure 3.3. Imposing minimum duration constraint avoids identification of peaks of duration less than the minimum duration. The pitch related and spurious peaks are generally of smaller widths. Hence, imposing minimum duration of lengths slightly larger than their widths would avoid identification of such peaks. In our case, since the pitch related peaks are smoothed out in the mel-warped critical band spectrum, just to avoid spurious peak, a minimum duration of value equal to 2 is used.

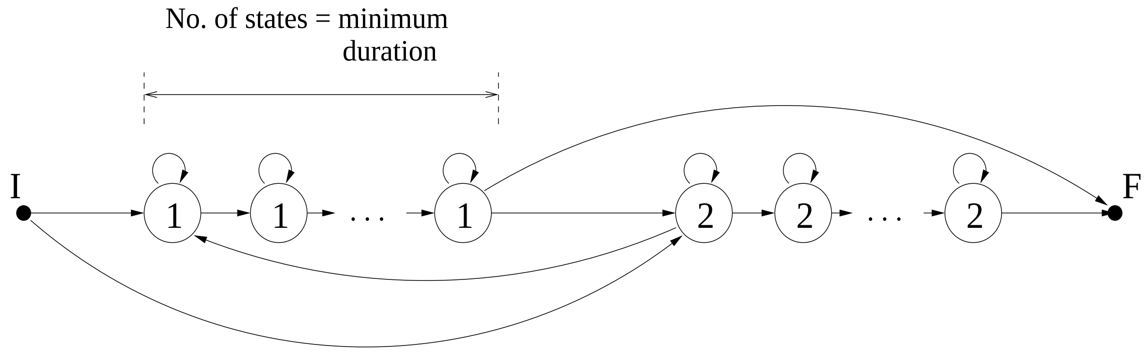

Figure 3.3. Frequency-based DP algorithm with minimum duration constraint to locate the spectral peaks.

3.4.2 Peak location estimation

Figure 3.4 shows an example mel-warped critical band spectrum of phoneme 'ih', and the estimated peak locations by the above algorithm. The minimum duration used for the states in the above algorithm is 2.

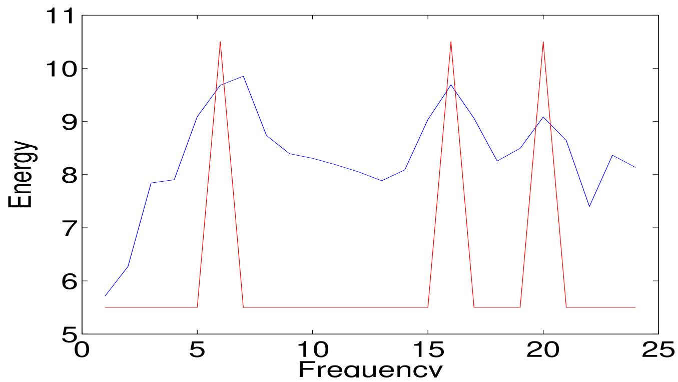

Figure 3.4. Spikes show the locations of peaks identified in an example mel-warped critical band spectrum corresponding to phoneme 'ih', by the frequency-based DP algorithm.

Figure 3.5 shows mel-warped critical band spectrogram of a speech sample utterance taken from the OGI Numbers95 database. Figure 3.6 shows locations of the peaks identified by the above explained algorithm. It can also be seen from the figures that there is a close resemblance between the actual spectral peak trajectories and the trajectories of the peak locations identified by the algorithm, especially in the speech regions of the spectrum.

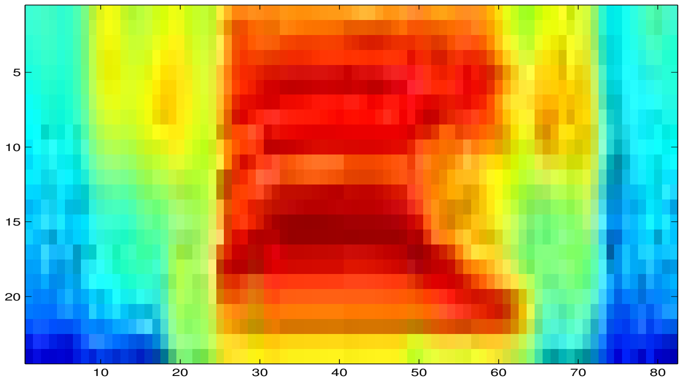

Figure 3.5. Mel-warped critical band spectrogram of a sample speech utterance taken from OGI Numbers95 database.

A close examination of the peaks identified will reveal that the number of peaks identified in the spectrogram is not constant over time. This is as a result of the ergodic model of the DP algorithm in Figures 3.2 and 3.3. This provides it an ability to locate as many available peaks in the spectrum, satisfying only the constraint of minimum duration. The number of peaks estimated varies from

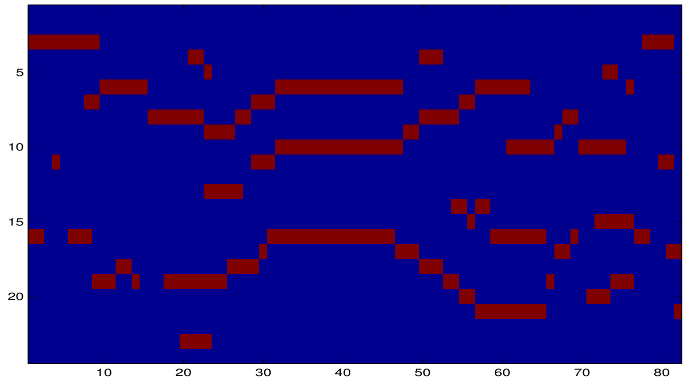

Figure 3.6. Peak locations identified from the mel-warped critical bank spectrogram of a sample speech utterance (of the Figure 3.5), by the frequency-based DP algorithm.

two to five. An additional observation from the figure is that the peaks identified in the speech regions form a smooth trajectory over time, whereas the peaks in the silence regions do not form smooth trajectories characterizing their behavior.

As our final goal is to use the peak location information to improve the noise robustness, an important case to consider is the peak location estimation by the DP algorithm in noisy speech. Figure 3.7 shows mel-warped critical band spectrogram of the sample speech utterance, of the Figure 3.5, added with factory noise from Noisex92 database at 6dB SNR. The results of the peak identification on this noisy spectrogram by the DP algorithm is given in Figure 3.8. As can be seen from the figure, the peak locations estimated are disturbed and not same as the one in Figure 3.6. However, there is a close resemblance in the speech regions of the spectrum.

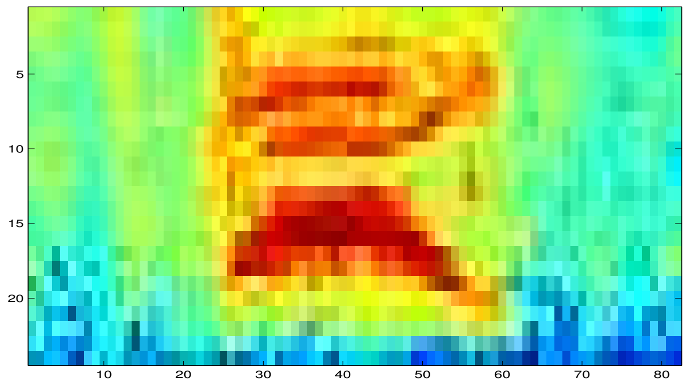

Figure 3.7. Mel-warped critical band spectrogram of the sample speech utterance (same as in Figure 3.5) taken from OGI Numbers95 database corrupted by factory noise from Noisex92 database at 6dB SNR.

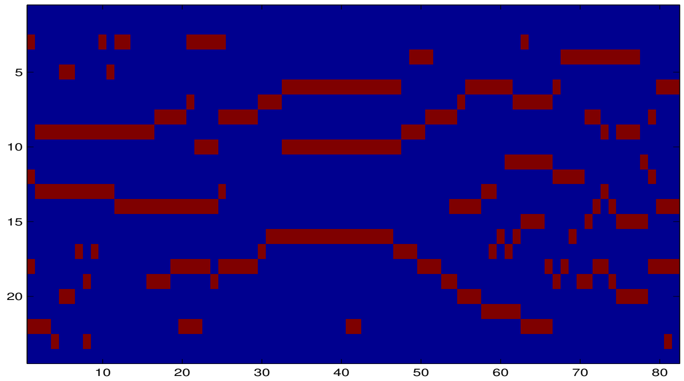

Figure 3.8. Peak locations identified from the mel-warped critical bank spectrogram of the noisy speech utterance (of the Figure 3.7), by the frequency-based DP algorithm.

3.4.3 Extension of the DP algorithm - Learning distinct regions

The frequency-based DP algorithm explained above estimates the spectral peaks based on the spectral slope values of single time frame. It basically discriminates between the positive and negative sloped regions of the spectrum, along with minimum duration constraints, in order to locate the peaks. An interesting extension of this algorithm is to make the states learn and discriminate between, more general, time-frequency (TF) patterns in the spectrogram and use them for locating the peaks. The use of TF patterns is expected to impose temporal constraints during the peak location estimation, thereby yielding a smoother estimate of peak locations over time. As explained in the next section, HMM/ANN (Bourlard and Morgan, 1993) provides a nice framework for doing this, as the MLP used in HMM/ANN for state emission modeling has demonstrated ability to effectively handle the temporal context.

3.5 HMM/ANN based algorithm

The use of HMM/ANN for spectral peak estimation is basically motivated by the HMM2 based peak identification, explained in the earlier sections of this chapter. As we have seen, HMM2 uses frequency HMMs to locate the spectral peaks. In the current case, a simple HMM/ANN is employed along the frequency axis of the spectrum for estimating the peak locations. As well known, HMM/ANN use multi-layer perceptron (MLP) for emission modeling (in contrast to the HMM2 case, where in the frequency HMMs, GMMs for the same task). The use of MLP provides an additional flexibility to use more general, time-frequency (TF) patterns in the spectrogram, as shown in Figure 3.9, for the peak estimation task. This is because the MLP has been shown to be more effective in handling the temporal contextual information (Bourlard and Morgan, 1993). The inclusion of such temporal contextual information is expected to impose temporal constraints during the peak identification, which in turn is expected to result in smoother estimates of the peak locations over time. This is not the case in the previously explained frequency-based DP algorithm which considers only single frame for the peak estimation task, thus there is a possibility of unrealistic variation in the estimated peak locations from one frame to the other.

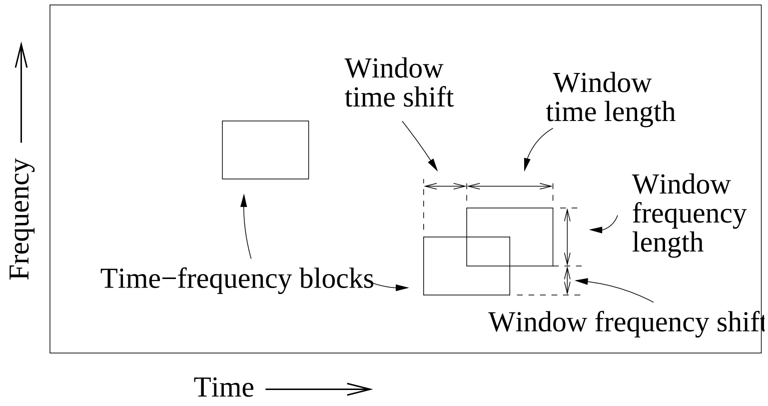

Figure 3.9. Illustration of time-frequency blocks as seen by the HMM/ANN states in the spectrogram.

3.5.1 Strategy

The strategy used in HMM/ANN based algorithm for peak location estimation is similar to the one used for frequency-based DP algorithm. In DP algorithm, distinct positive and negative sloped regions in the spectrum are identified in order to estimate the peak locations. In the current case, instead of using single time frame spectral energy values, the aim is to use, more general, TF pattern in the spectrogram to estimate the peak locations. To achieve this first of all, the HMM/ANN states should learn distinct TF patterns in the spectrogram. Suppose, that the topology of the HMM/ANN is the same as the one used for frequency-based DP algorithm given in Section 3.4. Now, suppose along the frequency axis, the TF patterns before the spectral peaks are modeled by the first state and the TF patters after the spectral peaks are modeled by second state. With this, a

Viterbi alignment of the HMM/ANN along the frequency axis will yield peak locations as points of transition from the first state to the second state. However the training of HMM/ANN to make it learn distinct TF regions raises a few issues, which are explained in the next subsection.

3.5.2 Issues

As explained in Section 3.4 mel-frequency critical bank spectrum is used for the peak location estimation task. This is because the pitch information is reasonably suppressed in the mel-frequency critical bank spectrum. Additionally, states of HMM/ANN are imposed with minimum duration constraints as shown in Figure 3.3 to avoid spurious peaks locations.

For the training of the HMM/ANN there is no transcription available for discriminating the spectral regions into hypothesized classes, i.e., TF patterns before the spectral peaks (positive sloped TF patterns) and TF patterns after the spectral peaks (negative sloped TF patterns). In this sense, the training of the HMM/ANN needs to be unsupervised. The convergence of such unsupervised training into segmentation of hypothesized regions is not always guaranteed. However, an use of slope spectrum facilitates this to a certain extent. Additionally, the topological constraints of the HMM/ANN along with minimum duration constraints, as given in Figure 3.3, is expected to further facilitate the convergence. The peak identification results given in subsection 3.5.4 indeed show that MLP training converges to the classification of spectrum into hypothesized regions.

The implementation details of HMM/ANN used for peak location estimation task is given in the next subsection.

3.5.3 Implementation details

As mentioned in previous subsections, the topology of the HMM/ANN is same as in the Figure 3.3. The emission modeling for the states is performed using MLP of following specifications: The input layer size is same as the TF pattern size used. For example, if the time and frequency widths of TF pattern are w_t and w_f respectively, then the input layer size is $w_t \times w_f$. The output layer size is 2 corresponding to the number of distinct HMM/ANN states. Hidden layer size is fixed at 20-50. Mel-frequency critical bank spectrum is used as the spectral representation. The minimum duration used for the states in the Figure 3.3 is 2.

3.5.4 Peak location estimation

Assuming the HMM/ANN has been trained on the spectrogram of several utterances, the main factor that affects the peak location estimation performance is the size of the TF blocks. In order to have a comparison with the results of frequency-based DP algorithm, for the first case, single coefficients (i.e., $w_t = 1$ and $w_f = 1$) are considered as the TF patterns. As mentioned before minimum state duration used is 2. The equivalent results of the Figures 3.4 and 3.6 for the case of HMM/ANN are given in Figures 3.10 and 3.11, respectively. As can be seen from these figures, when compared to the frequency-based DP algorithm, the HMM/ANN based peak estimation seems to miss a few peak locations. In the Figure 3.10 one of the prominent peak locations has been missed when compared to the Figure 3.4. The Figure 3.11, seems to have located prominent peak locations in the Figure 3.5. However, low energy level peaks are missed out.

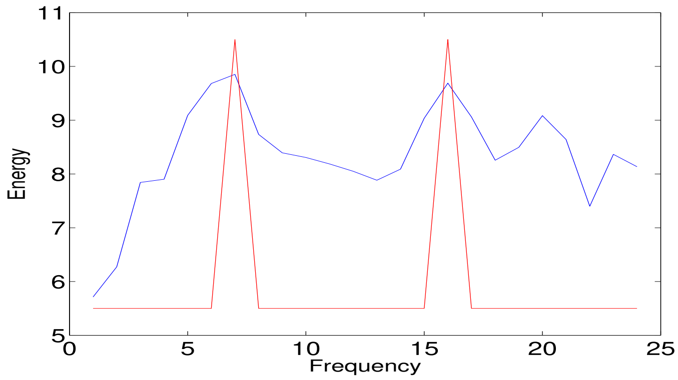

Figure 3.10. Spikes show the locations of peaks identified in an example mel-warped critical band spectrum corresponding to phoneme 'ih', by the HMM/ANN based algorithm, when the time-frequency block of size $w_t = 1$ and $w_f = 1$ is used.

The equivalent result of the Figure 3.8 for noisy speech is given in Figure 3.12. The peak locations estimated in this case also has picked noisy peaks as can be seen for the Figure 3.8. However, from our arguments in the previous sections, increasing the size of TF block is expected to result in smoother estimates of peak locations over time, which is checked in the next case.

Figures 3.13 and 3.14 show plots of peak locations estimated in clean and noisy speech spectrograms when the TF pattern size is: $w_t = 3$ and $w_f = 1$. Similarly, Figures 3.15 and 3.16 give the peak locations estimated in clean and noisy speech spectrograms when the TF pattern size is: $w_t = 5$ and $w_f = 2$. In actual, it is hard to compare these figures visually and draw much conclu-

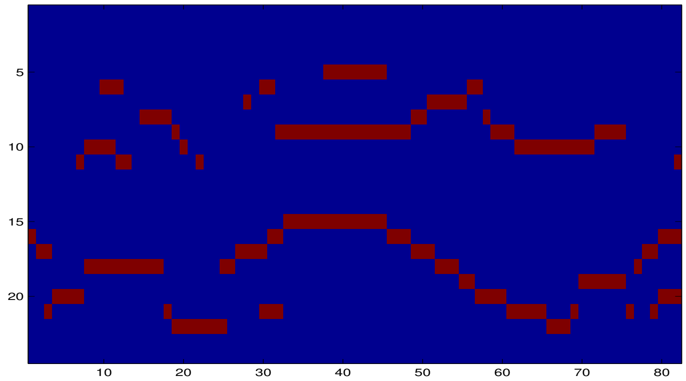

Figure 3.11. Peak locations identified from the mel-warped critical bank spectrogram of a sample speech utterance, by the HMM/ANN based algorithm, when the time-frequency block of size $w_t = 1$ and $w_f = 1$ is used.

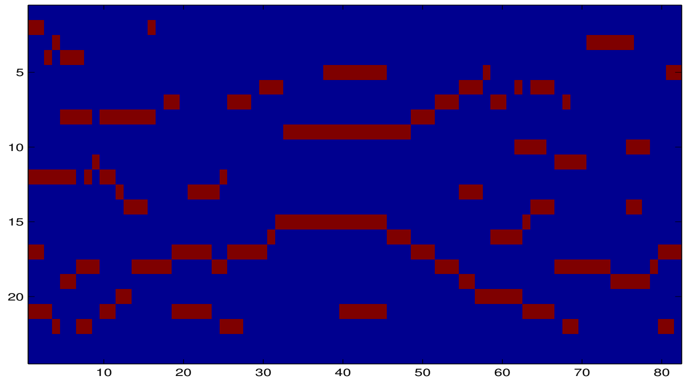

Figure 3.12. Peak locations identified from the mel-warped critical bank spectrogram of the noisy speech utterance, by the HMM/ANN based algorithm, when the time-frequency block of size $w_t = 1$ and $w_f = 1$ is used.

sions. However, one major conclusion that can be made is the fact that with the increase in the size of the TF pattern used for peak identification, locations estimated seems to get more constrained, i.e., only very prominent peaks seem to get identified. Also the temporal trajectories of such peaks estimated are more smooth across time. Comparing the Figures 3.12, 3.14, and 3.16, this is true for the peak location estimation in the noisy speech spectrogram also.

The actual evaluation of reliable information carried by these peak locations estimated is performed in the next chapter where such peak locations are used to compute a noise robust feature representation. As mentioned in the early sections of this chapter, the main purpose of the peak location estimation algorithms developed in this chapter is to use such information to develop noise robust feature representations.

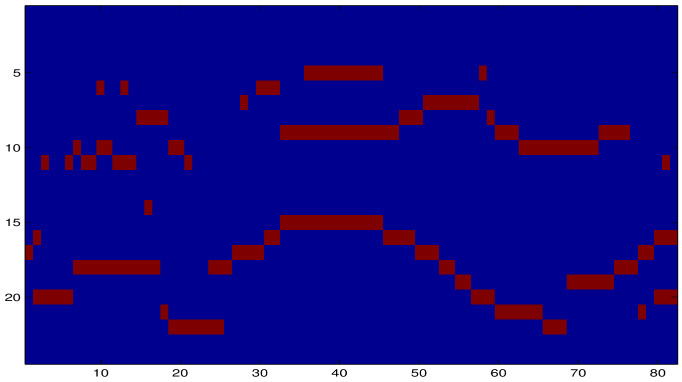

Figure 3.13. Peak locations identified from the mel-warped critical bank spectrogram of a sample speech utterance, by the HMM/ANN based algorithm, when the time-frequency block of size $w_t = 3$ and $w_f = 1$ is used.

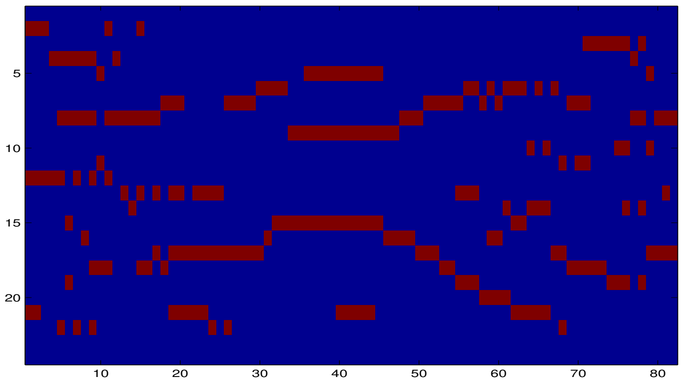

Figure 3.14. Peak locations identified from the mel-warped critical bank spectrogram of the noisy speech utterance, by the HMM/ANN based algorithm, when the time-frequency block of size $w_t = 3$ and $w_f = 1$ is used.

3.6 Conclusion

In this chapter we have developed two different algorithms for estimating the spectral peak locations. Both the algorithms are motivated by a previous work, where frequency based HMMs, in a general framework called HMM2, have been shown to be successful for estimating a fixed number of spectral peaks. The first method, developed in this chapter, referred to as frequency-based dynamic programming (DP) method, use spectral slope values of single time frame to estimate peaks. Whereas, second method, referred to as HMM/ANN based peak estimation algorithm, use, more general, time-frequency (TF) patterns in the spectrum for such task. The use of TF patterns is

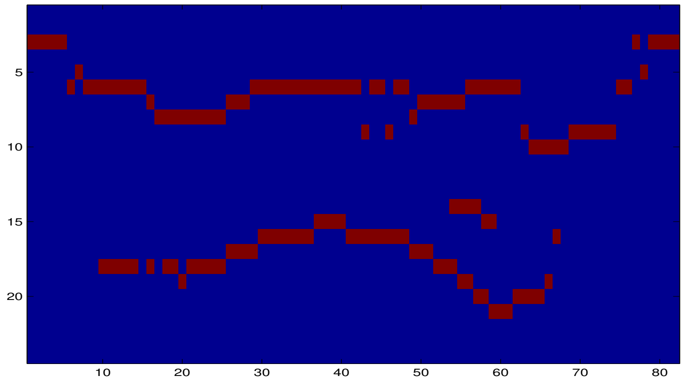

Figure 3.15. Peak locations identified from the mel-warped critical bank spectrogram of a sample speech utterance, by the HMM/ANN based algorithm, when the time-frequency block of size $w_t = 5$ and $w_f = 2$ is used.

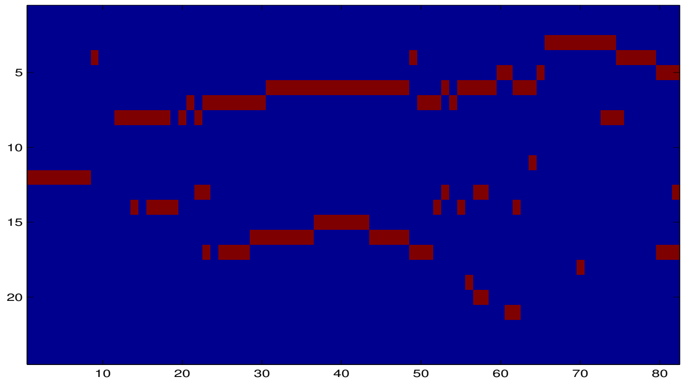

Figure 3.16. Peak locations identified from the mel-warped critical bank spectrogram of the noisy speech utterance, by the HMM/ANN based algorithm, when the time-frequency block of size $w_t = 5$ and $w_f = 2$ is used.

expected impose temporal constraints during the peak location estimation. A few plots of peak locations estimated in an example spectrogram were given. However, it is very difficult and, in fact, not valid to draw conclusions based on such plots. The evaluation of such peak locations estimated is actually performed in the next chapter, where the peak locations estimated are used to compute noise robust feature representations. A final point to mention about the peak estimation algorithm developed is this chapter are: These algorithms distinguish themselves from most of the previous algorithms developed in the literature by the fact that the number of peak locations estimated are not restricted to fixed number. This is a tailor-made aspect of these algorithms to them suitable for the noise robust feature extraction algorithms developed in the next chapter,

Chapter 4

Spectro-Temporal Activity Pattern (STAP) features

4.1 Using peak location information to improve the noise robustness

Assuming the availability of the spectral peak location information, the next problem to look upon is: How to use such information to improve the noise robustness. As mentioned in the previous chapters, our strategy in this thesis to improve the noise robustness is motivated by a perceptual phenomenon observed in the human auditory processing system called noise masking (Moore, 1997). As a result of noise masking, information that are unreliable are masked or discarded while recognizing the sound in the presence of noise.

Another interesting aspect of the human auditory system is the processing of local time-frequency patterns in the incoming signals during the process of recognizing sounds. Physiological studies conducted on mammalian auditory cortex show evidences for recognition of local spectro-temporal patterns in the signal by auditory cortical neurons (Depireux *et al.*, 2001). This is quite contrast to the feature extraction schemes which pay attention to the spectral representations over the entire span of the frequency axis and over a limited span of the temporal axis. For example, standard features used for speech recognition, such as MFCC or PLP cepstrum typically represent the spectral envelope of a short segment of the speech signal and the temporal characteristics are typically modeled in a relatively weak manner through the use of derivatives (Furui, 1986) of the static features or through the use of temporal contextual information as done in HMM/ANN system (Bourlard and Morgan, 1993). On the other hand, few recently proposed successful features for speech recognition such as TRAPS (Hermansky, 2003), MCMS (Tyagi *et al.*, 2003), and FDLP (Athineos and Ellis, 2003), mainly represent the temporal characteristics of the speech, and in this case, the frequency characteristics are modeled by feeding in the recognizer with temporal patterns extracted from the entire span of the frequency range.

In this chapter we develop a new feature extraction approach, inspired by the two above mentioned interesting aspects of the human auditory processing (noise masking and local time-frequency pattern processing), and explore its effectiveness for noise robustness.

4.2 Parameterizing the information around spectral peaks

The regions around the spectral peaks are less sensitive to noise as they constitute the part of the signal that has high SNR. On the other hand, spectral valleys constitute low SNR part of the speech, and hence are more sensitive to the noise. Thus a simple masking of the spectral coefficients in the non-peak regions of the spectrum is expected to result in improved noise robustness. The resulting feature vector in this case would constitute spectral energy values only from the regions around the spectral peak locations. Now based on the knowledge from the local time-frequency processing by the human auditory system, a better scheme to model these regions around the spectral peaks is to parameterize the local time-frequency patterns around the spectral peaks. The resulting feature vector in the case will constitute parameters describing the activity within local time-frequency patterns (i.e., the energy surface of the patterns) around the spectral peaks. We refer to this parameterization as spectro-temporal activity pattern (STAP) features.

Time-frequency parameterization

The parameters that are actually considered for describing the activity within the local time-frequency patterns are the following:

1. Frequency index of the peak location, which is also the center point of the time-frequency

pattern, denoted by L.

2. Energy level at the peak location or the average energy level of the whole time-frequency pattern around the peak location, denoted by E.

3. Delta of energy within the time-frequency pattern around the peak location along the time axis, denoted by $\Delta_t E$.

4. Acceleration of the energy within the time-frequency pattern around the peak location along the time axis, denoted by $\Delta_t^2 E$.

5. Delta of energy within the time-frequency pattern around peak location along the frequency axis, denoted by $\Delta_f E$.

6. Acceleration of energy within the time-frequency pattern around the peak location along the frequency axis, denoted by $\Delta_f^2 E$.

4.3 STAP feature

The STAP features use one or more of the above described parameters extracted from the local time-frequency patterns around the spectral peaks as its feature components. Such use of information only from the regions around the spectral peaks is expected to result in improved noise robustness. However, as a result of the masking of the non-peak spectral components, which also carry information for clean speech recognition, an inferior clean speech recognition performance is expected.

An important case to take a look at in case of STAP features is: as the peak identification algorithm can yield varying number of peak locations over time, the total number of time-frequency patterns considered for parameterization changes over time. This leads to a STAP feature sequence whose dimension change over time. However, the conventional speech recognition systems, which can handle only uniform dimensional feature sequence, can not handle the STAP features in this form. Thus they need to be converted into uniform dimensional features, some how.

4.3.1 Uniform dimensional STAP features

A simple method by which STAP features can be converted into uniform dimensional vectors is to assign zeros to the parameters describing the time-frequency patterns around the non-peak locations, and include them in the feature representation (in other words, masks, whose values are non-zeros only at the peak locations, are applied on the complete time-frequency representation of the parameters). For example, the part of the uniform dimensional STAP feature corresponding to the parameter E is obtained by masking non-peak locations in the spectrogram to zeros. Figure 4.1 shows a sequence of such uniform dimensional feature vectors, carrying just the E information, computed using spectrogram shown in the Figure 3.5 and peak locations shown in the Figure 3.6. These features, in fact, have both L and E information, as the frequency index of the peak locations are also encoded in them. In a similar manner, part of the feature corresponding to $\Delta_t E$ is obtained by applying the mask on the delta spectrogram, and so on. As we can see from the figure 4.1, this way of parameterization introduces a large number of zeros in the feature vector. In actual, though the feature dimension appears to be large, the actual dimension of the useful part of the feature is very small, corresponding to the number of peak location identified, which is typically 2-5.

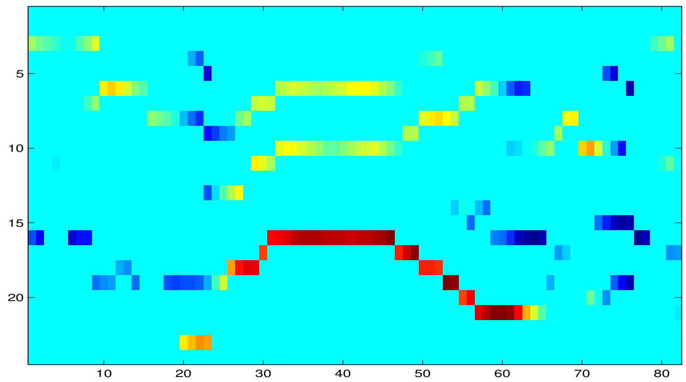

Figure 4.1. Sequence of STAP features where only L and E information are used. The differences in the intensity level indicate the differences in the energy.

4.3.2 STAP features dimension

STAP features used for experimental purposes are basically extracted from the 24 dimensional mel-warped critical band spectrogram. As we have seen above, uniform dimensional STAP feature has many of its components as zeros. This in addition with minimum duration constraint imposed by the peak picking algorithms (both frequency-based dynamic programming algorithm and HMM/ANN based algorithm in previous chapter) allows a down-sampling of the STAP features. Thus, in our case, each local time-frequency pattern describing parameter contributes 12 dimensions to the final STAP feature. Out of the pattern describing parameters, L and E can be encoded in a single 12 dimensional vector. Hence, the use of all the parameters, as listed in section 4.2, in the STAP feature would make its dimension 60. However, as explained in the previous subsection many components of the feature (typically 35-45) are zeros.

4.3.3 Analogies to missing data approach

The masking of non-peak locations to zeros for obtaining the uniform dimensional STAP features draw analogies between them and the missing data approach, which was explained in the Section 2.3.6. Missing data approach masks the unreliable spectral coefficients in the spectrogram and consider it as missing data. Such missing data is handled during the recognition either by marginalizing over the missing data or by estimating the missing data based on the reliable data (Cooke *et al.*, 2001; Raj *et al.*, 2001). However, in STAP approach regions around non-peak locations are completely discarded and a supposedly better parameterization of the regions around spectral peaks is obtained by computing parameters describing the activity of the local time-frequency pattern around.

4.4 Handling the feature correlation

As explained in the previous section, STAP features are derived from the spectral representations by masking non-peak locations to zeros. This form of feature vector is highly unsuitable for HMM/GMM speech recognition system, because first of all the spectral components are highly correlated and the commonly used diagonal covariance matrix for emission modeling GMMs do not support it. Additionally, the presence of a large number of zero is also unrealistic. Thus the STAP features need to be transformed to a form where the components are decorrelated. A similar example, of unsuitability of HMM/GMM, with GMMs using diagonal covariance, is for the case where mel-frequency critical band spectrum is used as feature vector. However, MFCC obtained by performing DCT on mel-frequency critical band spectrum to decorrelate the feature components suits well for the HMM/GMM systems.

The unsuitability of HMM/GMM, with diagonal covariance, for features with correlation between the components is best illustrated by the mel-warped critical bank spectrum and its linearly transformed version MFCC. MFCC is better suited for HMM/GMM system than its spectral equivalent, because DCT performed to obtain MFCC decorrelates the feature components to some extent.

A commonly used tool, in literature, that is more better suited for the decorrelation of STAP features is principle component analysis (PCA). However, since the transformation involved in PCA is computed in an unsupervised manner, the speech specific aspects of the STAP feature may have less influence in governing the transformation computation. An alternative tool is linear discriminant analysis (LDA), which finds out a set of orthogonal bases for transformation sub-space along the direction of maximum possible speech class discriminatory information. A previous work where LDA has been shown to be successful for speech recognition task can be found in (Haeb-Umbach and Ney, 1992; Aubert *et al.*, 1993). A detailed explanation of the LDA is given in Appendix A. As we have seen in section 4.3.2, the dimension of the STAP feature, incorporated with all the time-frequency pattern describing parameters, is 60. Performing LDA yields feature of dimension equal to the number of classes under consideration minus one. In our case, the context-independent phonemes constitute the classes, which is 27 in number. Hence the LDA transformed version of STAP feature, referred to as L-STAP, is 26 in dimension.

Evaluation of the STAP features

In the following sections of this chapter the STAP features are evaluated both for the clean and noisy speech conditions. An important note to consider here is the fact that the STAP approach can result in different feature representations based on the peak location estimation algorithm employed. In the previous chapter, we have developed two different algorithms for locating the spectral peaks, namely 1) frequency-based dynamic programming algorithm, and 2) HMM/ANN based algorithm. Accordingly, we evaluate the two forms of the STAP features, computed based on these peak estimation algorithms. Evaluation of the STAP feature also serve as an evaluation of the effectiveness of the peak location estimation algorithms.

In the following sections, L-STAP-DP denote the L-STAP features computed using peak locations estimated by the frequency-based dynamic programming algorithm (explained in the Section 3.4). L-STAP-$w_t w_f$-HA denote the STAP feature computed using peak locations estimated by the HMM/ANN based algorithm (explained in the Section 3.5). As we have seen in the Section 3.5, the main factors that affect the peak estimation in HMM/ANN algorithm is the time-frequency pattern size denoted by $w_t \times w_f$. Three different cases of $w_t \times w_f$ are considered: 1) $w_t = 1$ and $w_f = 1$, 2) $w_t = 3$ and $w_f = 1$, and 3) $w_t = 5$ and $w_f = 2$.

4.5 Clean speech recognition performance of the STAP features

The effectiveness of L-STAP features for the speech recognition task is tested on the OGI Numbers95 database, using a HMM/GMM system. Table 4.1 shows clean speech performance comparison between the L-STAP feature, the MFCC feature, and the CJ-RASTA-PLP feature. All the time-frequency pattern activity describing parameters, mentioned in section 4.2, are used for computing the L-STAP feature. As can be seen from the table, the performances of all the L-STAP features are significantly inferior to that of the MFCC feature. However, yet the recognition performance achieved by the L-STAP feature is interesting considering the fact that the information used by it to achieve this performance is less compared to that used by the MFCC. The performance achieved with this less information signifies that fact that features generated from time-frequency patterns around the spectral peak carry quite a significant amount of information for speech recognition. Interestingly, among the L-STAP features L-STAP-DP gives the best recognition performance, and among the L-STAP-$w_t w_f$-HA features, feature with lowest time-frequency pattern size for peak location estimation task, performs the best. Before drawing any conclusions upon this it is worth to wait for the results given in the later sections.

Temporal context

The temporal characteristics of the regular features are modeled through delta and acceleration of the static feature coefficients. However, delta and acceleration cannot be computed with the direct STAP features as they are generated by an abrupt masking of the non-peak regions. To take care of this, we have introduced temporal information in a different manner explained as follows: STAP

Feature	Word Recognition Rate, %
L-STAP-DP	81.1
L-STAP-11-DP	72.9
L-STAP-31-DP	71.3
L-STAP-52-DP	74.7
MFCC	94.4
CJ-RASTA-PLP	90.2

Table 4.1. Performance comparison of L-STAP, MFCC, and CJ-RASTA-PLP features in HMM/GMM system. L-STAP feature is computed using all the parameters mentioned in section ?? extracted from the time frequency patterns around the spectral peaks.

features with some temporal context, i.e., a set of few preceding and following vectors, are taken as the feature vector for the current time. These features are then decorrelated and reduced in dimension using LDA. For a case of temporal context size equal to 9, in which case, the feature vector dimensionality becomes $9 \times 60 = 540$, resultant LDA transformed vector of 26 dimension is obtained. The recognition performance for these feature vectors is given in Table 4.2. As can be seen from the table, including the temporal context improves the recognition performance. An additional observation is, similar to the results of previous subsection, frequency-based dynamic programming algorithm for peak location estimation performs better than the HMM/ANN algorithm. Again as mentioned in the previous subsection we delay our conclusion upon this until we collect all the results. The clean speech recognition performance is still significantly lower than that of the MFCC feature. However, as our main concern in the development of the STAP features is noise robustness, their noise robust speech recognition performance is described in the next section.

Temporal context size, in frames	Word Recognition Rate, in %			
	L-STAP-DP	L-STAP-11-HA	L-STAP-31-HM	L-STAP-52-HM
9	83.2	79.1	78.7	78.5
0	81.1	72.9	71.3	74.7

Table 4.2. Performance comparison of L-STAP features with differing temporal contextual information, in HMM/GMM system. L-STAP feature is computed using all the parameters mentioned in section 4.2 extracted from the time frequency patterns around the spectral peaks.

4.6 Noise robustness of STAP feature

Figures 4.2, 4.3, and 4.4 show performance comparison of noisy speech recognition using L-STAP-DP, MFCC, and CJ-RASTA-PLP features for various noise conditions at different noise levels (comparative study and discussion on L-STAP-$w_t w_f$-HA features are given separately in the later part of this section, as they are inferior to the L-STAP-DP in the noisy speech conditions as like the clean speech condition). Figure 4.2 gives the comparison when the speech is corrupted by additive factory noise. Figure 4.3 gives comparison for the lynx noise and figure 4.4 for the car noise. As can be seen from the figures, L-STAP-DP feature gives a significantly better recognition performance than the MFCC feature in high noise conditions. However, for low noise conditions it is inferior to the MFCC feature, and except for high factory noise levels, it is inferior to the CJ-RASTA-PLP feature. This can be attributed to the fact that the clean speech recognition performance of the L-STAP-DP feature itself is low to start with. Hence in the presence of noise, the performance degrades further and hence the noise robustness property of the STAP feature is able to show up well only in the high noise conditions. The noise robustness of the L-STAP-DP features can in fact be seen from the relatively slower degradation (relatively flatter curve) of its speech recognition performance curve with increasing noise level. To improve the noise robustness of STAP feature further, a better solution is to improve its clean speech recognition performance. Additionally, LDA, used to decorrelate the feature components of L-STAP-DP, is certainly not the best solution as it is still a linear technique and hence a projection of the feature space onto a linear sub-space may lead to loss of some speech discriminatory information. In chapter 6, we will see that a nonlinear equivalent of the LDA called TANDEM approach is able to improve the clean speech recognition performance of the STAP feature, and hence is able to utilize its noise robustness characteristics better.

An additional factor to note about the STAP feature is that its noise robustness is also heavily dependent upon the spectral peak location estimation, which is more likely to be prone to noise. Thus the parameterization scheme we use to compute the STAP feature can still be more robust if the spectral peak locations are more reliably estimated in the presence of noise. This is in fact shown to be true in the next chapter (Section 5.7), where a more reliable estimation of peak location is shown to result in further improvement in the noise robustness of the L-STAP-DP feature.

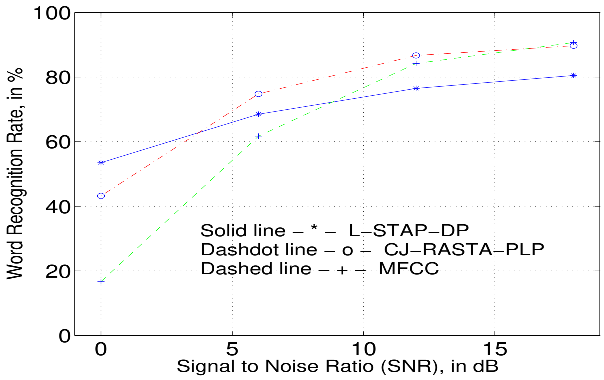

Figure 4.2. Performance comparison between the L-STAP-DP, CJ-RASTA-PLP, and MFCC features for various noise levels of the factory noise.

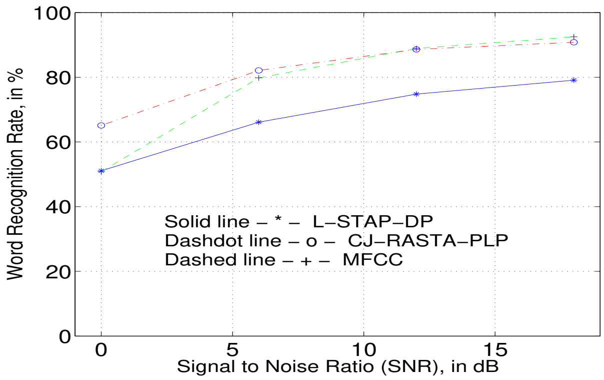

Figure 4.3. Performance comparison between the L-STAP-DP, CJ-RASTA-PLP, and MFCC features for various noise levels of the lynx noise.

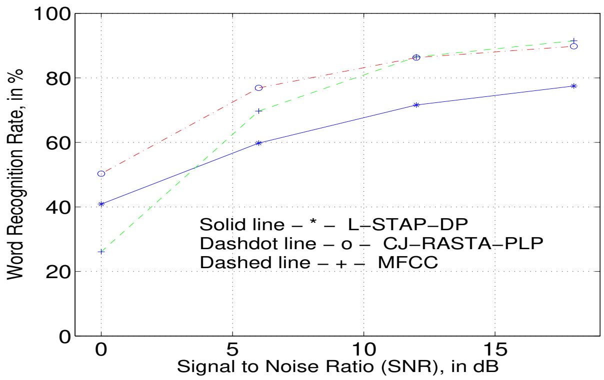

Figure 4.4. Performance comparison between the L-STAP-DP, CJ-RASTA-PLP, and MFCC features for various noise levels of the car noise.

Frequency-based DP algorithm vs HMM/ANN based algorithm

Figure 4.5 shows a comparison of recognition performances of speech corrupted by factory noise, for L-STAP-DP, L-STAP-11-HA, L-STAP-31-HA, and L-STAP-52-HA features (similar trends are observed when speech is corrupted by other types of noises, and hence such results are not given separately). As can be seen from the Figure L-STAP-DP feature is comparatively better than all the L-STAP-$w_t w_f$-HA features. As we have seen in the previous sections, similar is the case for the clean speech recognition performances also. This raises questions about the usefulness of the HMM/ANN based peak location estimation algorithm, as the basic difference between these features is the peak estimation algorithm. At this point one crucial distinguishing aspect to consider, in the case of HMM/ANN based peak location estimation algorithm, is the fact that MLP used actually learns the distribution of the distinct time-frequency patterns in the spectrogram during the training. Then during peak location estimation with a test spectrogram the posterior probabilities of time-frequency patterns are found to finally locate the peaks. In such a case, the use of mel-scaled filter-bank spectrum may not be the best suitable spectrum for the peak estimation task. An energy normalized spectrum where large variabilities in the energy levels are suppressed may be a better choice. This is indeed the case. As we will see in the next chapter (Section 5.7), an use of energy normalized spectrum with enhanced spectral peaks and smoothed spectral valleys for peak estimation task result in better recognition performance of the L-STAP-$w_t w_f$-HA features.

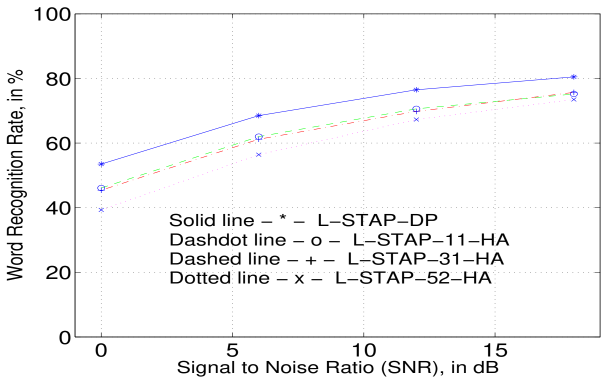

Figure 4.5. Performance comparison between the L-STAP-DP, L-STAP-11-HA, L-STAP-31-HA, and L-STAP-52-HA features for various noise levels of the factory noise.

4.7 STAP features in HMM/ANN system

As we have seen in the previous section, LDA is a linear technique and may not be good enough to handle the underlying complex feature correlations, which may lead to poor recognition performance of the L-STAP features. An interesting alternative to consider in such a case will be to evaluate the clean speech recognition performance of the STAP features in a HMM/ANN framework. This is because the MLP used in HMM/ANN system can handle the feature correlation and the temporal context well (STAP feature computed with peak location information obtained using frequency-based dynamic programming algorithm is only used for this study as it gives better performance than the STAP feature computed with HMM/ANN based peak estimation algorithm). The description of HMM/ANN system used to evaluate clean speech recognition performance of STAP feature is as follows: The MLP has 27 output units corresponding to the number of context-independent phonemes. Hidden layer size is proportional to the feature vector dimension and the input layer size is equal to the feature dimension multiplied by the context length. Table 4.3 shows performance comparison between the STAP, MFCC, and CJ-RASTA-PLP features in HMM/ANN system. STAP feature use all the time-frequency activity describing parameters as mentioned in section 4.2. Thus its feature dimension is 60. The temporal context size is 9. As can be seen from the table, the performance of the STAP feature improves in the HMM/ANN system than the L-STAP feature in HMM/GMM system. However, the clean speech recognition performance is yet inferior to that of the MFCC feature. Also the performance of MFCC in HMM/ANN is inferior than the HMM/GMM, because HMM/GMM can more effectively do triphone modeling. This, in fact, signifies the fact that LDA is not very appropriate to the STAP features. Hence, if their potential is used properly in HMM/GMM system there is a scope for improving them further in HMM/GMM system. In chapter 6, we will see that TANDEM approach provides a nice framework for doing this.

Feature	Word Recognition Rate, %
STAP	86.1
MFCC	91.9
CJ-RASTA-PLP	91.9

Table 4.3. Performance comparison of STAP, MFCC, and CJ-RASTA-PLP features in HMM/ANN system. STAP feature include all the parameters mentioned in section 4.2, extracted from the time frequency patterns around the spectral peaks.

Temporal context in STAP feature

Looking at the Figure 4.1, the STAP features can be expected to be more effective than the original spectrogram, in modeling the time-trajectories of the prominent time-frequency activities. A simple method by which this can be verified is to use STAP feature in HMM/ANN system with increased temporal context. Since the STAP features has a relatively less information than its capacity (most of its components are zeros) and thus an additional source of information incorporated by increasing the temporal context can be modeled better. Furthermore, an improved modeling with increase temporal context can be achieved in this case because the information that we consider as disturbing are also masked to zeros. Table 4.4 shows performance comparison between the STAP features and the MFCC feature when the temporal context is increase to 19. The results show an improvement in the recognition performance for the case of STAP features, whereas for the MFCC it is not the case. The absolute recognition performance obtained with STAP feature is interesting considering the fact that it still uses less information from the speech signal as compared to the MFCC.

Feature	Word Recognition Rate, %
STAP	87.2
MFCC	91.0

Table 4.4. Performance comparison of STAP and MFCC features in HMM/ANN system, when the input temporal context size is 19.

4.8 Evaluation of importance of STAP parameters

Another interesting case to look at is the evaluation of the contributions of the various pattern describing parameters, as listed in section 4.2, for the overall performance of the STAP feature. These parameters basically parameterize different aspects of the time-frequency patterns. Table 4.5 gives results of the experiments conducted to evaluate their relative importance. It basically gives a comparison of the speech recognition performances when the activity describing parameters incorporated in the STAP feature is varied. First column in the table gives description of the features used and the third column gives the word error rates. It is clear from the results that incorporation of more and more information about the activity of the time-frequency patterns around the spectral peaks improves the speech recognition performance. However, as given in the second column of the table, with more parameters the feature dimension of the STAP features increase more. However, as explained in section 4.3.1, this is because of the requirement of uniform dimensional feature vectors for the speech recognition systems. In actual, the amount of useful information in the STAP feature is very less. The results are given for both the cases when the temporal context size at the input of the MLP are 9 and 19. Similar trends are observed in both the cases.

Feature used	Dimension	% Word Recognition Rate in clean speech for MLP input context size	
		9	19
$\{L\}$	12	40.8	52.2
$\{L, E\}$	12	75.2	79.9
$\{L, E, \Delta_t E, \Delta_t^2 E\}$	36	83.4	85.6
$\{L, E, \Delta_t E, \Delta_t^2 E, \Delta_f E, \Delta_f^2 E\}$	60	86.1	87.2

Table 4.5. Comparison of the speech recognition performances of STAP features incorporated with various time-frequency pattern activity describing parameters.

4.9 Conclusion

Inspired by the two interesting aspects of the human auditory processing system, namely the noise masking and the local time-frequency processing, we have developed a new noise robust feature representation for speech recognition task called spectro-temporal activity pattern (STAP) features. In STAP approach, parameters extracted from local time-frequency patterns around the spectral peaks, describing the activity pattern within those patterns (i.e., the energy surface), are used as features. The effectiveness of the STAP features depend crucially upon the peak location estimation. Two peak location estimation algorithms, described in Chapter 3, namely frequency-based dynamic programming algorithm and HMM/ANN based algorithm, have been used to compute the STAP features. The STAP features have been evaluated both on the clean speech and noise corrupted speech. These evaluations actually serve as evaluation of both the STAP approach and peak location estimation algorithm.

From the result of the experiments, for peak-location estimation the frequency-based dynamic programming algorithm is better than the HMM/ANN based algorithm. As explained in Section 4.6, variations in the energy levels of the spectrum used for peak location estimation could be the reason for the inferior performance of the HMM/ANN system. In the next chapter (in Section 5.7), we will see that with an energy normalized spectrum, also with enhanced spectral peaks and smoothed valleys, HMM/ANN based peak location estimation algorithm is able to achieve performance close to that of the frequency-based dynamic programming algorithm.

Taking a overall look at the speech recognition performances of the STAP features in comparison with MFCC and CJ-RASTA-PLP features, it is possible to conclude the following: STAP features are relatively more robust in high noise condition. However, they are significantly inferior both in clean and low noise levels, which makes them highly unusable as stand alone features in the speech recognition system. The reason for this is the fact that non-peak regions which are masked to zeros in STAP features also carry useful information for the clean speech recognition. Thus an abrupt masking of these components may not be the right solution. In the next chapter, we will see an alternative soft masking approach, in order to improve the noise robustness. In addition to soft masking, this approach also avoids the explicit estimation of spectral peaks locations, thereby saving the feature vector from the problems faced with peak location estimation.

However, STAP feature do not end here. As we have seen in Section 4.6, the right solution to improve the overall performance of the STAP features is to improve its clean speech recognition performance. LDA used in this chapter, to make the STAP feature usable in the HMM/GMM system, is a linear technique, and may not be able to deal with the underlying complex feature correlations. As we will see in Chapter 6, a nonlinear equivalent of LDA called TANDEM approach is best suited for STAP features, and is able to show an overall improved recognition performance.

Chapter 5

Phase AutoCorrelation (PAC) features

In the previous chapter, STAP features computed based on spectral components only from regions around the spectral peaks has been shown to improve the noise robustness. This supports the fact that an useful noise robust information exists in regions around the spectral peaks. However, non-peak regions also carry significant amount of information, which is not noise robust, but useful for clean speech recognition. This is evident from the clean speech recognition performance of the STAP features which show a significant degradation in clean speech conditions when compared to standard features. This points to the fact that completely discarding information from regions around the non-peaks is not a right solution. A more appropriate solution would be to follow a soft-masking approach where non-peak regions are not discarded completely during the computation of the feature vectors, but are relatively deemphasized. Another factor to consider about the STAP features is that it is sensitive to the peak location estimation algorithm. This is evident from the significant differences in the performance observed (in the previous chapter) while using two different algorithms for peak location estimation.

In this chapter, we develop a new class of features, referred to as phase autocorrelation (PAC) features, that provide a nice solution to overcome the above mentioned problems. In contrast to the other noise robust methods that work at the spectral domain, the PAC approach addresses the problem of noise robustness at autocorrelation domain. The PAC uses phase (i.e., angle) variation of signal vectors over time as a measure of correlation (referred to as phase autocorrelation), as opposed to the regular autocorrelation where dot product of the time-delayed signal vectors is used as a measure of correlation. As will be explained with more details, such an use of PAC has an effect of enhancing the peaks and smoothing out the valleys, in the spectral domain. Interestingly, such an enhancement and smoothing are performed without explicit estimation of the peak locations, thus making the feature vectors independent of the peak estimation algorithm.

In the next section, we explain the regular autocorrelation, from which the traditional features are extracted, and its short comings in the presence of noise.

5.1 Autocorrelation

Feature extraction block in a typical speech recognition system divides the speech signal $s[n]$ into a sequence of frames given by,

$$\{s_0[n], s_1[n], ..., s_t[n], ..., s_{T-1}[n]\}$$

where T is the total number of frames. t^{th} frame is given by,

$$s_t[n] = \{s[Kt+0], s[Kt+1], ..., s[Kt+N-1]\}$$

where N is the frame length and K the frame shift. Traditional feature vectors are typically derived from the power spectrum of the signal frame. The power spectrum is obtained by first performing a discrete Fourier transform (DFT) of the signal frame samples and then taking the squared magnitude of the resulting coefficients for various frequencies (Rabiner and Schafer, 1978). DFT intrinsically assumes each frame $s_t[n]$ to be part of a periodic signal $\tilde{s}_t[n]$ (Oppenheim and Schafer, 1975) defined as:

$$\tilde{s}_t[n] = \sum_{k=-\infty}^{+\infty} s_t[n+kN] \tag{5.1}$$

The time-domain Fourier equivalent of power spectrum is the autocorrelation function. Autocorrelation, $R[n]$, over length N, can be computed from the periodic sequence in (**??**), using equation given as follows:

$$R[k] = \sum_{n=0}^{N-1} \tilde{s}_t[n]\tilde{s}_t[n+k], \quad k = 0, 1, ..., N-1. \tag{5.2}$$

The above operation of autocorrelation basically removes the phase differences between various sinusoidal components in the signal to yield $R[k]$. Looking at another view, the above equation gives a measure of correlation by computing dot product between two vectors in N dimensional space, as given by:

$$\begin{aligned} \mathbf{x}_0 &= \{\tilde{s}_t[0], \tilde{s}_t[1], ..., \tilde{s}_t[N-1]\} \\ \mathbf{x}_k &= \{\tilde{s}_t[k], ..., \tilde{s}_t[N-1], \tilde{s}_t[0], ..., \tilde{s}_t[k-1]\} \end{aligned} \tag{5.3}$$

$$R[k] = \mathbf{x}_0^T \mathbf{x}_k \tag{5.4}$$

If the samples spaced at an interval of k are highly correlated, $\mathbf{x}_0$ will be closer to $\mathbf{x}_k$ in the N dimensional space and hence will result in a higher value of the dot product. An alternative form of (5.4) is,

$$R[k] = \|\mathbf{x}\|^2 \cos(\theta_k) \tag{5.5}$$

where $\|\mathbf{x}\|^2 = \|\mathbf{x_0}\|^2 = \|\mathbf{x_k}\|^2$ represents the energy of the frame, which actually is a squared magnitude of the component vectors, and θ_k the angle between the vectors $\mathbf{x_0}$ and $\mathbf{x_k}$ in N dimensional space.

Noise sensitivity

In the presence of an external additive noise, denoted by $r[n]$, the resultant signal becomes $s^n[n] = s[n] + r[n]$. The autocorrelation, $R^n[k]$, for t^{th} frame of $s^n[n]$, denoted by $s_t^n[n]$, is the dot product between two vectors:

$$\mathbf{x}_0^n = \{\tilde{s}_t^n[0], \tilde{s}_t^n[1], ..., \tilde{s}_t^n[N-1]\}$$

$$\mathbf{x}_k^n = \{\tilde{s}_t^n[k], ..., \tilde{s}_t^n[N-1], \tilde{s}_t^n[0], ..., \tilde{s}_t^n[k-1]\}$$

where $\tilde{s}_t^n[n]$ is the periodic signal obtained from the frame $s_t^n[n]$. This $R^n[k]$ is clearly a function of the noise component present in the speech signal. An 2-D illustration of the effect of noise is given in Figure 5.1. In the presence of noise, the magnitudes of $\mathbf{x}_0^n$ and $\mathbf{x}_1^n$ and the angle between them undergo change causing variations in the dot product.

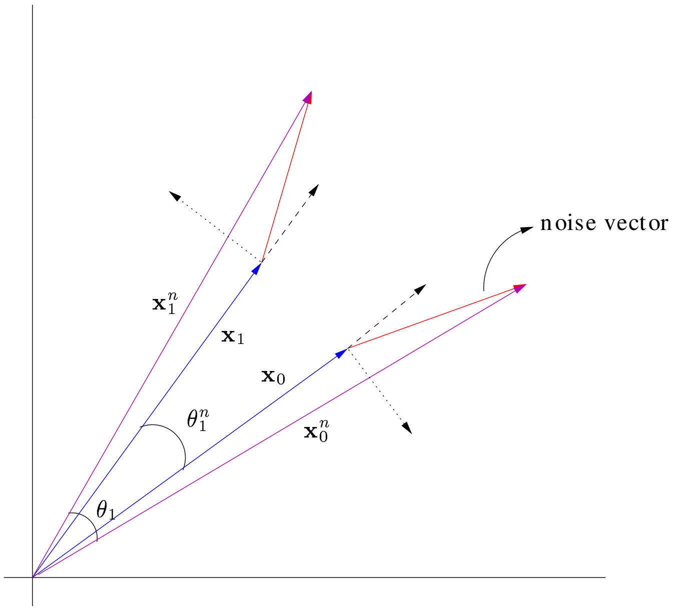

Figure 5.1. A 2-D illustration of how additive noise affects the autocorrelation function of the speech frames. The direction denoted by dashed and dotted arrows, gives the directions along and orthogonal to the speech vector, respectively.

5.2 Phase autocorrelation

In an attempt to reduce the sensitivity of the correlation coefficients to the external noise, we propose here a new measure of autocorrelation, where the angles θ_k, as appear in (5.5), are used as a measure of correlation. The resulting new set of correlation coefficients $P[k]$ are given as follows:

$$P[k] = \theta_k = \cos^{-1}\left(\frac{R[k]}{||\mathbf{x}||^2}\right) \tag{5.6}$$

This new measure of correlation is referred to as 'Phase AutoCorrelation' (PAC) (Ikbal *et al.*, 2003,a), as it gives a measure of phase, i.e., angle, variation of signal frame over time. From (5.6), the computation of PAC coefficients from the autocorrelation coefficients involve two operations:

1. Energy normalization, to compute autocorrelation normalized by instantaneous energy.

2. Inverse cosine, to nonlinearly transform the energy normalized autocorrelation coefficients into PAC coefficients.

These two operations thus convert the dot product of the speech vectors, as typically done during the computation of the autocorrelation coefficients, into angle between the vectors. This use of angle for correlation measure is motivated by the fact that the angle gets less affected by external additive noise than the dot product (Mansour and Juang, 1988).

As illustrated in Figure 5.1, both the angle and energy undergo change in the presence of noise. $R[k]$ depends on both the frame energy and the angle between the vectors, where as $P[k]$ depends only on the angle. Consequently, $P[k]$ are expected to be less susceptible to the external noise than the $R[k]$. Considering a special case, when the noise vector is in the same direction as the signal vector (along the direction of the dashed arrow in Figure 5.1), $P[k]$ do not get altered at all whereas $R[k]$ gets altered. However, when the noise vector is orthogonal to the signal vector (along the direction of the dotted arrow in in Figure 5.1), both $P[k]$ and $R[k]$ get altered to the same extent.

Energy normalized autocorrelation vs PAC

An interesting case to look now is the energy normalized autocorrelation coefficients as given below:

$$R_n[k] = \cos(\theta_k) = \frac{R[k]}{||\mathbf{x}||^2} \tag{5.7}$$

Ideally, even the use of energy normalized autocorrelation coefficients should result in noise robustness, since it also depends just on θ_k. This is indeed the case and experimental results given in the later sections of this chapter confirm this. However, the inverse cosine performed to compute θ_k also turns out to be an important operation, since an improved robustness is achieved while using θ_k as correlation coefficients. The inverse cosine operation leads to enhancement of the spectral peaks and smoothing out of the spectral valleys. As discussed in the starting of this chapter, this soft masking approach (where peaks are emphasized and valleys are deemphasized) is expected to be comparatively better than the abrupt masking, as done in STAP approach. However, PAC also has a drawback of inferior performance in clean speech. This is because, θ_k does not have the frame energy information which is crucial for the clean speech recognition. In addition, the smoothing of the valleys also leads to loss of information. Yet, the fact that PAC is a simple approach to achieve noise robustness makes it an interesting approach to consider further. The effects of the use of θ_k in the spectral domain is discussed in detail in the next section.

5.3 PAC spectrum

Frequency domain Fourier equivalent of the PAC coefficients is called the 'PAC power spectrum'. Equation (5.6) yields the values of $P[k]$ in the range 0 to π, for the input energy normalized autocorrelation values in the range $+1$ to -1. This causes the PAC power spectrum to have an unrealistically high value at zero frequency. To avoid this, $P[k]$ are transformed to $P_n[k]$ according to equation:

$$P_n[k] = 1 - \frac{2}{\pi} P[k] \tag{5.8}$$

Using (5.6), (5.7), and (5.8), we get

$$P_n[k] = 1 - \frac{2}{\pi} \cos^{-1}(R_n[k]) \tag{5.9}$$

Figure 5.2 shows a plot of this equation, which yields $P_n[k]$ values in the range $+1$ to -1.

Applying DFT analysis equation (Oppenheim and Schafer, 1975) on $R_n[k]$ and $P_n[k]$ will yield the energy normalized power spectrum, $S_n[l]$, and the PAC power spectrum, $S_p[l]$, as given below:

$$S_n[l] = \sum_{k=0}^{N-1} R_n[k] \exp(-j \frac{2\pi}{N} kl) \tag{5.10}$$

$$S_p[l] = \sum_{k=0}^{N-1} P_n[k] \exp(-j \frac{2\pi}{N} kl) \tag{5.11}$$

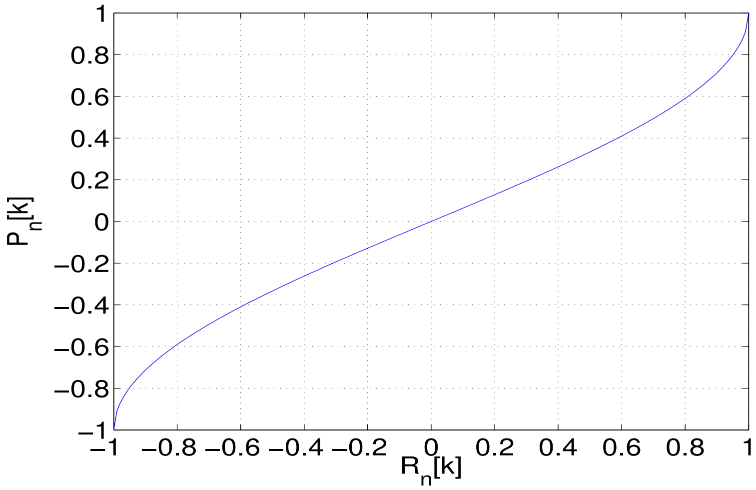

Figure 5.2. Normalized inverse cosine function.

5.3.1 PAC spectrum vs energy normalized spectrum

Because of the nonlinear relationship between $P_n[k]$ and $R_n[k]$ in (5.9), it is not possible to find a closed form relationship between $S_p[l]$ and $S_n[l]$. However, an empirical analysis of both the spectra leads to explanation of some interesting aspects of their relation, explained as follows: Figures 5.3 and 5.4 show plots of regular power spectrum (energy normalized) and PAC power spectrum respectively, for a sample frame of phoneme 'ih'. A visual observation of these spectra show that the PAC spectrum has its peaks enhanced than the peaks of the energy normalized power spectrum. The reason for this is the inverse cosine operation performed during the computation of the PAC coefficients. Figure 5.2 shows an increase in slope for the higher magnitudes of the x-axis. As a result, any variation in the values of x-axis near +/- 1 is magnified in the y-axis. Typically, the initial few coefficients of the autocorrelation are high in magnitude, and hence the variations within these coefficients are enhanced. These initial coefficients mainly decide the shape of the spectral envelope, as they constitute the slow varying part in the corresponding spectral domain. As a result, the shape of the spectral envelope, and hence the spectral peaks, are better enhanced in the PAC spectrum. On the other hand, when the autocorrelation coefficients are close to zero, which is typically the case in noisy vectors, the inverse cosine do not enhance the variation across them. An additional observation that can be made from the figures is the smoothing out of the fine details in the spectral valleys. This can be attributed to the fact that $P_n[k]$ and $R_n[k]$ have nonlinear relationship (according to (5.9)) and as a result, in the spectral domain, every frequency gets harmonics from other frequencies. The effect of this is more prominent in the valleys because of their low magnitude levels.

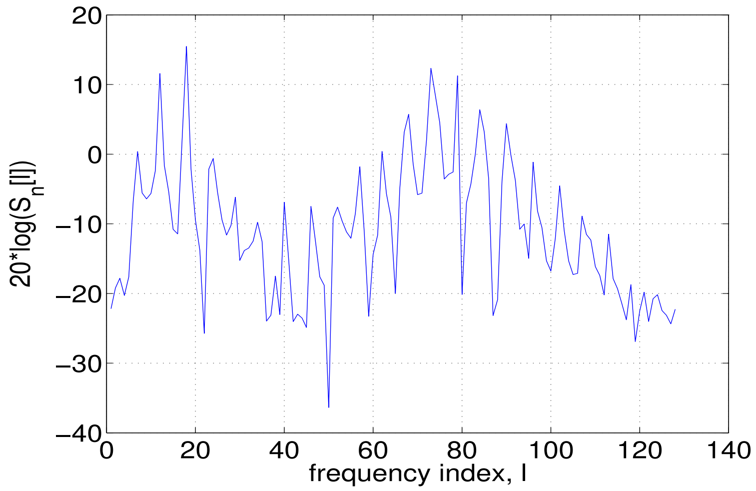

Figure 5.3. Logarithm of the energy normalized power spectrum, $S_n[l]$, for a frame of phoneme 'ih'.

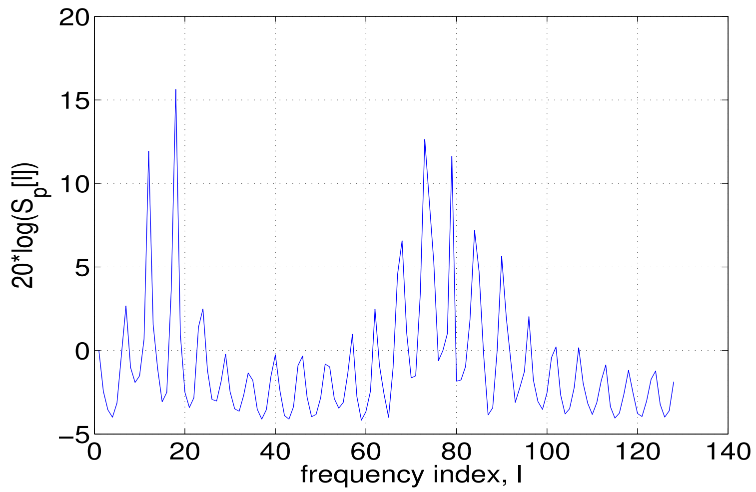

Figure 5.4. Logarithm of the PAC power spectrum, $S_p[l]$, for a frame of phoneme 'ih'.

The peak enhancement and the smoothing of the valleys in PAC power spectrum is further illustrated by Figure 5.5, showing the distribution of the PAC spectral power against the energy normalized spectral power for an example utterance. Each point in the figure corresponds to a particular frequency, with the corresponding x and y coordinates giving the spectral powers of the energy normalized spectra and PAC spectra respectively. It is clear from the figure that for higher power values, the relationship between energy normalized and PAC spectra is linear. Whereas for the lower power values, a larger range of the spectral axis is compressed within a small range of the PAC spectral axis.

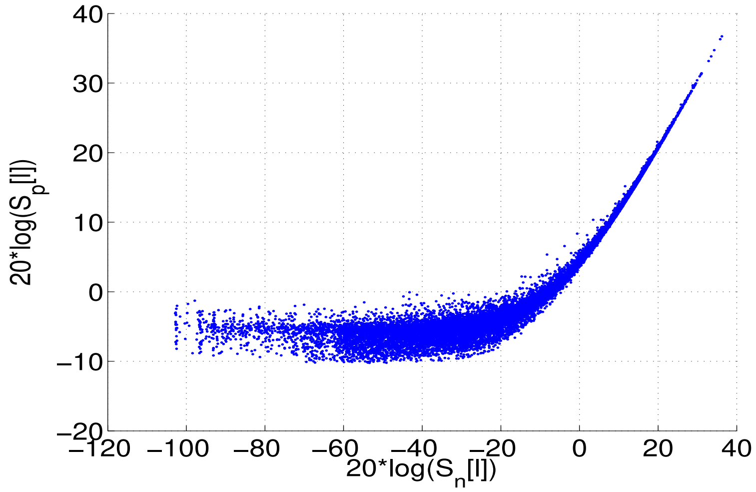

Figure 5.5. Distribution of the energy normalized spectral power against the PAC spectral power for an example speech utterance from OGI Numbers95 database.

5.3.2 Noise robustness of PAC spectrum

Spectral peaks constitute high signal to noise ratio (SNR) regions in the spectrum. Hence, the relative enhancement of the peaks in the PAC power spectrum leads to an improvement in the noise robustness. The noise robustness of the PAC spectrum is illustrated in Figures 5.6 and 5.7. Figure 5.6 shows a plot of Euclidean distances between spectra of clean speech and the spectra of same speech corrupted by an additive factory noise at 6dB SNR, over an example utterance. Figure 5.7 shows a similar plot for the PAC spectra. In order to have a fair comparison, the magnitudes of both the spectra are normalized to same range of values by mean removal and variance normalization. It is clear from the figures that the PAC spectrum is less affected by additive noise than the regular spectrum.

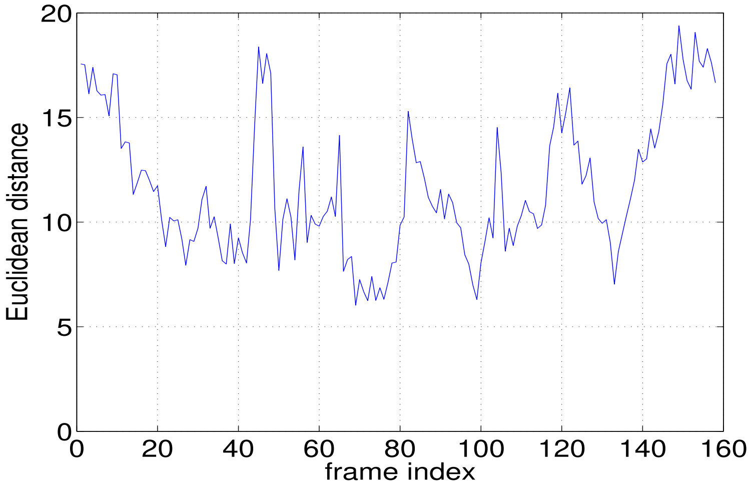

Figure 5.6. Euclidean distance between energy normalized spectra of clean speech and 6 dB additive factory noise corrupted speech for an example speech utterance from OGI Numbers95 database.

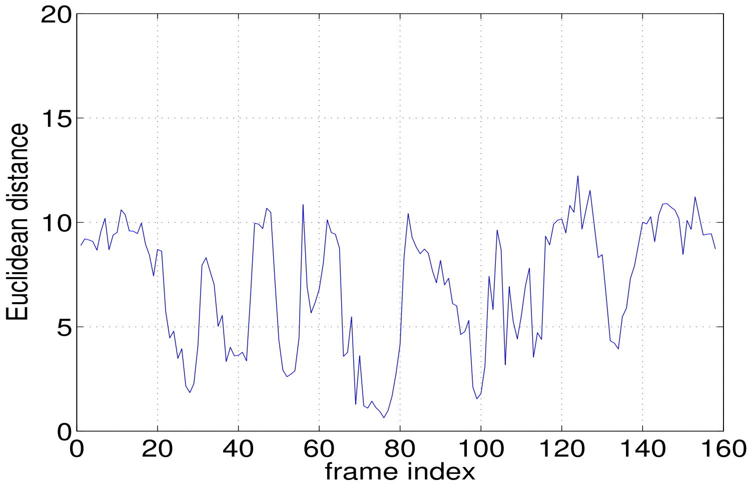

Figure 5.7. Euclidean distance between the PAC spectra of clean speech and 6dB additive factory noise corrupted speech for an example speech utterance from OGI Numbers95 database.

5.4 PAC features

An entire class of features, which are usually extracted from the regular spectrum, can now be extracted from the PAC spectrum. These features are referred to as 'PAC features' (Ikbal *et al.*, 2003). MFCC derived from the PAC spectrum is called the PAC-MFCC. The PAC features are expected to be more noise robust than their spectral equivalent. In addition to being robust, these features may also carry information that is complementary to that of the regular features as they went through a different processing procedure.

5.5 Performance of the PAC features

5.5.1 Noisy speech performance

Figures 5.8, 5.9, and 5.10 confirm the noise robustness of the PAC-MFCC. These figures compare the recognition performance of the PAC-MFCC with regular MFCC and CJ-RASTA-PLP features for various noise conditions at different noise levels. Figure 5.8 gives the comparison when the speech is corrupted by the factory noise. Figure 5.9 gives comparison for the lynx noise and figure 5.10 for the car noise.

From the figures, it is clear that in the presence of noise the performance of PAC-MFCC is significantly better than the regular MFCC. Additionally, the noise robustness of PAC-MFCC is comparable to that of the CJ-RASTA-PLP. Interestingly, in extreme noise conditions, PAC-MFCC is even more robust than the CJ-RASTA-PLP. And comparing these Figures with respective Figures 4.2, 4.3, and 4.4, in the previous chapter, it can be seen that PAC-MFCC is more robust to different kinds of noises than the L-STAP-DP feature (explained in previous chapter). However, we delay our conclusion about this until we compare their clean speech recognition performances also, in Section 5.5.2.

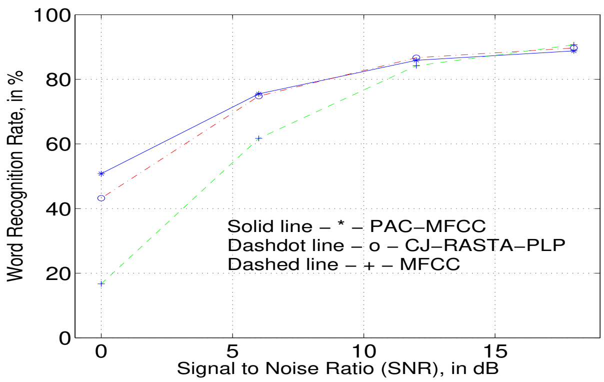

Figure 5.8. Performance comparison of PAC-MFCC with regular MFCC and CJ-RASTA-PLP for various noise levels of OGI Numbers95 database corrupted by an additive factory noise.

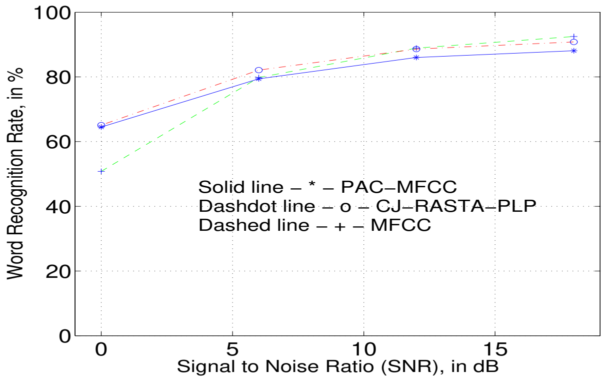

Figure 5.9. Performance comparison of PAC-MFCC with regular MFCC and CJ-RASTA-PLP for various noise levels of OGI Numbers95 database corrupted by an additive lynx noise.

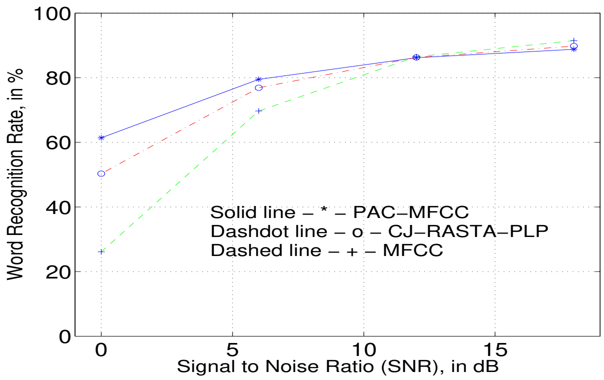

Figure 5.10. Performance comparison of PAC-MFCC with regular MFCC and CJ-RASTA-PLP for various noise levels of OGI Numbers95 database corrupted by an additive car noise.

Energy normalized MFCC vs PAC

As we have noted in section 5.2, the use of $R_n[k]$, as appear in (5.7), as autocorrelation coefficients should also result in an improved robustness. MFCC derived by using $R_n[k]$ as autocorrelation coefficients is called the energy normalized MFCC. Figure 5.11 gives a performance comparison between MFCC, energy normalized MFCC, and PAC-MFCC for various noise levels of factory noise corrupted speech. As can be seen from the figure, energy normalized MFCC is more noise robust than the regular MFCC. However, PAC-MFCCs are more noise robust than the energy normalized MFCCs. This is as a result of the inverse cosine operation performed to compute the PAC coefficients. As discussed in section 5.3.1, inverse cosine operation leads to enhancement of the spectral peaks and smoothing out of the spectral valleys, which improves the noise robustness further.

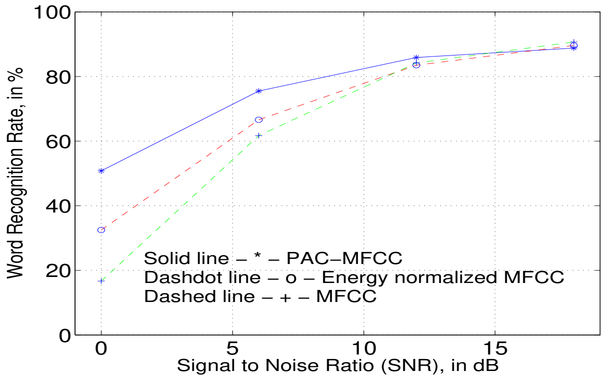

Figure 5.11. Performance comparison of energy normalized MFCC with regular MFCC and PAC-MFCC for various noise levels of OGI Numbers95 database corrupted by an additive factory noise.

5.5.2 Clean speech performance

Table 5.1 gives a performance comparison of PAC-MFCC with regular MFCC and CJ-RASTA-PLP for clean speech. In spite of their robustness to noise, PAC features are inferior to the regular features in clean speech. However, its performance is better than the clean speech recognition performance of L-STAP-DP feature as given in the Table 4.2. This shows that the soft-masking strategy followed in PAC approach is better than the abrupt masking of non-peak spectral regions as done in STAP approach. However, even soft-masking leads to clean speech degradation when compared to the standard features. Such a degradation in clean speech performance seems to be unavoidable in most of the noise robust techniques. For example, in the Table 5.1, the recognition performance of CJ-RASTA-PLP is also inferior to that of the MFCC in clean speech. However, the degradation of PAC-MFCC is more significant, and a more specific reason for this is the fact that (as explained in Section 5.2, the computation of PAC also involves energy normalization in addition to the inverse cosine. As we have seen in Section 5.5.1, such energy normalization helps in improving the noise robustness. However, it hurts the clean speech recognition performance as the energy constitutes an important source of information.

The inferior performance of the PAC features in clean speech makes them impossible to use as a stand alone feature in speech recognition systems. In the next two sections, we study more closely

Feature	Word Recognition Rate, % acc.
PAC MFCC	87.8
MFCC	94.4
CJ-RASTA-PLP	90.2

Table 5.1. Performance comparison of PAC-MFCC with regular MFCC and CJ-RASTA-PLP for clean speech of OGI Numbers95 database.

the effect of energy normalization and inverse cosine transformation on the PAC spectrum, and try to improve the clean speech recognition performance of the PAC features.

5.6 Improving the PAC feature in clean speech

5.6.1 Energy normalization

Energy normalization performed during the computation of the PAC coefficients is important from two aspects. First, the inverse cosine transformation requires the autocorrelation values to be in the range +/-1. Second, energy normalization also contributes to the robustness of the feature vector, as energy changes with the addition of noise. This is in fact, evident from the discussion in the Section 5.5.1 where energy normalized MFCC is shown to be more noise robust than the regular MFCC.

However, energy constitutes an important source of information for recognition of the clean speech. This is illustrated by the performance comparison in the first two rows of the Table 5.2, giving the clean speech recognition performances of the MFCC and energy normalized MFCC. This points to the fact that, the clean speech recognition performance of the PAC-MFCC can be improved by incorporating the energy information back in it, in a sophisticated manner. Row 3 of the Table 5.2 gives clean speech recognition performance of the PAC-MFCC, and row 4 gives the performance when the frame energy is appended to the PAC-MFCC as an additional coefficient (Ikbal *et al.*, 2003a). From the table, appending the energy with PAC-MFCC has resulted in a significant improvement of 3.8% for clean speech. However, the energy appended PAC-MFCC is yet inferior to the regular MFCC. This is because of the inverse cosine operation performed during the PAC computation, which while enhancing the peaks also smooths out the spectral valleys.

In the presence of noise, the incorporation of energy is expected to decrease the noise robustness of the PAC-MFCC. Figure 5.12 shows a performance comparison between the PAC-MFCC, energy

Feature	Word Recognition Rate, % acc.
MFCC	94.4
Energy normalized MFCC	91.7
PAC-MFCC	87.8
Energy appended PAC-MFCC	91.6

Table 5.2. Performance comparison of energy appended PAC-MFCC with PAC-MFCC and MFCC for clean speech of OGI Numbers95 database.

normalized MFCC, and energy appended PAC-MFCC for various noise levels of the factory noise. Interestingly, the performance of the energy appended PAC-MFCC is very close to the original PAC-MFCC, and significantly better than energy normalized MFCC. This is quite contrast to the regular MFCC where, as a result of the presence of the energy, the performance degrades more significantly in noise. The robustness of the energy appended PAC-MFCC can be attributed to the fact that here the energy is completely decoupled from the feature and is introduced as a single coefficient along with the feature. The HMM/GMM system seems to be not very sensitive to the noise related variability observed in a single coefficient than when the variability is observed in the entire feature vector. A behavior similar to this can be found in (Stephenson *et al.*, 2003) where performance improvement is achieved while energy is used as an auxiliary variable.

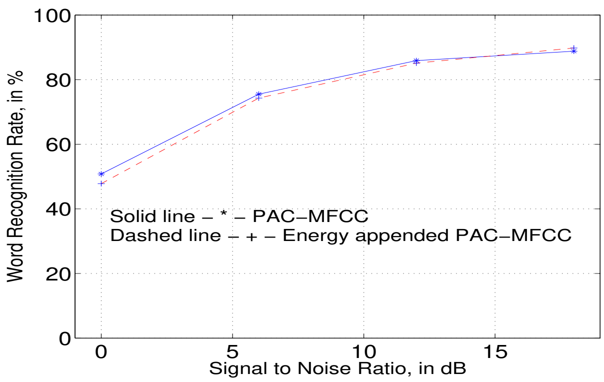

Figure 5.12. Performance comparison of energy appended PAC-MFCC with PAC-MFCC for various noise levels of OGI Numbers95 database corrupted by additive factory noise.

However, from the clean speech recognition performance shown in table 5.2, the incorporation of energy is not sufficient for improving the PAC features in clean speech. This is because of the inverse cosine operation. In the next section, we analyze the effects of inverse cosine operation.

5.6.2 Inverse cosine

As explained in Section 5.3.1, inverse cosine function enhances the PAC spectral peaks and smooths out the spectral valleys. This results in improved noise robustness as the spectral peaks are less sensitive to noise. However, in the clean speech, this results in degradation of the recognition performance, as can be seen from the Table 5.2. The regular MFCC features and the energy appended PAC MFCC features carry the same information except for the fact that inverse cosine operation is performed additionally in the later case. This causes 2.8% drop in recognition rate for clean speech. This raises questions about optimality of the inverse cosine function for PAC computation. In this section we explore alternative nonlinear functions to the inverse cosine function, that have similar characteristics as the inverse cosine for improving the noise robustness, but also would not hurt clean speech recognition performance.

Figure 5.13 shows a few examples of alternate functions we consider. In the figure, functions plotted by solid lines are linear and inverse cosine. Those plotted by dotted and dashed lines are alternate functions that yet have the shape of the inverse cosine but differ in the magnitudes. The family of dashed curves are specified by the values of variable f from -1 to $+1$. When $f = -1$ the function is linear and when $f = +1$ function is inverse cosine. All the functions in between are specified by f values between -1 to $+1$.

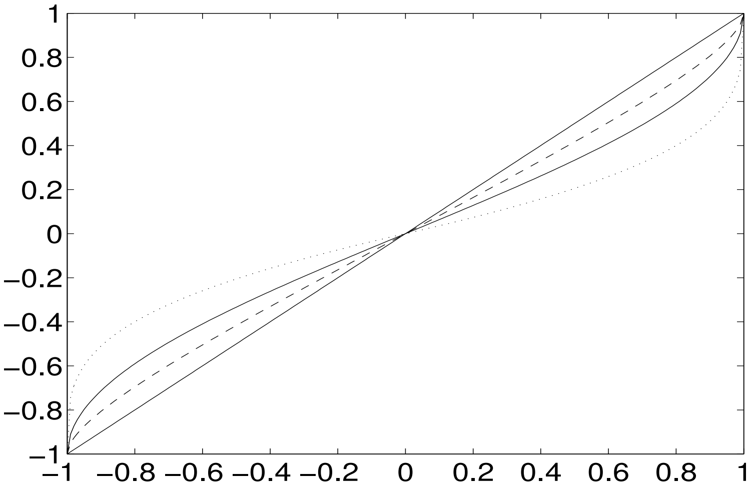

Figure 5.13. Alternative nonlinear functions to the inverse cosine.

The function plotted with dotted line looks interesting for our current investigation because its slope is larger than inverse cosine for larger values of x. Hence, according to the argument in Section 5.3.1, this function should enhance the spectral peaks even better. Unfortunately, this function do not yield better performance both for clean and noisy speech. The recognition performances obtained are 87.0% for clean speech and 71.8% for 6dB factory noise corrupted speech. The reason for this could be the fact that such transformation could severely modify the autocorrelation, magnifying even a simple variation in the autocorrelations, thus making them unsuitable for further processing. This turns our attention to a set of functions shown by the dashed line, because they cause milder modifications than the inverse cosine. Figure 5.14 shows plots of recognition performance for the clean speech and the 6dB factory noise corrupted speech, for various values of f. For clean speech, with highest recognition performance for $f = -1.0$, which corresponds to energy normalized MFCC, the performance drops down gradually with increasing f and reaches a low value when $f = 1.0$, which corresponds to the PAC-MFCC. This leads to a conclusion that all the nonlinear transformations hurt the recognition performance of clean speech. The milder the nonlinearity, lesser the degradation. But the nonlinear transformation certainly helps in the noisy speech. As we can see from the Figure, even for the lower values of f, the recognition performance is reasonably better than the linear transformation.

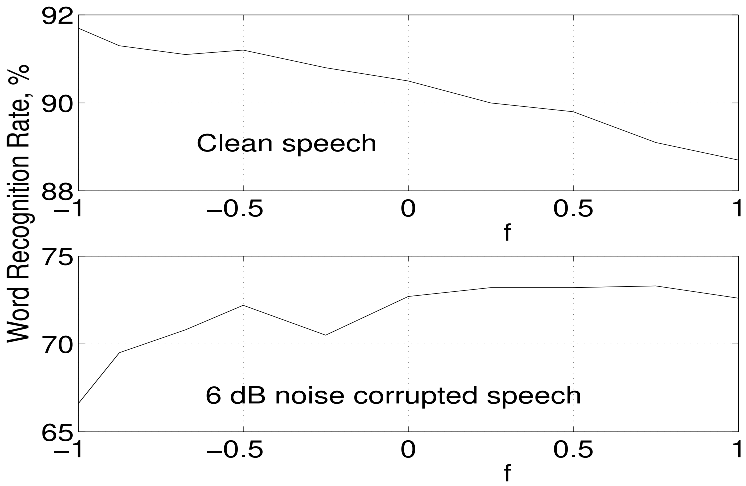

Figure 5.14. Recognition performances of the alternative nonlinear functions.

Preliminary conclusion

Although not all possible alternatives to inverse cosine are explored, the above analysis point to a fact that it is difficult to achieve an improvement in noise robustness without hurting the clean speech recognition performance.

Before concluding this chapter, the next chapter considers an interesting case, where the PAC spectrum is used for the peak location estimation task in computing the STAP (explained in the previous chapter) feature. As the PAC spectrum is energy normalized and its peaks are enhanced and valleys are smoothed than the regular spectrum, using it for peak location estimation is expected to result in more reliable peak location information.

5.7 PAC spectrum for peak identification in STAP

We have seen (in Section 5.3.1) that the peaks of PAC spectrum are more enhanced than its spectral counterpart, and their valleys of are smoothed out. Additionally, they are energy normalized also. With these properties an interesting trial one would like to make is to use the PAC spectrum in spectral peak location estimation for the STAP feature computation (explained in Chapter 4). The resulting features are referred to as PSTAP features. Both the peak location estimation algorithm, developed in Chapter 3, namely, 1) frequency-based dynamic programming algorithm, and 2) HMM/ANN based algorithm, are tested with PAC spectrum, and the corresponding L-PSTAP features computed (L-PSTAP-DP and L-PSTAP-31-HA[1], explained in Chapter 4) are evaluated.

5.7.1 Frequency-based dynamic programming algorithm

Table 5.3 shows comparison between the clean speech recognition performance when PAC spectrum and regular spectrum are used to estimate the spectral peaks in computing the L-STAP-DP features (explained in Section 4.8). Figures 5.15, 5.16, and 5.17 shows comparison of their performances for various noise levels of speech corrupted by factory noise, lynx noise, and car noise respectively. As can be seen from the Table and the Figures, the overall performance is better while using PAC spectrum (L-PSTAP-DP feature) for peak location identification than while using regular spectrum

[1]Here only time-frequency patters of size $w_t = 3$ and $w_f = 1$ are considered, as they gave the best results.

(L-STAP-DP). This illustrate the reliability of the PAC spectrum when compared to the regular spectrum for the peak location identification task.

Spectrum used for peak identification	Word Recognition Rate, %
PAC spectrum	84.6
regular spectrum	83.2

Table 5.3. Performance comparison when the PAC spectrum (L-PSTAP-DP) and regular spectrum (L-STAP-DP) are used as input to the frequency-based dynamic programming algorithm for peak location estimation.

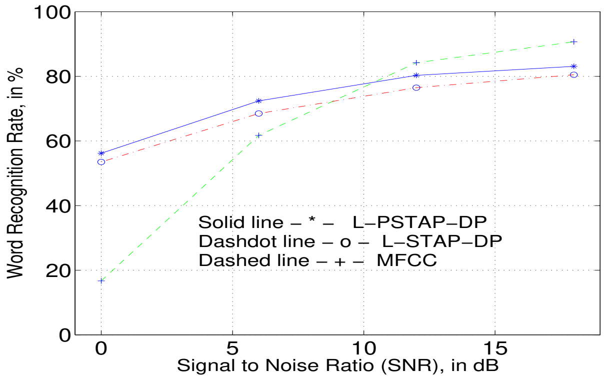

Figure 5.15. Performance comparison when the STAP features computed based on peak location estimation with PAC spectrum (L-PSTAP-DP) and regular spectrum(L-STAP-DP, and MFCC features for various noise levels of the factory noise.

5.7.2 HMM/ANN based algorithm

Table 5.4 shows comparison between the clean speech recognition performance when the PAC spectrum and the regular spectrum are used to estimate the spectral peaks in computing the L-STAP-31-HA features (explained in Section 4.8). Figures 5.18, 5.19, and 5.20 shows comparison of their performances for various noise levels of speech corrupted by factory noise, lynx noise, and car noise respectively. Again as can be seen from the Table and the Figures, the overall performance is better while using PAC spectrum (L-PSTAP-31-HA) for peak location identification than while using regular spectrum (L-STAP-31-HA). This again illustrates the reliability of the PAC spectrum compared to the regular spectrum for the peak location identification task. However, similar to

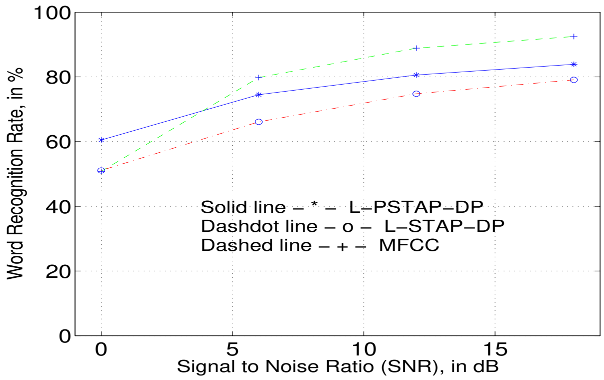

Figure 5.16. Performance comparison when the peak location estimation is done with PAC spectrum (L-PSTAP-DP) and regular spectrum (L-STAP-DP), and MFCC features for various noise levels of the factory noise.

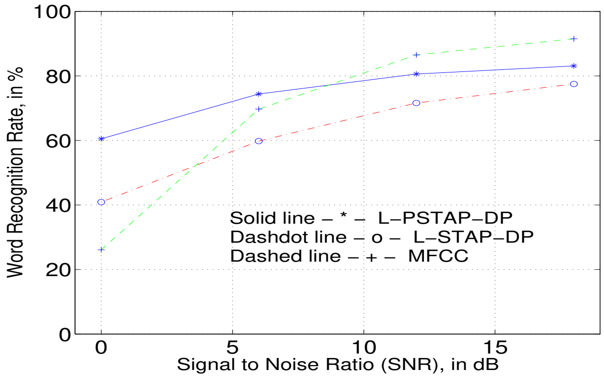

Figure 5.17. Performance comparison when the peak location estimation is done with PAC spectrum (L-PSTAP-DP) and regular spectrum (L-STAP-DP), and MFCC features for various noise levels of the factory noise.

the results while using regular spectrum, even while using the PAC spectrum, the performance of the HMM/ANN based algorithm is inferior to that of the frequency-based dynamic programming algorithm.

Spectrum used for peak identification	Word Recognition Rate, %
PAC spectrum	84.1
regular spectrum	78.7

Table 5.4. Performance comparison when the PAC spectrum (L-PSTAP-31-HM) and regular spectrum (L-STAP-31-HA) are used as input to the HMM/ANN based algorithm for peak location estimation.

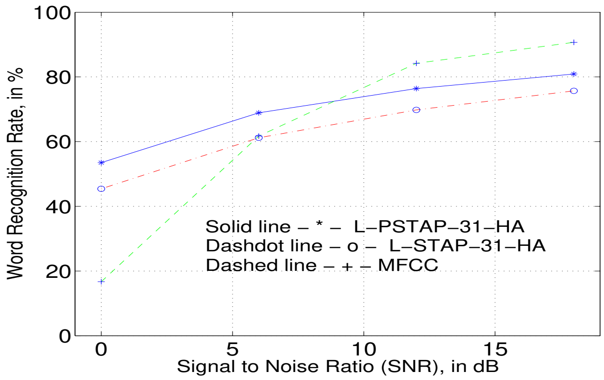

Figure 5.18. Performance comparison when the peak location estimation is done with PAC spectrum (L-PSTAP-31-HA) and regular spectrum (L-STAP-31-HA), and MFCC features for various noise levels of the factory noise.

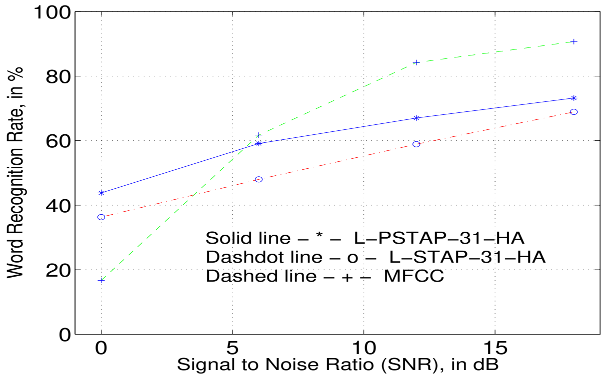

Figure 5.19. Performance comparison when the peak location estimation is done with PAC spectrum (L-PSTAP-31-HA) and regular spectrum (L-STAP-31-HA), and MFCC features for various noise levels of the lynx noise.

5.8 Conclusion

In this chapter we have introduced a new class of noise robust features called phase autocorrelation (PAC) features. These features are derived from an alternative measure to the autocorrelation called phase autocorrelation. PAC used angle variation of the signal frame over time as a measure of correlation, as opposed to the regular autocorrelation which computes correlation as a dot product between the time-delayed signal vectors. The use of angle for the correlation measure makes the PAC more robust to noise than its regular autocorrelation counterpart. This is because in the presence of an additive disturbance angle gets less disturbed than the dot product. The use of angle as a measure of correlation has an interesting effect of enhancing the peaks and smoothing of the

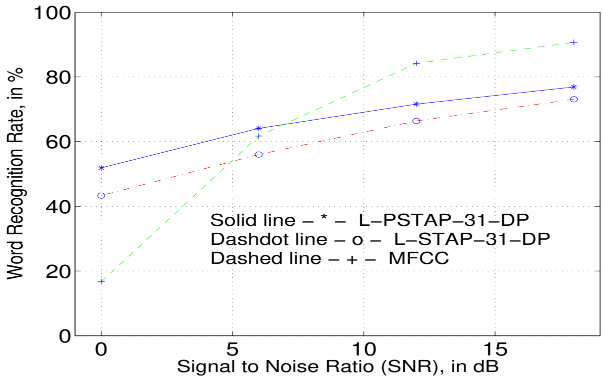

Figure 5.20. Performance comparison when the peak location estimation is done with PAC spectrum (L-PSTAP-31-HA) and regular spectrum (L-STAP-31-HA), and MFCC features for various noise levels of the car noise.

valleys in the spectral domain. This serves as a soft-masking technique to deemphasis the spectral valley regions, which is our main motivation in developing the PAC. Experimental results show that such soft-masking is able to perform better in both clean and noisy speech conditions, when compared to the STAP approach (developed in the previous Chapter) where a hard-masking strategy is followed by completely discarding the information from the spectral valley regions. Apart from this, the use of PAC spectrum in the peak location estimation task for computing the L-STAP features has resulted in improved clean and noisy speech recognition performance.

However, as discussed in Section 5.6.2, even soft-masking hurts the clean speech recognition performance and result in inferior performance of the PAC features when compared to the standard features. Similar characteristics are observed in most of the noise robust techniques developed in the past. Thinking more about this we arrive at the following conclusion. The reason for the degradation could be the fact that performing some external modifications to the spectrum or features using the limited knowledge we have gained about the speech or human speech perception may not be enough. The underlying complexities of the speech always lead to hurting one factor while trying to improve the other. For example, we know from perceptual studies that relatively weaker spectral components are masked in human auditory system while recognizing sound. But we don't understand how exactly it is performed. Thus an external design of noise robust algorithm may not be able to result in improved noise robustness without hurting clean speech recognition performance. In the next chapter, we analyze an interesting alternative that is based on data-driven feature extraction.

Chapter 6

Noise Robustness Analysis of TANDEM Approach

6.1 Introduction

Most of the feature based noise robust algorithms utilize external knowledge about the effect of noise on the speech in order to device an appropriate algorithm that would result in improved noise robustness. Such external knowledge is typically inferred from the human speech perception experiments. However, the underlying complexity of the speech process limits us from gaining the complete knowledge required to design externally an effective noise robust algorithm meeting the actual requirements. Sometimes, algorithms designed based on limited knowledge may also hurt class discriminatory information, leading to degradation of clean speech recognition performance. For example, the STAP and PAC features (explained in the Chapters 4 and 5), use the knowledge (gained from the human perception systems) that in the presence of noise, relatively weaker spectral components are masked. However, various factors taken into account and the actual procedure of how this is done in human perception system is not known. Thus the externally designed transformations (based on the partial knowledge) to perform the masking has resulted in significant degradation of the clean speech recognition performance. This is infact the case with most of the noise robust approaches.

An ideal solution in such scenario would be to have an algorithm that can learn an appropriate transformation required for improving the noise robustness from the training data, satisfying constraints such as keeping the clean speech recognition performance intact. Such methods are referred to as data-driven approaches for noise robustness. In this chapter, we analyze a recently proposed data-driven approach called TANDEM approach (Hermansky *et al.*, 2000; Ellis *et al.*, 2001) for the case of noise robustness. TANDEM approach is basically a nonlinear equivalent of the linear discriminant analysis (LDA). LDA (Duda and Hart, 1973; Haeb-Umbach and Ney, 1992) (explained in Appendix A), projects the input feature space onto a linear sub-space whose axes are along the maximum possible sound discriminatory information. Similarly, as we will see in later sections of this chapter, TANDEM projects feature space onto a nonlinear subspace along maximum possible sound discriminatory information. The analysis of TANDEM approach for noise robustness has led to understanding of quite a few interesting aspects of it. Interestingly, TANDEM approach can also be used as an integration tool for combining several feature streams, which will be discussed in detail in the next chapter. In the next Section, we give a brief description of the TANDEM approach.

6.2 TANDEM approach

TANDEM approach combines two major approaches for speech recognition namely: 1) the HMM/GMM approach, and 2) HMM/ANN approach (Hermansky *et al.*, 2000; Ellis *et al.*, 2001). In this way it combines the discriminant training abilities and temporal context modeling abilities of the HMM/ANN approach with the HMM/GMM approach to generally yield an improvement in the recognition performance. Figure 6.1 gives an illustration of the TANDEM approach. As can be seen from the figure, it has two emission probability models, one the MLP and other the GMM, in TANDEM. However, the MLP in this case is not used for the emission modeling. Instead it is used for the feature transformation. It actually acts as a means to perform a data-driven feature transformation of the input feature. The output of the MLP, which is supposed to be a better feature representation, is further transformed with logarithmic transformation and Karhunen-Loeve (KL) transformation, and given as an input feature to the GMM of HMM/GMM system. We refer to this transformed feature representation as the *TANDEM representation* of the input feature.

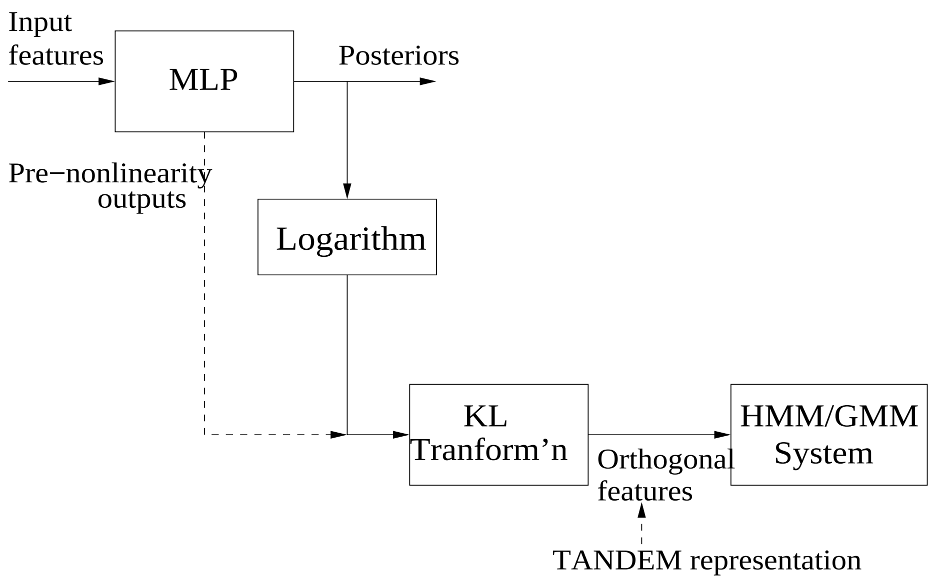

Figure 6.1. Illustration of the TANDEM system. Transformed posterior outputs of the MLP constitute the TANDEM representation of the input feature.

The feature transformation performed by the MLP is basically learned from the training data, on which it is trained in a supervised, discriminative classifier mode, with output classes being the context-independent phonemes. As known from the theory of HMM/ANN approach for speech recognition, when a MLP is trained in discriminative classifier mode, it learns the estimates of the posterior probability distribution of the input feature vector space (Bourlard and Morgan, 1993). Hence, the output of the MLP are basically the estimates of the posterior probabilities of the phoneme classes. The posterior probabilities, if estimated accurately, actually correspond to the best source of phonetic class information. Hence they are used as input features to the HMM/GMM system. However, it is not possible to use them directly as feature input to the GMM, because the MLP output distribution is highly skewed and also the output components are highly correlated. Thus, before feeding into GMMs, as illustrated in the Figure 6.1, the posterior probabilities are passed through the following transformations (Hermansky *et al.*, 2000):

1. Logarithmic transformation of the softmax outputs to modify the skewness of the posterior probability distribution. This can also be achieved by directly taking the pre-nonlinearity outputs of the MLP[1], as illustrated in the Figure 6.1.

[1]Throughout this work the prenonlinearity outputs are used for computing the TANDEM representations, as they per-

2. Karhunen-Loeve (KL) transform to decorrelate the features.

The output of these transformations, which constitute the TANDEM representation, is fed as the input to the GMM. The TANDEM system has been shown to perform significantly better than both of the component systems, HMM/GMM and HMM/ANN, in clean speech (Hermansky *et al.*, 2000; Ellis *et al.*, 2001; Ikbal *et al.*, 2004a). In the context of multicondition training, TANDEM approach has also been shown to be useful for noise robustness (Hermansky *et al.*, 2000). However, it has not yet been used and analyzed for the case where training is performed entirely on the clean speech. In the next section, we analyze this and in the sections later on we evaluate the noise robustness of the TANDEM approach.

6.3 Noise robustness of TANDEM representations

While training the MLP in a supervised, discriminative classifier mode, it actually performs a non-linear discriminant analysis (NLDA) of the input feature space. Each component of the MLP output can be seen as a projection of the input feature space onto a nonlinear axis corresponding to its output node. Because of the discriminative training, the MLP learns such nonlinear axis along the direction of maximum class discriminatory information, i.e., along the nonlinear axis of a particular output node, the corresponding phoneme class can be discriminated from the other phoneme classes to the maximum possible extent. As a result, the nonlinear sub-space constituted by the nonlinear axes of all the output nodes of the MLP, which is actually the MLP output space, is along the maximum possible class discriminatory information. Thus, after the training, when the MLP is used for feature transformation, it basically projects the input feature vector onto the nonlinear subspace along the maximum possible class discriminatory information.

As we know, any transformation that involve projection onto a subspace will lead to retention of information only along that space and lose of all other information. Hence, the projection of the input feature space onto the space at the output of the MLP will retain mainly the class discriminatory information. All other information will be either reduced or removed completely, depending upon whether they are partially along the output space or not. The complete removal happens when they are in the orthogonal direction (in a nonlinear sense). This is explained well by a linear

form better in terms of recognition than while using the logarithmic posteriors. The reason for the inferior performance of logarithmic posteriors could be the fact that the softmax is a many-to-one function.

equivalent of such operation, as given in Figure 6.2, which shows a 2-D example that is commonly used to explain the LDA. As can be seen from the figure, there are two classes. The direction shown by the solid line arrow corresponds to the direction of maximum possible class discriminatory information. Projecting the 2-D data onto the axis along this arrow would lead to retention of class discriminant information, but lose of other information. For example, projection will reduce, in the new space, the information along the direction denoted by the dash-dot arrow and will completely remove the information along the direction of the dashed line arrow.

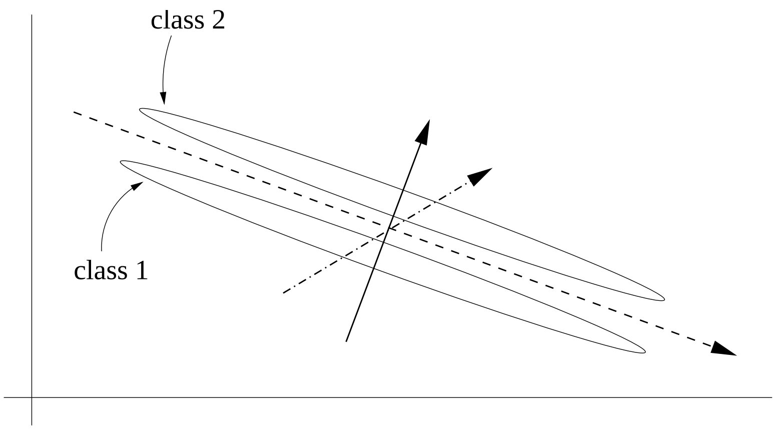

Figure 6.2. 2-D illustration of noise reduction while projecting along maximum possible class discriminatory information.

Thus, the transformation performed by the MLP is expected to reduce, at its outputs space, any information other than the class discriminatory information. As shown in simple 2-D illustration given in Figure 6.3, the noise related variability and any other disturbing variability will also get reduced if they are not along the sub-space of maximum class discriminatory information. A reduction in noise related variability will lead to improvement in noise robustness of the TANDEM representation. Thus the TANDEM representations can be claimed to be noise robust if it can be shown that the noise information is actually not along the space of maximum class discriminatory information. In the next section, we try to show this through a linear equivalent case, by performing LDA on the input feature space of the clean and noisy speech[2].

[2]The underlying complexity of the speech and noise processes makes it almost impossible to show this directly for the nonlinear case.

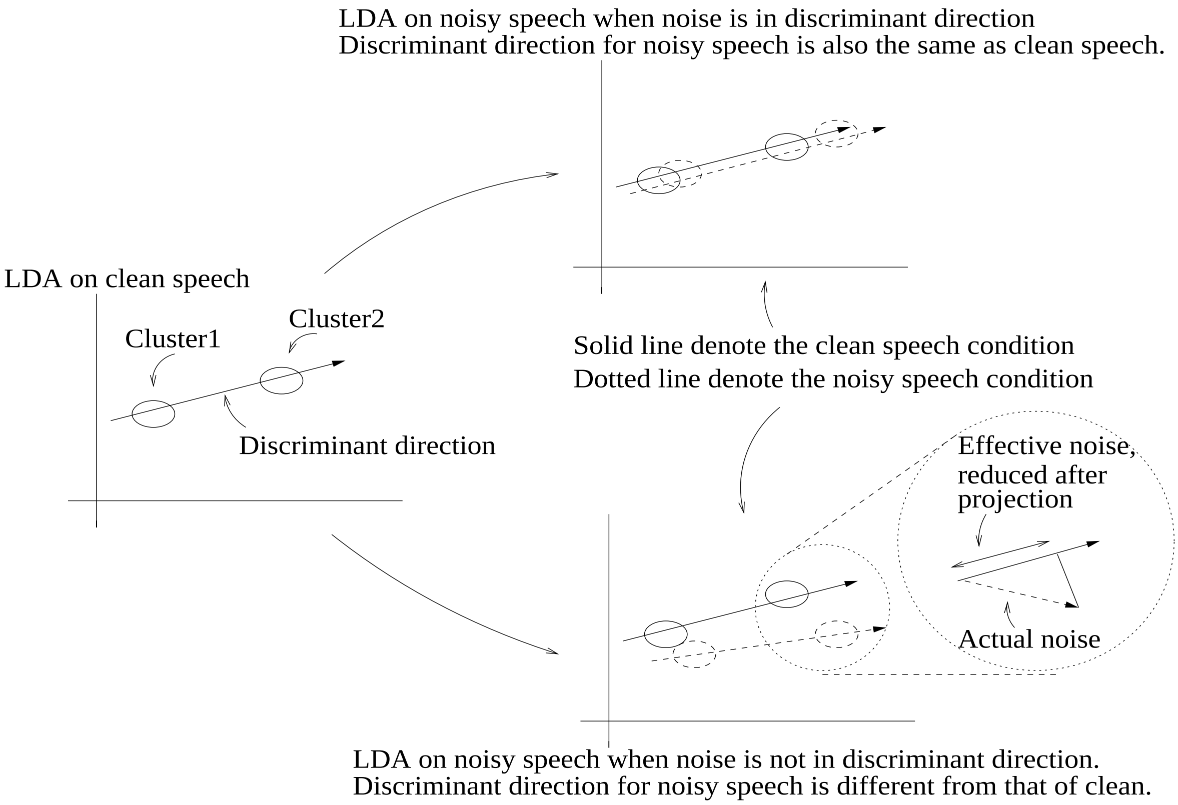

Figure 6.3. A 2-D illustration of how the class discriminant direction is affected when the noise is not along it. A projection onto clean speech discriminant direction will reduce the noise variability.

Simple noise analysis through LDA

Figure 6.4 shows results of LDA performed on a 2-D feature space, constituted by the 2nd and 3rd components of the MFCC feature, for clean as well as noise corrupted speech of various noise levels. The database used for this experiment is OGI Number95 database. For noisy speech experiments, factory noise from the Noisex92 database is added with the clean speech. In order to have an equivalent linear case of the nonlinear discriminant analysis performed by the MLP, a simpler two class problem is considered, where all the feature vectors belonging to phoneme 'ih' are treated as 'class 1' and the feature vectors of all other phonemes are treated as 'class 2'. Performing a LDA on the 2-D feature space, with this class information, will yield a principal direction along which the discrimination between the classes is the maximum. The value of Fisher discriminant ratio for such direction will be the highest. Figure 6.4 shows such directions identified by the LDA for clean speech as well as noise corrupted speech for various noise levels. As can be seen from the figure, the direction of the maximum discriminatory information changes in the presence of noise. Interestingly, the angle of deflection of such direction for noisy speech from that of the clean speech also increases gradually with the increase of the noise level. This can happen only when the noise disturbance is not along the direction of the class discriminatory information. This supports our claim atleast for the current 2-D, two class problem.

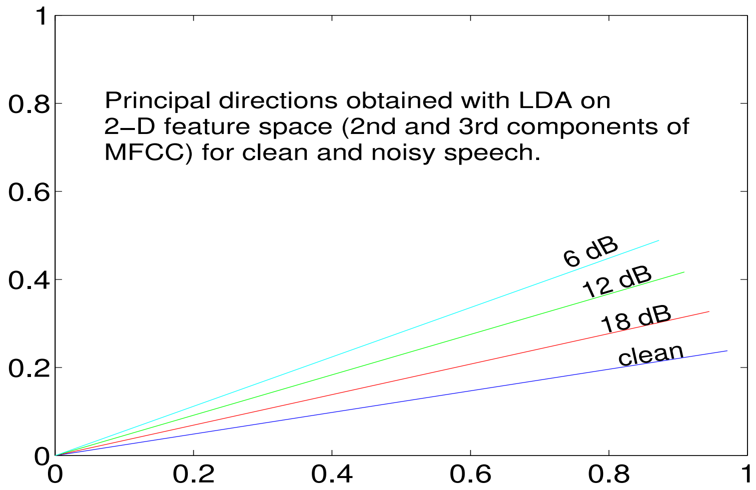

Figure 6.4. 2-D illustration to show that the noise information is not along the direction of the class discriminatory information. The deflection of the principal discriminant direction for various levels of noise supports this.

Figure 6.5 shows result of a similar experiment performed on the full (39-D) MFCC feature space. The figure shows angle of deflection of the direction of highest discrimination, in 39-D space, for various noise levels of noisy speech when compared to the clean speech. The two curves in the Figure corresponds to the angle deflections for two different types of noises, one factory noise and the other lynx noise. As can be seen from the figure, in the full MFCC feature vector space also the direction of maximum class discriminatory information gets deflected in the presence of noise, and the angle of deflection increases gradually with the noise level. Interestingly, it can also be seen that for a particular noise level, the angle of deflection for the factory noise, which is a full-band nonstationary noise, is more when compared to the lynx noise, which is a colored noise. This result also supports the fact that noise information is not along direction of the maximum possible class discriminatory information.

The trends seen in the above experiments for the linear case can be assumed to generalize for the

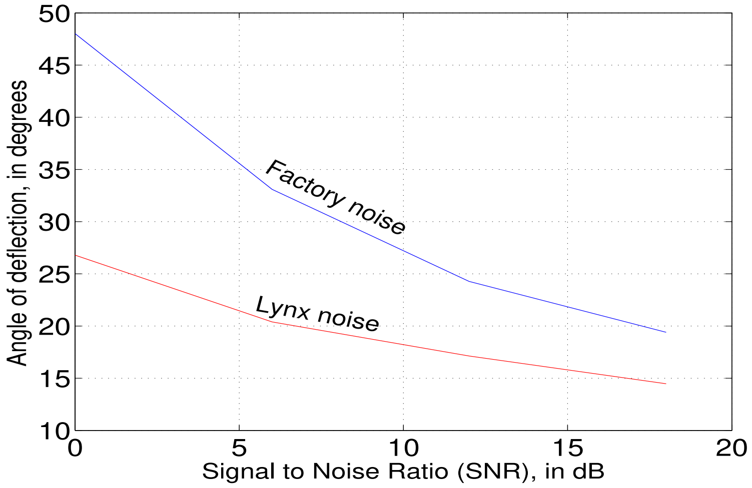

Figure 6.5. Angle of deflection of the principal discriminant direction (computed in 39-D feature vector space) in the presence of the noise supports our claim that TANDEM representations are noise robust. Angle of deflection for lynx noise is milder than that for the factory noise.

nonlinear cases also, i.e., noise information in the feature space can be assumed to be not along the nonlinear sub-space of maximum class discriminant information. In such case, the transformation by the MLP will lead to reduction of the noise information at its output. Hence the TANDEM representations derived from the outputs of the MLP can be expected to be more robust. This is infact shown to be the case in (Ikbal *et al.*, 2004a,b). In the next section, we further evaluate the noise robustness of the TANDEM representations through noisy speech recognition experiments.

6.4 Experimental evaluation of noise robustness of TANDEM representations

Figures 6.6, 6.7, and 6.8 show results of noisy speech recognition experiments conducted using TANDEM representations of the MFCC feature, denoted by T-MFCC. These experiments are conducted on OGI Numbers95 speech database corrupted by various types of noises from Noisex92 database. Figure 6.6 shows results for various noise levels of factory noise, Figure 6.7 for lynx noise, and Figure 6.8 for car noise. As can be seen from the figures, TANDEM representation of the MFCC feature show a significantly improved robustness to all noise types.

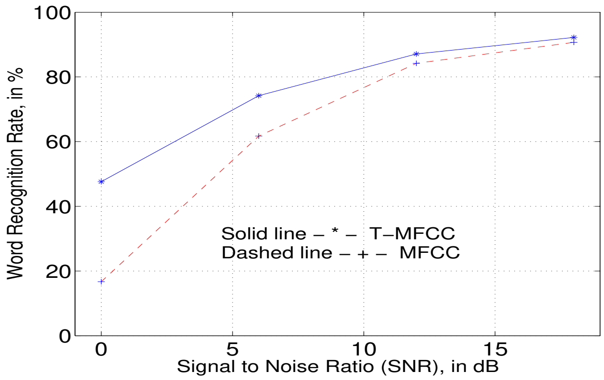

Figure 6.6. Performance comparison between MFCC and its TANDEM representation for various noise levels of factory noise.

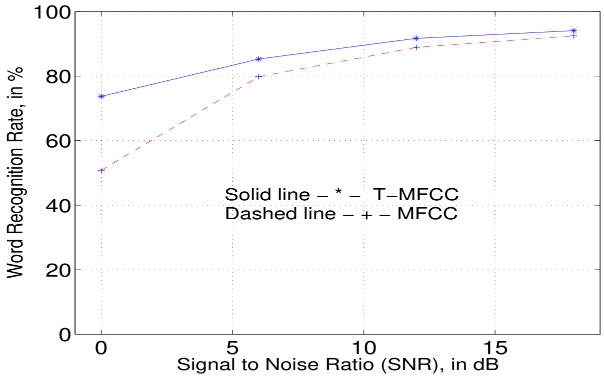

Figure 6.7. Performance comparison between MFCC and its TANDEM representation for various noise levels of lynx noise.

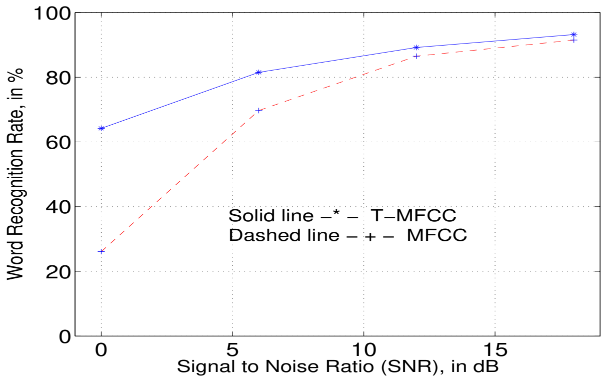

Figure 6.8. Performance comparison between MFCC and its TANDEM representation for various noise levels of car noise.

A comment on clean speech recognition

Apart from the improved noise robustness, in the previous literature, TANDEM representation has also been shown to improve the clean speech recognition performance significantly (Hermansky *et al.*, 2000). The reason for this improvement can also be given in lines similar to that of the noise robustness improvement. The variability caused by various sources that usually cause degradation of clean speech recognition performance, such as the speaker differences, can also expected to be reduced in the TANDEM representation, because of the transformation performed by the MLP. Table 6.1 gives a comparison of the clean speech recognition performances of the MFCC and its TANDEM representation.

Feature	Word Recognition Rate, %
T-MFCC	95.3
MFCC	94.4

Table 6.1. Comparison of the speech recognition performance of the MFCC and TANDEM representation of the MFCC (denoted by T-MFCC) for clean speech of OGI Numbers95 database.

6.5 TANDEM representations of STAP and PAC features

STAP and PAC features stands in as interesting candidates for the TANDEM approach, as their robustness can be improved further in their TANDEM representations (Ikbal *et al.*, 2004a,b). Moreover, with the TANDEM approach, there is a better scope for improving their clean speech recognition performance also. This is because, any of the disturbing variabilities that are introduced by the externally applied transformations during the computation of the STAP and PAC features can be expected to be reduced in their TANDEM representations.

6.5.1 Clean speech recognition

Table 6.2 gives a comparison of the clean speech recognition performances of the L-PSTAP-DP (explained in Section 5.7.1), L-PSTAP-31-HA (explained in Section 5.7.2), and PAC-MFCC (explained in Section 5.4) features and their TANDEM representations, as denoted by T-PSTAP-DP, T-PSTAP-31-HA, and T-PAC-MFCC respectively. An important point to note here is the fact that as the MLP used in TANDEM can better handle the feature correlation and the zeros present in the original STAP features (as explained in Section 4.7), original STAP features are used directly to compute their TANDEM representation, not their LDA transformed versions. As can be seen from the table, the clean speech recognition performances of all the features improve significantly in their TANDEM representations.

Feature	Word Recognition Rate, in %		Feature
	Original feature	TANDEM representation	
L-PSTAP-DP →	84.6	90.7	← T-PSTAP-DP
L-PSTAP-31-HA →	84.1	89.4	← T-PSTAP-31-HA
PAC-MFCC →	87.8	90.0	← T-PAC-MFCC

Table 6.2. Performance comparison between L-PSTAP-DP, L-PSTAP-31-HM, and PAC-MFCC features and their corresponding TANDEM representation, denoted by T-PSTAP-DP, T-PSTAP-31-HA, and T-PAC-MFCC, respectively. An important thing to note here is the fact that TANDEM equivalents of L-PSTAP-DP and L-PSTAP-31-HA are obtained directly from their the corresponding original STAP features, not from the LDA transformed features.

6.5.2 Noisy speech recognition

Figures 6.9 through 6.17 show performance comparison between L-PSTAP-DP, L-PSTAP-31-HA, and PAC-MFCC, and their TANDEM counterparts T-PSTAP-DP, T-PSTAP-31-HA, and T-PAC-MFCC for various noise levels of factory, lynx, and car noises. From the figures, it can be seen that the robustness of these features is further improved in their TANDEM representations. Comparing between L-PSTAP-DP and L-PSTAP-31-HA features (both for the clean and noisy speech recognition performances), except for a few cases, L-PSTAP-DP feature is marginally better. The reason for this could be the fact that HMM/ANN based peak location estimation algorithm used to compute the L-PSTAP-31-HA feature is more sensitive to noise, which can be reasonably understood by the fact that HMM/ANN based method use the actual distribution of the energy values (which varies in the presence of noise) in the spectrum for peak location estimation, whereas frequency-based dynamic programming algorithm do not. In the later chapters, only T-PSTAP-DP is used for further experiments, as it gives the best recognition performance.

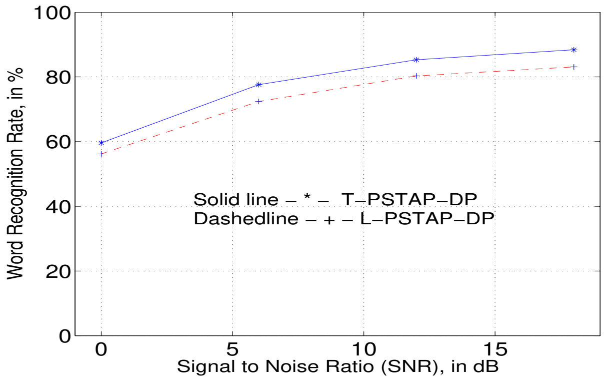

Figure 6.9. Performance comparison between T-PSTAP-DP and L-PSTAP-DP for various noise levels of factory noise.

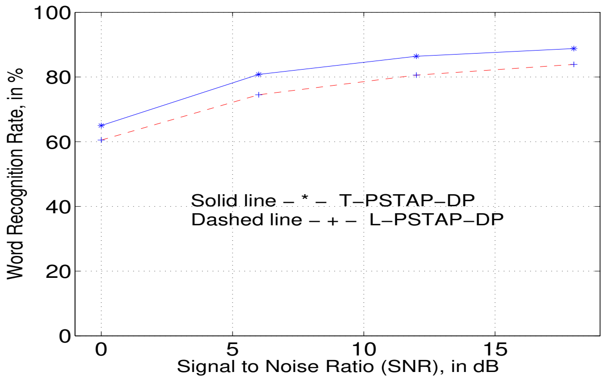

Figure 6.10. Performance comparison between T-PSTAP-DP and L-PSTAP-DP for various noise levels of lynx noise.

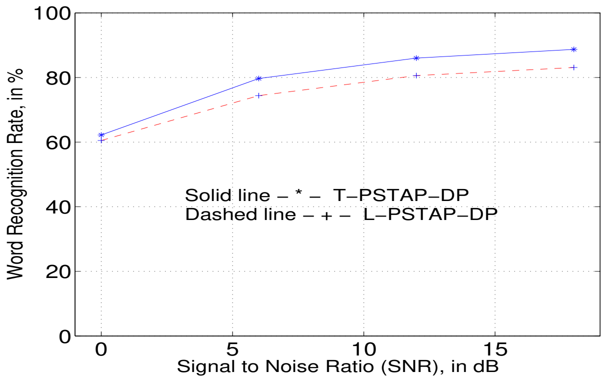

Figure 6.11. Performance comparison between T-PSTAP-DP and L-PSTAP-DP for various noise levels of car noise.

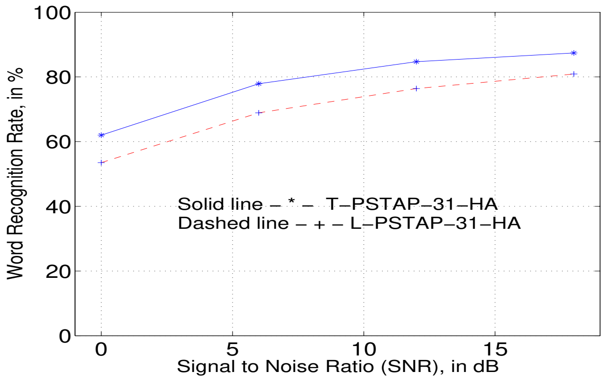

Figure 6.12. Performance comparison between T-PSTAP-31-HA and L-PSTAP-31-HA for various noise levels of factory noise.

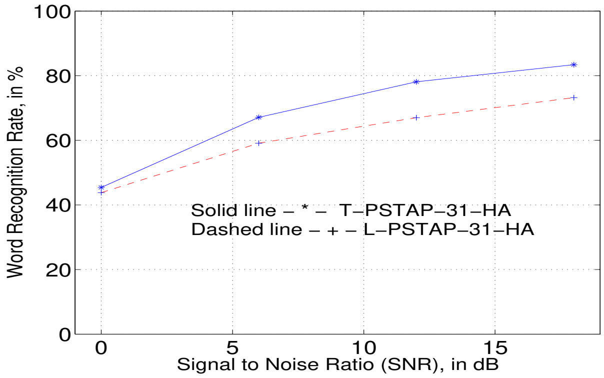

Figure 6.13. Performance comparison between T-PSTAP-31-HA and L-PSTAP-31-HA for various noise levels of lynx noise.

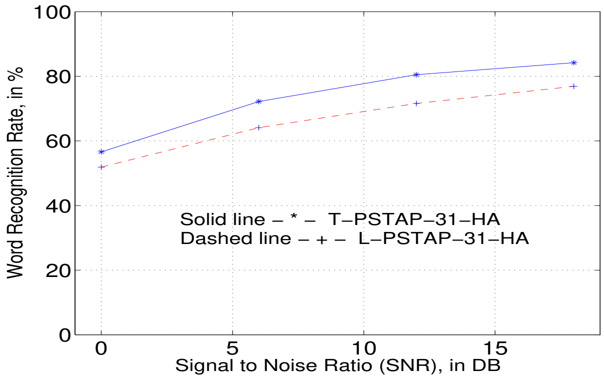

Figure 6.14. Performance comparison between T-PSTAP-31-HA and L-PSTAP-31-HA for various noise levels of car noise.

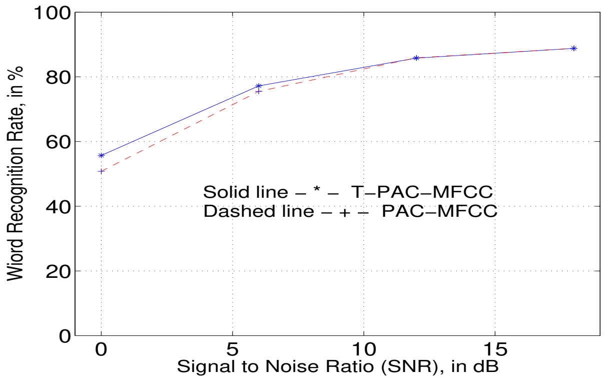

Figure 6.15. Performance comparison between PAC-MFCC and its TANDEM representation for various noise levels of factory noise.

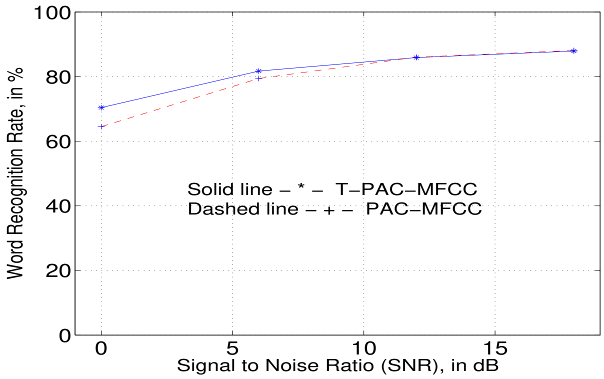

Figure 6.16. Performance comparison between PAC-MFCC and its TANDEM representation for various noise levels of lynx noise.

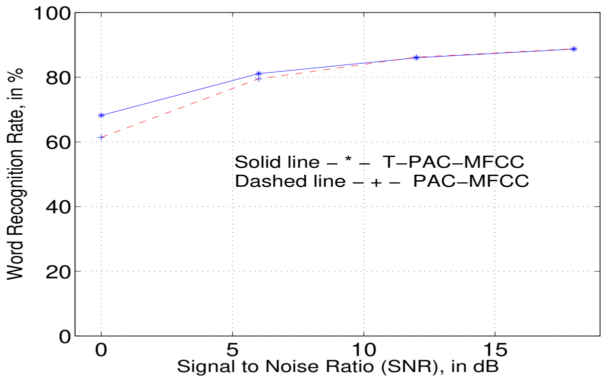

Figure 6.17. Performance comparison between PAC-MFCC and its TANDEM representation for various noise levels of car noise.

6.6 Conclusion

In this chapter, we analyzed and evaluated the TANDEM approach for noisy speech recognition. The feature transformation performed by the MLP used in TANDEM, (which is learned from training set in a data-driven manner, through a discriminative training), is able to keep the clean speech recognition performance intact (infact improved!), while improving the noisy speech recognition of the features. This is in contrast to the externally designed transformations based on the knowledge from human perception system, as we have seen in Chapters 4 and 5, respectively for the STAP and PAC features, where clean speech recognition performance degrades, while improving the noise robustness. This is because during the discriminative training, MLP learns to project the input feature space onto a nonlinear sub-space along the direction of maximum class discriminatory information. Such a projection retains the class discriminatory information while suppressing the noise variability. Evaluating the TANDEM representations of several features (MFCC, STAP, and PAC-MFCC) consistently show that TANDEM representations are more noise robust, with improved clean speech recognition performance also.

The analysis of the TANDEM noise robustness has lead to another interesting use of TANDEM approach, namely, using it as a tool for feature integration, which is explained in the next Chapter. The STAP, PAC, and TANDEM approaches arrive at their own noise robust representations through different processing schemes. This provides a scope for an existence of complementary information between these features. In the next chapter, we explain the combination of these features in a TANDEM framework, utilizing the possible complementary information, to further improve the overall robustness in all conditions.

Chapter 7

Evidence Combination In TANDEM Approach

7.1 Introduction

In this chapter, we explore the possibilities of combining multiple feature streams in TANDEM framework. This topic attains significance in the context of this thesis, from several angles, explained as follows: 1) The nonlinear transformation performed by the MLP, to project the input space onto the maximum possible class discriminatory space (as explained in the previous chapter, in Section 6.3) provides good scope for an adaptive integration of the multiple feature streams at its input. Additionally, as the TANDEM representations (explained in the previous chapter, in Section 6.2) are basically transformations of the posterior probabilities, various multistream adaptive posterior combination techniques developed in the previous literature (Misra *et al.*, 2003) can be utilized to combine the TANDEM representations. 2) As we have seen respectively in Chapters 4, 5, and 6, STAP, PAC, and TANDEM approaches arrive at their own noise robust representations through different processing schemes. This provides good scope for an existence of complementary information between the corresponding features, which can further be utilized to improve the overall robustness through an adaptive combination. 2) As we have seen respectively in Chapters 4 and 5, STAP and PAC approaches, utilizing externally designed transformations for improving the noise robustness, hurt the clean speech recognition performance. An ideal solution to improve their clean speech recognition performance, in such scenario, is to look for an alternative and better source of evidence for clean speech, and combine it with the evidences from STAP and PAC features. As we know, traditional features like MFCC provides better evidence in clean speech. Thus, if the combination framework can adaptively give higher weighting to the evidences from MFCC in clean speech and can give higher weighting to the evidences from the STAP and PAC features in noisy speech, then the resulting system will have an improved recognition performance in all the conditions.

As illustrated in Figure 7.1, traditional methods to combine the feature streams do the combination either at the feature level or at the statistical model level. The combination using TANDEM approach, as described in the next section, comes under the feature level combination.

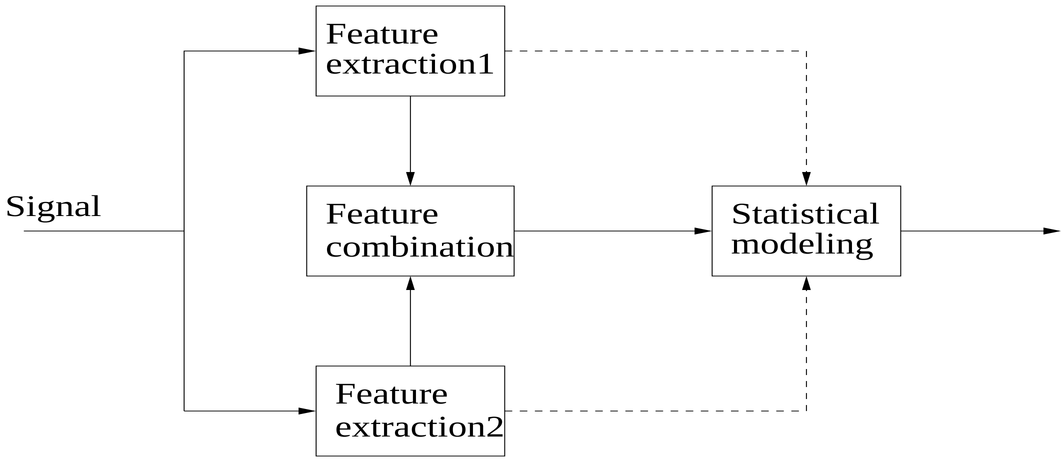

Figure 7.1. Illustration of multiple feature stream combination. Solid line arrows denote the path of feature level combination and dotted line arrows denote the path of statistical model level combination.

7.2 Feature combination in TANDEM framework

TANDEM approach (described in the previous chapter) provides a nice framework for combining multiple feature streams. Using TANDEM approach, the input feature streams can be transformed to a combined TANDEM representation, which can further be used as input feature to the HMM/GMM system. The combined TANDEM representation can be obtained by:

1. Training the MLP with combined (concatenated) feature streams at its input.

2. Adaptive combination of the independent TANDEM representations obtained from each of the feature streams.

Each of these methods are explained in detail in the next two sections.

7.3 Combination at the input of the MLP

Feature combination at the input of the MLP in the TANDEM framework is illustrated in Figure 7.2. In order to learn an appropriate combination of the input feature streams, the MLP is trained with all the feature streams at its input. From the explanation given in the previous chapter (Section 6.3), such a training will make the MLP to learn a transformation that maps the combined multiple feature input space onto a space of maximum class discriminatory information. Thus the output of the MLP is an appropriate combination of the input feature streams. As shown in the illustration figure, the combined TANDEM representation is obtained by decorrelating the pre-nonlinearity output of the MLP using KL transformation.

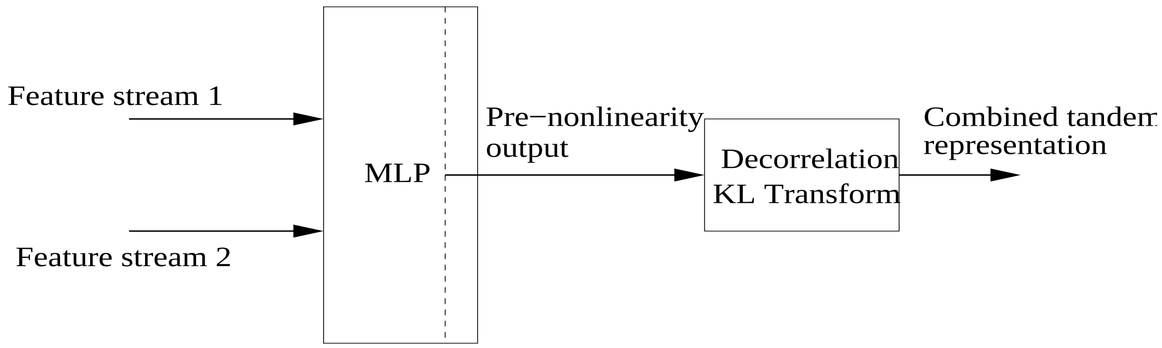

Figure 7.2. Illustration of multiple feature combination at the input of MLP in a TANDEM system. KL transformed prenonlinearity outputs of MLP is the combined TANDEM representation.

7.4 Adaptive combination of individual TANDEM representations

As we have seen in the previous chapter (in Section 6.2), TANDEM representations are basically transformations of the posterior probabilities obtained at the output of a discriminatively trained MLP. Thus, as illustrated in Figure 7.3, various posterior probability combination techniques, reported in the previous literature, can very well be used to combine the individual TANDEM representations also.

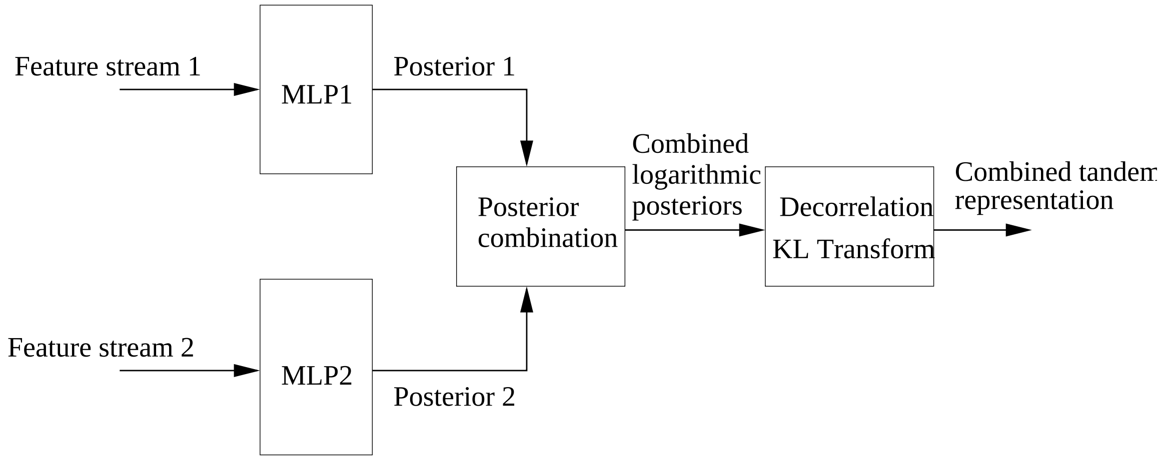

Figure 7.3. Illustration of combination of individual TANDEM representations. Combined TANDEM representation is the combination of the logarithmic posteriors followed by a KL transformation.

Multi-stream posterior combination is explained in the next subsection.

7.4.1 Multi-stream posterior combination

If $\mathbf{x}_i$ denote the i^{th} feature stream and θ_i denote the parameters of the MLP trained with $\mathbf{x}_i$, then the posteriors $P(q_k|\mathbf{x}_i, \theta_i)$ obtained at the MLP outputs can be combined to get the resultant posteriors according to equation:

$$P(q_k|\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_I) = \sum_{i=1}^{I} w_i P(q_k|\mathbf{x}_i, \theta_i) \tag{7.1}$$

where I denote the total number of feature streams to be combined, and w_i denote combination weight assigned to the i^{th} feature stream, satisfying constraint $\sum_{i=1}^{I} w_i = 1$. In the previous literature (Hagen, 2001), it has been shown that instead of combining the direct posteriors the combination of the logarithms of the posteriors, as given by the equation below, is practically more effective (although there is no theoretical justification to it).

$$\log P(q_k|\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_I) = \sum_{i=1}^{I} w_i \log P(q_k|\mathbf{x}_i, \theta_i) \tag{7.2}$$

Taking a close look at the above equations, the direct posterior combination is an arithmetic mean of the posteriors, where as the logarithmic posterior combination is a geometric mean. In this thesis work, we consider only the logarithmic posterior probability combination, as it gives better recognition performance than the direct posterior combination (Ikbal *et al.*, 2004a).

The combination weights, w_i, in (7.1) and (7.2), decides how much to rely upon the corresponding feature streams while computing the combined posterior. The closer the weight value is to 1, the more we rely upon the corresponding feature stream. Thus, adaptively changing the values of these weights depending upon the reliability of the feature streams for each frame is expected to yield the best possible recognition performance in all conditions.

The central problem in multi-stream posterior combination is to estimate the reliability of the feature streams in order to compute the appropriate combination weights. An entropy based method to perform this is explained in next subsection.

7.4.2 Entropy based reliability estimation

Unreliable feature at the input of the MLP leads to uncertain estimation of the posterior probabilities at its output. An information-theoretic measure called entropy gives a measure of the uncertainty in the estimation of the posterior probability distribution. Thus entropy of the posterior probability distribution estimated at the output of the MLP can be used as a measure to estimate the reliability of the feature stream. This has been suggested in (Okawa *et al.*, 1998) and utilized and verified in (Misra *et al.*, 2003), where weights based on entropy values have been assigned to different feature streams in a multi-stream framework for combining the posterior probabilities. The effectiveness of such an entropy based combination of the posterior probabilities has been further demonstrated in (Ikbal *et al.*, 2004). We use a similar method to combine the TANDEM representations of the STAP and PAC features with MFCC feature, which is explained in the next subsection.

7.4.3 Entropy based combination of TANDEM representations

If there are K classes (which in our case corresponds to the number of context-independent phonemes), the entropy of the posterior probability distribution at the output of the MLP for i^{th} feature stream can be computed according to equation:

$$h_i = -\sum_{k=1}^{K} P(q_k|\mathbf{x}_i, \theta_i)\log_2 P(q_k|\mathbf{x}_i, \theta_i) \tag{7.3}$$

The closer the entropy value is to zero, more reliable the posterior probability distribution is. Hence, a normalized inverse value of the entropy can be used to compute the adaptive weighting factor (Misra *et al.*, 2003), as follows :

$$w_i = \frac{1/h_i}{\sum_{j-1}^{I} 1/h_j} \tag{7.4}$$

As illustrated in Figure 7.4, the weights computed based on the entropy values are first used to combine the logarithmic posteriors. Then the combined logarithmic posteriors are decorrelated using KL transformation to obtain the combined TANDEM representation.

In the following sections we discuss the experimental results of the combination of the STAP and PAC features with MFCC using above explained TANDEM-based combination techniques.

7.5 Evaluation of TANDEM-based feature combination

The combination of the STAP and PAC features with MFCC features performed using the above explained combination techniques in TANDEM framework is evaluated using Numbers95 database. STAP features used in the experiments are computed using peak locations estimated from the mel-warped critical bank PAC spectrum using frequency-based dynamic programming algorithm

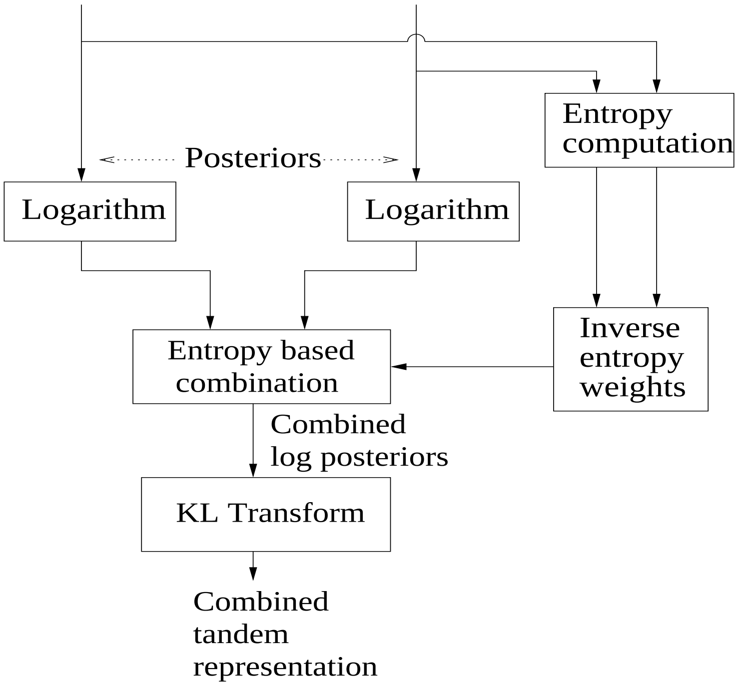

Figure 7.4. Entropy based combination of TANDEM representations.

(resulting in PSTAP-DP feature, as explained in Section 5.7.1)[1]. The performance of the individual features given in the previous chapters are again given in Table 7.1, for easy reference and comparison with the combination performances.

Feature	% Word Recognition Rate for SNR			
	clean	12 dB	6 dB	0 dB
T-PSTAP-DP	90.7	85.3	77.6	59.6
T-PAC-MFCC	90.0	85.8	77.2	55.7
T-MFCC	95.3	87.1	74.2	47.6

Table 7.1. Comparison of the speech recognition performances of TANDEM representations of STAP (T-PSTAP-DP), PAC-MFCC (T-PAC-MFCC) and MFCC (T-MFCC) for clean speech and noisy speech with additive factory noise levels of 12 dB, 6 dB, and 0 dB SNRs.

The results of the combination experiments are presented and discussed in the following two subsections.

[1]From now on, for the noisy speech case, speech corrupted with factory noise is only considered. Also, the results from now on are all given in Tables, because the values we compare are more closer to each other and thus the comparison in Figures wont give a good idea.

7.5.1 Combination at MLP input

Table 7.2 gives results of the experiments when the features are combined at the input of the MLP. Three rows in the table gives performance of combined TANDEM representations of 1) STAP and MFCC, 2) PAC-MFCC and MFCC, and 3) STAP, PAC-MFCC, and MFCC, respectively, for various noise levels of additive factory noise. As can be seen from the table, combination method is able to utilize the complementary information between the feature streams in order to improve the recognition accuracy even better than the best performing feature, for noisy conditions. The combination all the three features is able to achieve a mild improvement over the pairwise combination.

	% Word Recognition Rate for SNR			
Feature	clean	12 dB	6 dB	0 dB
T-PSTAP-DP+MFCC	94.0	88.5	80.1	61.4
T-PAC-MFCC+MFCC	94.9	88.7	79.3	56.5
T-PSTAP-DP+PAC-MFCC+MFCC	94.5	89.1	80.3	61.9

Table 7.2. Speech recognition performances of feature combination in TANDEM framework, at the input of the MLP. T-PSTAP-DP+MFCC represents the combination of PSTAP-DP feature with MFCC, T-PAC-MFCC+MFCC represents the combination of PAC-MFCC with MFCC, and T-PSTAP-DP-PAC-MFCC-MFCC represents the combination of PSTAP-DP, PAC-MFCC, and MFCC. Results shown are for clean speech and factory noise corrupted speech with noise levels of 12 dB, 6 dB, and 0 dB SNRs.

7.5.2 Entropy based combination of TANDEM representations

Table 7.3 gives results of the experiments when the individual TANDEM representations are combined using weights that are computed based on entropy. Again, the three rows in the table gives performance of combinations of individual TANDEM representations of 1) STAP and MFCC, 2) PAC-MFCC and MFCC, and 3) STAP, PAC-MFCC, and MFCC, respectively, for various noise levels of additive factory noise. As can be seen from the table, similar to the combination at MLP input, entropy based combination of individual TANDEM representations also improve the recognition accuracy even better than the best performing feature stream, for noisy conditions. The difference to be noted in the current case is that, logarithmic posteriors are used for computing the combined TANDEM representation, whereas when combination is done at the input, prenonlinearity outputs of the MLP are used for the computation of combined TANDEM representation.

Feature	% Word Recognition Rate for SNR			
	clean	12 dB	6 dB	0 dB
T-PSTAP-DP+T-MFCC	93.6	87.4	79.1	61.1
T-PAC-MFCC+T-MFCC	94.6	88.3	79.2	55.1
T-PSTAP-DP+T-PAC-MFCC+T-MFCC	93.9	89.8	83.2	66.1

Table 7.3. Speech recognition performances of combinations individual TANDEM representations. T-PSTAP-DP+T-MFCC represents the combination of TANDEM representations of the PSTAP-DP feature and the MFCC, T-PAC-MFCC+T-MFCC represents the combination of TANDEM representations of the PAC-MFCC and the MFCC, and T-PSTAP-DP+T-PAC-MFCC+T-MFCC represents the combination of TANDEM representations of the PSTAP-DP, PAC-MFCC, and MFCC features. Results shown are for clean speech and factory noise corrupted speech with noise levels of 12 dB, 6 dB, and 0 dB SNRs.

7.6 Conclusion

In this chapter we have presented two methods to combine multiple feature streams in the TANDEM framework. In the first case, nonlinear transformation performed by the MLP, projecting the input space onto a sub-space of maximum class discriminatory information, is used to combine the concatenated features at the input of the MLP. In the second case, individual TANDEM representations obtained from the feature streams are combined through an entropy based combination techniques. Both methods were shown to be able to utilize complementary information between the feature streams in order to achieve improved recognition performance, especially in noisy conditions.

Moreover, we have also shown that such combination methods are able to alleviate the drawback of the STAP and PAC features, namely, the inferior recognition performance in clean speech, by combining them with the standard features.

In the next chapter, we use a different database (widely used by the noise robustness research community) called Aurora-2, to repeat a few critical experiments conducted throughout this chapter, to see whether the corresponding ideas tested were holding on in another database.

Chapter 8

Experiments on Aurora database

In this chapter, a few critical experiments performed throughout this thesis are repeated on a different database called AURORA-2, to confirm the validity of the corresponding ideas evaluated, on an independent database. A description of the Aurora-2 database and a recognition system supplied with the database for common evaluation of the front-end processing are given in the next two sections section. For comparison purpose, a front-end selected by ETSI Aurora group as ETSI standard for advanced distributed speech recognition (DSR) is described in the following section. Later sections discuss the experiments conducted and the results obtained using techniques developed in this thesis.

8.1 Aurora-2 database

Aurora-2 database (as described in (Hirsch and Pearce, 2000)) is designed to evaluate the performance of speech recognition algorithms in noisy conditions. It is a connected digits database for speaker-independent recognition task. The noisy conditions involve both additive noise condition and combination of additive and convolutional distortions. However, as this thesis deals mainly with additive noise, experiments reported in this chapter are conducted only on the additive noise database.

TIDigits database (Leonard, 1984) is used as a basic speech database for Aurora-2, which contains the recordings of male and female US-American adults speaking isolated digits and sequences of upto 7 digits. To simulate the telephone speech, the original data, sampled at 20 kHz, has been down-sampled to 8 kHz after low-pass filtering the speech to extract the spectral content between 0 and 4 kHz. The low-pass filter used is G.712 standard whose frequency characteristics have been defined by the ITU (Hirsch and Pearce, 2000). The down-sampled, low-pass filtered, data constitute the clean speech data.

8.1.1 Noise description

Various noises are added artificially to the clean speech at various SNR levels to generate the noisy speech versions. The SNRs used are 20dB, 15dB, 10dB, 5dB, and 0dB. The noise types used (representing the most probable telecommunication application scenarios) are: 1) Suburban train, 2) Crowd of people (babble), 3) Car. 4) Exhibition hall, 5) Restaurant, 6) Street, 7) Airport, and 8) Train station. Noises such as car noise and exhibition hall noise are stationary, while street noise and airport noise are non-stationary. The long-term spectral characteristics of these noises can be found from (Hirsch and Pearce, 2000).

8.1.2 Training database

In addition to performing training on clean speech and recognition on noisy speech, Aurora-2 database also allows multicondition training, which involves training of the system with a subset of the noise types mentioned above and recognition with all the noise conditions. However, the multi-condition training do not fall in the scope of this thesis, and hence wont be discussed further. The clean training database consists of 8440 utterances containing recordings of 55 male and 55 female adult speakers.

8.1.3 Test database

The test data consists of 3 sets, the first two constituting data with matched channel conditions and the third set constituting data with channel mismatch conditions. As the channel mismatch condition do not fall in the scope of this thesis, we do not consider the third set. The first two sets are explained as follows: Each set consists of 4004 clean speech utterance divided into 4 parts of 1001 utterances each. Each part is added with one of the noises mentioned in the Section 8.1.1. For the first test set (test set A), noises used are suburban train, babble, car, and exhibition hall. For the second set (test set B), noises used are restaurant, street, airport, and train station. As mentioned above, these noises are added at SNRs of 20dB, 15dB, 10dB, 5 dB, and 0dB. Furthermore, clean speech without adding these noises constitute the sixth test condition. Hence, each test set consists of $1001 \times 4 \times 6 = 24024$ utterances.

8.2 Recognition system

For an identical evaluation of the new noise robust speech extraction schemes a predefined set up of HTK (Young *et al.*, 1992) based recognizer is provided with Aurora database. The experiments reported in this chapter are performed using this recognizer. The description of the recognizer is as follows: whole word HMMs, with 16 states per word, are used to model the digits. The states are connected in a simple left-to-right fashion, without any state skips. Mixture of 3 Gaussians per state, with diagonal covariance matrix, are used to model the emission. In addition to the word models, there are two pause models defined. The first one is called 'sil' consisting of 3 states, to model the pause before and after the utterance, with mixture 6 Gaussian per state. The second pause model is called 'sp' to model the pauses between words, consisting of single state with mixture of 6 Gaussians.

The training procedure for the above system can be found in (Hirsch and Pearce, 2000). During recognition, an utterance is modeled by any sequence of digits with the possibility of a 'sil' model at the beginning and the end and a 'sp' model between two digits.

Reporting the results

General practice used in the previous literature to report results of experiments performed using Aurora-2 database is: 1) to report all the recognition accuracies in a table for all noise conditions and the noise levels, 2) relative improvement, for all the noise conditions and noise levels, in comparison to a baseline system recognition accuracies provided with Aurora-2 database (can be seen in (Hirsch and Pearce, 2000), and 3) average relative overall improvement across all the noise conditions for the noise levels from 20 dB until 0 dB in comparison to the baseline system provided with Aurora-2 database. In this chapter, absolute recognition accuracies (the first case above) and the average relative overall improvement (the third case above) are reported for all the experiments involving techniques developed in this thesis. For comparison purpose, average relative overall improvements obtained with ETSI Aurora standard front-end are reported in the next section.

8.3 ETSI Aurora standard for advanced front-end

In February 2002, ETSI Aurora group has selected a noise robust front-end developed jointly by Motorolla Inc., France telecom, and Alcatel as a ETSI standard for advanced distributed speech recognition (DSR) front-end (Macho *et al.*, 2002). This front-end has been demonstrated to yield best overall performance among all the candidates participated in the standardization task. This front-end basically calculates noise-reduced cepstral features from the incoming digital signal in 4 steps explained as follows:

1. Two-stage mel-warped Wiener filter noise reduction: This is a combination of two-stage Wiener filter scheme developed in (Agarwal and Cheng, 1999) and time-domain noise reduction scheme described in (Noe *et al.*, 2001). It aims to reduce the noise in the signal by two (similar but not identical) passes of wiener filter. This two pass approach gives more flexibility in Wiener filter design to achieve a non-linear behavior difficult to accomplish with a single-pass Wiener filter. The input signal is first filtered with the Wiener filter designed during the first pass and its output signal goes as input signal to the second pass. In each pass the signal spectrum is estimated from a Hanning windowed frame of 25 msec frame length and 10 msec frame shift. Then a power spectral density (PSD) mean is computed by averaging the time-frequency blocks of 2 frames time length and 2 frequency indices. From the resultant spectrum, Wiener filter frequency characteristics is estimated. During the first pass a speech/non-speech decision from energy-based voice activity detection (VAD) is also used in estimating the Wiener filter characteristics. The details of Wiener filter frequency characteristics estimation can be found in (Macho *et al.*, 2002). Having estimated the frequency characteristics, the Wiener filter impulse response is obtained using a mel-warped inverse cosine transform. Then the denoised signal is obtained by convolving the noisy input signal with the Wiener filter impulse response. During the second stage an additional operation called gain factorization is performed where an aggressive noise reduction is performed for the purely noisy frames and

less aggressive noise reduction is performed for the frames containing speech.

2. SNR-dependent waveform processing: This uses the fact that SNR within the noisy speech period is variable, which is because in the voiced segments of the speech signal the speech waveform exhibits quasi-periodic maxima due to the glottal excitation. Thus the high SNR portions of the waveform are emphasized and the low SNR waveform portions are deemphasized by a weighting function (Macho and Cheng, 2001).

3. Cepstrum computation: Cepstrum computation differs from the regular cepstrum computation by the fact that the preemphasis coefficient used is 0.9 instead of 0.97 and the power spectrum is used instead of magnitude spectrum in the filter-bank integration.

4. Blind equalization: This relies on the least mean square algorithm (Mauuary, 1998), which minimizes the mean square error computed as a difference between the current and the target cepstrum. The target cepstrum corresponds to the cepstrum of a flat spectrum. This step reduces the convolutional distortion caused by the use of different microphones in training of acoustic models and testing.

The cepstral features computed with the above steps are further appended with the energy coefficient and the delta features and passed through a feature selection module before being used for recognition. In the feature selection module, features from non-speech regions of the speech are dropped, because they cause most of the insertion errors as a result of mismatch between the non-speech portion of the signal and the silence model.

Table 8.1 gives the percentage overall relative improvement obtained with the ETSI Aurora standard front-end when compared to the baseline feature performance provided along with the Aurora-2 database.

Feature	% improvement over Aurora-2 front end		
	Test set A	Test set B	Average
ETSI Aurora standard	70.04	74.94	72.44

Table 8.1. Average relative improvement (over 4 kinds of noise in each set, and SNR varying from 20 dB to 0 dB) achieved by the ETSI Aurora standard for noise robust front end over the baseline feature provided by the Aurora-2.

8.4 Recognition performance on Aurora-2 database

This section presents and discusses a repetition of the experiments presented throughout this thesis, on Aurora-2 database. In order to confirm the trend observed earlier on OGI Numbers95, just a few representative cases are taken and evaluated. The experiments performed are with the following features:

1. MFCC features, to serve as a baseline.
2. PAC-MFCC features, as explained in Section 5.5.
3. TANDEM representations of the MFCC features (T-MFCC), as explained in Section 6.4.
4. TANDEM representation of the PSTAP-DP features (T-PSTAP-DP), as explained in Section 6.5.
5. TANDEM representation of the PAC-MFCC feature (T-PAC-MFCC), as explained in Section 6.5.
6. Combination of the MFCC and PAC-MFCC features in the TANDEM framework, at the input of the MLP, as explained in Section 7.5.1.

Table 8.2 give a summary of all the results. It gives a percentage overall relative improvement obtained with features considered when compared to the baseline features provided by the Aurora-2 database, as mentioned in 8.2. As can be seen from the results, the MFCC feature we use is able to perform better than the baseline provided with Aurora-2 database. The reason for this could be that the MFCC features we use are cepstral mean normalized. Again from the Table 8.2, PAC-MFCC is more robust in noise conditions than the MFCC.

Similar to the trends observed with OGI Numbers95 database, T-PSTAP-DP and T-PAC-MFCC features show improved noise robustness. However, interestingly, a difference in the trend when compared to the trend observed in Numbers95 database is the that T-MFCC is much superior when compared to the other features. The reason for this is the fact that speech utterance of Aurora-2 are close to the microphone speech as the channel conditions are simulated by a filter, whereas Numbers95 is a realistic telephone speech database. Hence, class discriminatory information can be expected to be more prominent in Aurora-2 than the Numbers95. Thus TANDEM is able to utilize it better and show a significant improvement in the performance when compared to the other features. An improvement in robustness is observed also in T-PAC-MFCC, but not as prominent as the T-MFCC because the nonlinear transformation performed during the computation of the PAC-MFCC disturbs the speech class discriminatory information, which is shown by its inferior clean speech recognition performance.

The combination of MFCC and PAC-MFCC in the TANDEM framework, at the MLP input, is able to utilize complementary information and improve the robustness further. Comparing the combination results with the results of ETSI Aurora standard front-end given in Table 8.1, it can be seen that the techniques proposed in this thesis are able to achieve a reasonably good robustness, although the ETSI Aurora standard front-end remains the best.

An interesting point to note here is the following: the training of the MLP in TANDEM requires phoneme alignment information of the speech utterance. However, Aurora-2 do not provide phoneme alignment information. Hence, MLP trained on OGI Numbers95 database is used to extract the TANDEM features in Aurora-2 database. As can be observed from the Table 8.2 such features are still able to achieve good improvement in the robustness. A similar trend is reported in (Sivadas and Hermansky, 2004).

Tables 8.3 through 8.14 give word recognition accuracies of the features considered for all noise types and all noise levels.

Feature	% improvement over Aurora-2 front end		
	Test set A	Test set B	Average
MFCC	10.55	32.19	22.09
PAC-MFCC	24.64	34.23	29.75
T-MFCC	49.21	59.46	54.68
T-PSTAP-DP	23.51	36.69	30.54
T-PAC-MFCC	38.21	39.80	39.06
T-MFCC+PAC-MFCC	55.31	63.36	59.61

Table 8.2. Average relative improvement (over 4 kinds of noise in each set, and SNR varying from 20 dB to 0 dB) achieved by the noise robust techniques explored in this thesis over the baseline feature provided by the Aurora-2. The last line gives results of combination of MFCC and PAC-MFCC features in a TANDEM framework, when combination is performed at the MLP input.

SNR, in dB	Word Recognition Rate, in %				
	Subway	Babble	Car	Exhibition	Average
clean	98.80	99.09	98.63	99.20	98.93
20	96.28	97.40	96.63	96.33	96.66
15	90.76	93.83	91.02	90.47	91.52
10	73.13	81.71	71.85	72.91	74.90
5	42.98	52.42	36.27	38.97	42.66
0	22.47	24.00	20.19	18.76	21.36
Average (20dB-0dB)	65.12	69.87	63.19	63.49	65.42

Table 8.3. Word recognition rate for test set A of Aurora-2 database while using MFCC feature.

SNR, in dB	Word Recognition Rate, in %				
	Restaurent	Street	Airport	Train-station	Average
clean	98.80	99.09	98.63	99.20	98.93
20	97.42	96.92	97.17	97.59	97.28
15	94.50	92.35	94.21	93.12	93.55
10	83.82	76.21	84.52	79.67	81.06
5	57.60	46.31	56.19	48.69	52.20
0	27.51	22.67	29.53	23.82	25.88
Average (20dB-0dB)	72.17	66.89	72.32	68.58	69.99

Table 8.4. Word recognition rate for test set B of Aurora-2 database while using MFCC feature.

SNR, in dB	Word Recognition Rate, in %				
	Subway	Babble	Car	Exhibition	Average
clean	97.42	97.55	97.55	97.90	97.61
20	95.15	90.72	95.38	94.17	93.86
15	92.39	86.49	91.53	91.67	90.52
10	82.44	76.30	83.72	80.93	80.85
5	59.81	59.19	63.02	51.16	58.30
0	32.85	32.95	31.37	26.07	30.81
Average (20dB-0dB)	72.53	69.13	73.00	68.80	70.87

Table 8.5. Word recognition rate for test set A of Aurora-2 database while using PAC-MFCC feature.

8.5 Conclusion

Throughout this thesis, OGI Numbers95 database has been used to evaluate the approaches developed. In this chapter, a few representative experiments are repeated on a new database called Aurora-2 to confirm their validity. Conclusions based on the outcomes of the experiments are as follows:

SNR, in dB	Word Recognition Rate, in %				
	Restaurent	Street	Airport	Train-station	Average
clean	97.42	97.55	97.55	97.90	97.61
20	88.39	95.19	90.55	92.72	91.71
15	82.28	92.50	85.77	89.26	87.45
10	72.64	84.85	76.32	81.24	78.76
5	57.14	65.60	60.33	62.82	61.47
0	33.83	36.09	34.92	35.45	35.07
Average (20dB-0dB)	66.86	74.85	69.58	72.30	70.89

Table 8.6. Word recognition rate for test set B of Aurora-2 database while using PAC-MFCC feature.

SNR, in dB	Word Recognition Rate, in %				
	Subway	Babble	Car	Exhibition	Average
clean	98.86	98.64	98.84	99.20	98.89
20	96.87	97.52	97.41	96.64	97.11
15	94.14	95.34	95.94	93.30	94.68
10	87.26	90.27	91.23	85.13	88.47
5	73.13	74.33	79.72	68.44	73.91
0	49.49	45.95	54.46	40.76	47.67
Average (20dB-0dB)	80.18	80.68	83.75	76.85	80.37

Table 8.7. Word recognition rate for test set A of Aurora-2 database while using TANDEM representation of MFCC (T-MFCC) feature.

SNR, in dB	Word Recognition Rate, in %				
	Restaurent	Street	Airport	Train-station	Average
clean	98.86	98.64	98.84	99.20	98.89
20	97.27	96.98	97.44	97.72	97.35
15	94.53	95.04	95.56	96.17	95.33
10	88.27	89.99	90.99	91.18	90.11
5	71.85	75.45	78.68	78.25	76.06
0	45.38	51.78	55.06	55.59	51.45
Average (20dB-0dB)	79.46	81.85	83.55	83.38	82.06

Table 8.8. Word recognition rate for test set B of Aurora-2 database while using TANDEM representation of MFCC (T-MFCC) feature.

- The trend observed are similar to the ones observed in OGI Numbers95 database.
- An interesting case which drew our attention is, TANDEM representation of the MFCC is significantly better than the other features. As we have explained in Section 8.4, the reason for this is the fact Aurora-2 is more closer to the microphone speech (because the telephone

SNR, in dB	Word Recognition Rate, in %				
	Subway	Babble	Car	Exhibition	Average
clean	97.27	97.43	97.08	97.10	97.22
20	94.38	90.02	92.66	93.89	92.74
15	89.96	87.45	91.11	89.26	89.45
10	77.25	80.68	85.39	76.55	79.97
5	55.45	62.30	66.93	50.42	58.78
0	28.68	35.97	35.94	24.31	31.23
Average (20dB-0dB)	69.14	71.28	74.41	66.89	70.43

Table 8.9. Word recognition rate for test set A of Aurora-2 database while using TANDEM representation of PSTAP-DP (T-PSTAP-DP) feature.

SNR, in dB	Word Recognition Rate, in %				
	Restaurent	Street	Airport	Train-station	Average
clean	97.27	97.43	97.08	97.10	97.22
20	88.30	95.01	86.58	91.14	90.26
15	85.45	91.23	83.30	89.82	87.45
10	76.48	80.32	76.86	83.99	79.41
5	60.15	63.51	63.08	69.73	64.12
0	36.32	35.52	38.26	44.59	38.67
Average (20dB-0dB)	69.34	73.12	69.62	75.85	71.98

Table 8.10. Word recognition rate for test set B of Aurora-2 database while using TANDEM representation of PSTAP-DP (T-PSTAP-DP) feature.

SNR, in dB	Word Recognition Rate, in %				
	Subway	Babble	Car	Exhibition	Average
clean	97.36	97.19	97.61	97.28	97.36
20	95.95	89.60	95.35	94.42	93.83
15	93.18	84.92	92.13	91.27	90.38
10	87.63	76.21	86.40	83.18	83.36
5	74.49	60.79	75.01	65.26	68.89
0	51.55	36.73	49.30	38.88	44.12
Average (20dB-0dB)	80.56	69.65	79.64	74.60	76.11

Table 8.11. Word recognition rate for test set A of Aurora-2 database while using TANDEM representation of PAC-MFCC (T-PAC-MFCC) feature.

channel effect is simulated by a filter, whereas OGI Numbers95 database it collected on real telephone channels) and hence the speech class discriminatory information is expected to be more prominent in Aurora-2 than the Numbers95. Thus TANDEM is able to utilize it better and show a significant improvement in the performance.

SNR, in dB	Word Recognition Rate, in %				
	Restaurent	Street	Airport	Train-station	Average
clean	97.36	97.19	97.61	97.28	97.36
20	88.03	95.44	88.43	92.16	91.02
15	82.56	92.62	83.24	89.82	87.06
10	73.17	85.91	74.56	83.77	79.35
5	58.43	78.88	60.13	69.70	65.54
0	37.92	49.43	39.87	48.13	43.84
Average (20dB-0dB)	68.02	79.46	69.25	76.72	73.86

Table 8.12. Word recognition rate for test set B of Aurora-2 database while using TANDEM representation of PAC-MFCC (T-PAC-MFCC) feature.

SNR, in dB	Word Recognition Rate, in %				
	Subway	Babble	Car	Exhibition	Average
clean	99.14	98.73	99.02	99.17	99.02
20	97.45	97.64	97.52	97.35	97.49
15	95.70	96.10	96.12	94.94	95.72
10	89.10	91.02	92.84	88.55	90.38
5	76.91	77.09	82.67	73.03	77.43
0	53.82	48.34	61.29	46.99	52.61
Average (20dB-0dB)	82.60	82.04	86.09	80.17	82.72

Table 8.13. Word recognition rate for test set A of Aurora-2 database while using combination of MFCC and PAC-MFCC, in a TANDEM framework at the input of the MLP (T-MFCC + T-PAC-MFCC feature).

SNR, in dB	Word Recognition Rate, in %				
	Restaurent	Street	Airport	Train-station	Average
clean	99.14	98.73	99.02	99.17	99.02
20	97.70	97.58	97.26	98.12	97.67
15	95.00	96.01	95.68	96.85	95.89
10	89.16	90.60	91.47	93.18	91.10
5	74.12	78.78	80.20	82.69	78.95
0	48.08	54.87	57.17	61.18	55.33
Average (20dB-0dB)	80.81	83.57	84.36	86.40	83.79

Table 8.14. Word recognition rate for test set B of Aurora-2 database while using combination of MFCC and PAC-MFCC, in a TANDEM framework at the input of the MLP (T-MFCC + T-PAC-MFCC feature).

Chapter 9

Conclusion

9.1 Summary and conclusions

This thesis proposed a few new feature-based approaches for improving the noise robustness of automatic speech recognition systems. The central idea behind the development of these approaches is the fact that the noise robustness can be improved by emphasizing the part of the speech that is relatively more noise robust and/or deemphasizing or masking the part that is relatively more noise sensitive. Nonlinear transformations that perform such emphasis and deemphasis of different parts of speech, when applied to the spectrum or feature, have been explored.

Such a formulation of the approaches developed require a division of the speech into two components, one more robust to the noise and the other more sensitive to noise, so that they can be treated differently. This thesis explored two different possibilities of performing such division, namely 1) external division based on the knowledge about the speech, and 2) estimation of the division in a data-driven manner. Initial approaches used external division based on the knowledge that the peaks in spectral domain constitute the high signal to noise ratio (SNR) part of the speech. Later on, for the data-driven estimation, speech part corresponding to the sound class discriminatory information is used as the part that is relatively more robust to noise.

Considering the external division case first, where the spectral peaks are assumed to constitute relatively more robust part of the speech, two different strategies followed for the enhancement of the spectral peaks and the deemphasis of the spectral valley have lead to two different approaches.

In the first approach, the non-peak regions in the spectrum are completely masked to zeros, whereas in the second approach, a soft-masking procedure is followed, where non-peak regions of the spectrum are not discarded but are smoothed out. The first approach requires explicit specification of the peak locations in the spectrum in order to mask the non-peak locations. This thesis proposed two peak location estimation algorithms:

1. frequency-based dynamic programming (DP) algorithm,

2. HMM/ANN based algorithm.

Both the algorithms are motivated by a previous work on spectral peak estimation, where a frequency-based HMM, in general framework called HMM2 (Weber *et al.*, 2003a), is used for the estimation task. The frequency-based DP algorithm use spectral slope values of single time frame for the peak location estimation, while HMM/ANN based algorithm use, more general, distinct time-frequency (TF) patterns in the spectrogram for the same task. The use of TF patterns impose temporal constraints during the peak estimation. Both the algorithms differ from the previous peak estimation algorithms by the fact that the number of peak locations estimated is not fixed apriori. Some conclusions drawn about these peak estimation algorithms are as follows:

- The use of TF patterns require an unsupervised learning of the distinct TF patterns in the spectrogram. This makes the HMM/ANN based algorithm sensitive to the energy fluctuations in the TF patterns.

- Whereas the frequency-based DP algorithm is simple and independent of energy fluctuations, which gives it a marginal advantage over the HMM/ANN based algorithm.

- In the later part of the thesis it has been shown that an use of the energy normalized spectrogram with enhanced spectral peaks and smoothed spectral valley (referred to as PAC spectrogram) is able to improve the performance of the HMM/ANN based peak location estimation. However, frequency-based DP algorithm again performs marginally better in the PAC spectrum also.

Assuming the existence of reliable peak location estimates, the first approach proposed in this thesis to improve the noise robustness, namely the spectro-temporal activity pattern (STAP) approach, uses spectral components only from the regions around the spectral peaks for computing the features, referred to as STAP features. For an effective utilization of the information from the spectral components around the peaks, STAP approach has drawn inspiration from the outcomes of physiological studies conducted on mammalian auditory cortex, which show evidences for a processing of local time-frequency patterns by the cortical neurons (Depireux *et al.*, 2001). Accordingly, STAP features use parameters describing activity pattern (energy surface) within local time-frequency patterns around the spectral peaks as features. Experimental evaluation of the STAP features has lead to the following conclusions.

- STAP features are robust to noise.

- Interestingly, utilizing the spectral components only from regions around the spectral peaks is able to produce reasonable recognition performance. With an incorporation of more and more time-frequency pattern describing parameters the clean speech recognition performance also improves.

- However, the clean speech recognition performance of the STAP features is significantly inferior when compared to the standard features. The main reason for this is the complete masking of the non-peak regions of the spectrum, which also seem to carry significant information required for clean speech recognition.

- Additionally, the recognition performance of the STAP feature critically depends upon the proper functioning of the peak location estimation algorithm used, which is, in fact, evident from the different recognition recognition performances achieved with two different peak estimation algorithm.

The second approach proposed to improve the noise robustness is phase autocorrelation (PAC) approach, where a soft-masking procedure is used instead of discarding the non-peak spectral components completely. Additionally, the explicit peak location estimation is also avoided, thereby saving the features from being sensitive to the peak location estimation algorithm. In actual, PAC approach addresses the problem of noise robustness at the autocorrelation domain, as opposed to most of the other approaches which work at the spectral domain. Autocorrelation is a time-domain Fourier equivalent of the power spectrum. PAC approach uses phase (i.e., angle) variation of the signal vector over time as a measure of correlation, as opposed to the regular autocorrelation which uses dot product. This alternative measure of autocorrelation is referred to as PAC, and the features derived from it are referred to as PAC features. The use of angle is motivated by the fact that angle gets less disturbed in the presence of additive disturbance than the dot product. Interestingly, the use of PAC has an effect of emphasizing the peaks and smoothing out the valley, in the spectral domain, without explicitly estimating their locations. Experimental evaluation of the PAC features resulted in following conclusions:

- Soft-masking of the spectral valley components is better than the hard masking, which completely discards them.
- PAC features exhibit improved noise robustness.
- However, the softmasking also tends to degrade the clean speech recognition performance, resulting in inferior clean speech recognition performance of the PAC features when compared to the standard features.
- This points to a fact that externally designed transformations, which do not completely take into account the underlying complexity of the speech signal, may not be able to improve the robustness without hurting the clean speech recognition performance.
- Apart from the above, interestingly, the use of PAC spectrum in the peak location estimation task for the computing the STAP features has resulted in improved performances of STAP features both for clean and noisy speech. This is because the PAC spectrum is energy normalized and has enhanced spectral peaks and smoothed spectral valleys, thereby providing a nice alternative to the regular spectrum for reliable peak location estimation.

Considering the conclusions of the above approaches (degradation in clean speech), a better approach for improving the noise robustness will be to learn the relatively noise invariant part of speech from the speech data itself, in a data-driven manner, compromising between improving the noise robustness while keeping the clean speech recognition performance intact. An existing data-driven approach called TANDEM has been analyzed to validate this. The TANDEM approach uses MLP to perform a data-driven transformation of the input feature vectors. This transformation is learned by training the MLP in a supervised, discriminative mode, with phoneme labels as output classes. Such a training makes the MLP to perform a nonlinear discriminant analysis in the input feature space and thus makes it to learn a transformation that projects the input features onto a sub-space of maximum class discriminatory information. This projection is able to suppress the noise related variability, while keeping the speech discriminatory information intact. The analysis and experimental evaluation of the TANDEM approach for noise robustness has lead to the following conclusions.

- TANDEM approach is effective in improving the noise robustness, while also keeping the clean speech recognition performance intact, thus validating our claim that projecting onto the sub-space of maximum class discriminatory information would make a nice compromise between improving the clean speech recognition performance while keeping the clean speech recognition performance intact.

- Interestingly, (as also shown in the previous literature (Hermansky *et al.*, 2000)) TANDEM improves the clean speech recognition performance also significantly. The main reason for this could be the fact that similar to the noise variability, other variabilities that usually cause degradation in clean speech, such as speaker variability, are also suppressed during the projection onto the sub-space of maximum sound class discrimination.

- Interestingly, TANDEM is able to improve both the clean and noisy speech recognition performances of the STAP and PAC features, again because of the above mentioned reasons.

- More interestingly, as shown in Chapter 8, in Aurora-2 database the TANDEM representation of the MFCC feature is able to improve its noise robustness significantly than that of the STAP and PAC-MFCC features. This is in contrast to the results obtained with OGI Numbers95 database, as given in Chapter 6. The reason for this is the fact that Aurora-2 is closer to the microphone speech (because the telephone channel effect is simulated by a filter, whereas OGI Numbers95 database is collected on real telephone channels) and hence the speech class discriminatory information is more prominent in Aurora-2 database and are better utilized when used with TANDEM approach.

The noise robustness analysis of TANDEM has resulted in another interesting aspect of it namely, using it as an integration tool for combining the multiple feature streams. The transformation performed by the MLP has been used to adaptively combine the input feature streams by projecting the combined (concatenated) input space onto the sub-space of maximum class discriminatory information. Additionally, as the outputs of the MLP (that is trained in a discriminative mode) are basically posteriors, entropy based posterior combination (Misra *et al.*, 2003) has also been used to combine the features in TANDEM framework. Both the combination methods were used to combine the noise robust features developed in this thesis such as STAP and PAC features with standard feature such as MFCC. Experimental evaluation of the feature combination in TANDEM framework has resulted in conclusions as follows:

- Both the methods of combination (combination at the input of the MLP and the combination of posteriors) are able to utilize complementary information between the feature streams to improve over best performing stream in noisy conditions.
- The combination methods serves as a nice way of alleviating the drawbacks of STAP and PAC features to improve their clean speech recognition performance, while improving the noise robustness also.

9.2 Overall conclusions

Major overall conclusions from the work for this thesis are the following:

- Improvements in robustness obtained with the feature-based approaches developed in this thesis (based on nonlinear transformations of the spectrum or features) validates our hypothesis that an improvement in noise robustness can be achieved by emphasizing the part of the speech that is relatively more invariant to noise and/or deemphasizing the part of the speech that is relatively more sensitive to noise.
- Initial approaches, where an external division of the speech into relatively noise invariant and noise sensitive parts has been done, showed an improvement in noise robustness while resulting in inferior performance in clean speech when compared to the standard features. This points to the fact that externally designed transformations, which do not take a complete account of the underlying complexity of the speech signal, may not be able to improve the robustness without hurting the clean speech recognition performance.

- Analysis of TANDEM approach for the case of noise robustness points to the fact that, data-driven approaches provide a nice solution for the improving the noise robustness of the speech features, as they are able to make a nice compromise between improving the noise robustness while keeping the clean speech recognition performance intact. More specifically, in the TANDEM approach, projecting the input feature space onto a sub-space of maximum possible class discriminatory information is able to suppress the noise related variability, while keeping the speech discriminatory information intact.

- TANDEM provides a nice framework for combining multiple feature streams. An adaptive combination of multiple features streams in the TANDEM framework, result in a system that is uniformly robust in all the conditions, especially while combining noise robust features (which are good in noisy speech while not so good in clean speech) with standard features (which are good in clean speech but not so good in noisy speech).

9.3 Potential future directions

- Better parameterization of time-frequency patterns in STAP, instead of using the simple pattern describing parameters such as energy level, delta and acceleration of energy along time and frequencies.

- Adaptively varying the time-frequency pattern size to handle the underlying variabilities along time and frequency.

- Exploring appropriate output classes for the TANDEM, that will be able to do the noise suppression more effectively.

- Exploring the use of broader temporal context for the TANDEM that has potential to improve the noise robustness more effectively and also improve the clean speech recognition performance.

Appendix A

Linear Discriminant Analysis (LDA)

Suppose, there are N D-dimensional vectors, $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, belonging to C different classes. Let N_c denote the number of vectors in subset $\mathcal{X}_c$ that belong to class c. Then the aim of linear discriminant analysis (LDA) is to find a projection matrix of order $D \times (C-1)$, that when used to project the D-dimensional vector space to a space of dimension $C-1$, will retain the class discriminant information to a maximum possible extent. This can be achieved is the data points in the projected space have highest possible ratio between the between-class scatter and the within-class scatter.

The within-class scatter matrices in the original vector space is defined as follows:

$$S_w = \sum_{c=1}^{C} S_c$$

$$S_c = \sum_{\mathbf{x} \in \mathcal{X}_c} (\mathbf{x} - \mathbf{m}_c)(\mathbf{x} - \mathbf{m}_c)^{\mathrm{T}}$$

$$\mathbf{m}_c = \frac{1}{N_c} \sum_{\mathbf{x} \in \mathcal{X}_c} \mathbf{x}$$

Between-class scatter matrix:

$$S_b = \sum_{c=1}^{C} N_c(\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^{\mathrm{T}}$$

$$\mathbf{m} = \frac{1}{N}\sum_{c=1}^{C} N_c \mathbf{m}_c$$

Suppose W denote matrix of order $D \times (C-1)$ used to transform D-dimensional data to dimension $C-1$, using equation:

$$\mathbf{y} = W^{\mathbf{T}}\mathbf{x}$$

Then the scatter matrices $\tilde{S_w}$ and $\tilde{S_b}$ in the transformed space are given by:

$$\tilde{S_w} = W^{\mathbf{T}} S_w W$$

$$\tilde{S_b} = W^{\mathbf{T}} S_b W$$

Then the criterion function to find out the projection matrix is given as follows:

$$J(W) = \frac{|\tilde{S_b}|}{|\tilde{S_w}|} = \frac{|W^{\mathbf{T}} S_w W|}{|W^{\mathbf{T}} S_b W|}$$

Maximizing $J(W)$ with respect to W will give W that retains the maximum class discriminatory information. The solution for the above equation turns out to be as follows: The columns of W are the generalized eigen vectors that correspond to the largest eigen values of

$$S_b \mathbf{w}_i = \lambda S_w \mathbf{w}_i$$

Bibliography

A. Agarwal, and Y. M. Cheng, (1999) "Two-Stage Mel-Warped Wiener Filter for Robust Speech Recognition," in *Proc. of ASRU-99*, 1999.

P. Alexandre, B. Boudy, and P. Lockwood, (1993), "Root Homomorphic Deconvolution Schemes for Speech Processing in Car Environments," in *Proc. of ICASSP-93*, vol. 2, 1993, pp. 99-102.

J. B. Allen, (1994), "How Do Humans Process and Recognize Speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no: 4, pp. 567-577, 1994.

H. Athineos, and D.P.W. Ellis, (2003), "Frequency-Domain Linear Prediction for Temporal Features," in *Proc. of IEEE ASRU 2003 Workshop*, St. Thomas, Virgin Islands, USA. Nov-Dec, 2003.

X. Aubert, R. Haeb-Umbach, and H. Ney, (1993), "Continuous Mixture Densities and Linear Discriminant Analysis for Improved Context-Dependent Acoustic Models," in *Proc. of ICASSP-93*, vol. 2, pp. 648-651, Apr. 1993.

L.E. Baum, T. Petrie, G. Soules, and N. Weiss, (1970), "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains", *Ann. Math. Statistic*, Vol. 41, pp. 164-171, 1970.

S. Bengio, H.Bourlard, K. Weber, (2000), "An EM Algorithm for HMMs with Emission Distribution Represented by HMMs," *IDIAP Research Report*, IDIAP-RR 00-11, IDIAP Research Institute, Martigny, Switzerland, May 2000.

J. A. Bilmes, (1998), "A Gentle Tutorial of the EM algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models". *Report*, ICSI, Berkeley, CA, USA. TR-97-021, Apr. 1998.

C. Bishop, (1995), "Neural Networks for Pattern Recognition," *Claredon Press*, Oxford, 1995.

S. F. Boll, (1979), "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," in *Proc. of IEEE ASSP-27*, Apr.1979, pp. 113-120.

H. Bourlard, Y. Kamp, H. Ney, and C. J. Wellekens, (1985), "Speaker-Dependent Connected Speech Recognition via Dynamic Programming and Statistical Methods," in *Speech and Speaker Recognition (eds., M. R. Schroeder and Gottingen) Karger*, Basel, no. 2, pp. 115-148, 1985.

H. Bourlard, and Y. Kamp, (1988), "Autoassociation by Multi-Layer Perceptrons and Singular Value Decomposition," *Biological Cybernetics*, vol.59, pp. 291-294, 1988.

H. Bourlard, and N. Morgan, (1993), "Connectionist Speech Recognition: A Hybrid Approach," *The Kluwer International Series in Engineering and Computer Science*, Kluwer Academic Publishers, Boston, USA, vol. 247. 1993.

H. Bourlard, S. Dupont, H. Hermansky, and N. Morgan, (1996), "Towards Sub-Band Based Speech Recognition," in *Proc. of European Signal Processing Conference*, Trieste, Italy, pp. 1579-1582, Sep. 1996.

H. Bourlard, and S. Dupont, (1996a), "ASR Based on Independent Processing and Recombination of Partial Frequency Bands," in *Proc. of ICSLP-96*, Philadelphia, USA, Oct. 1996.

H. Bourlard, S. Dupont, and C. Ris, (1996b), "Multi-Stream Speech Recognition," *FPMS-TCTS*, Dec. 1996.

R. Cole, M. Noel, T. Lander, and T. Durham, (1995), "New Telephone Speech Corpora at CSLU," *in Proceedings of European Conference on Speech Communication and Technology*, 1995, vol. 1, pp. 821-824.

M. Cooke, A. Morris, and P. Green, (1997), "Missing Data Techniques for Robust Speech Recognition," in *Prof. of ICSLP-97*, pp. 863-866, 1997.

M. Cooke, P. Green, L. Josifovski, and A. Vizinho, (2001), "Robust Automatic Speech Recognition with Missing and Unreliable Data," *Speech Communication*, 34, pp. 267-285, 2001.

S. B. Davis, and P. Mermelstein, (1980), "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 28, 1980, pp. 357-366.

A. P. Dempster, N. M. Laird, and D. B. Rubin, (1977), "Maximum-Likelihood from Incomplete Data via the EM Algorithm". *Journal of Royal Statistical Society B*. Vol. 39, 1-38, 1977.

R. Duda, and P. Hart, (1973), "Pattern Classification and Scene Analysis," *John-Wiley & Sons*, 1973.

D. Depireux, J. Simon, D. Klein, and S. Shamma, (2001), "Spectro-Temporal Response Field Characterization with Dynamic Ripples in Ferret Primary Auditory Cortex." in *Journal of Neurophysiology*, 2001, 85:1220-1234.

D. Ellis, R. Singh, and S. Sivadas, (2001), "Tandem Acoustic Modeling in Large-Vocabulary Recognition," *in Proc. of ICASSP-01*, Salt Lake City, Utah, USA, May 2001.

H. Fletcher, (1953), "Speech and Hearing in Communication," *Van Nostrand*, New York, (ASA Edition, ed. J. B. Allen), 1953.

K. I. Funahashi, (1994), "On the Approximate Realization of Continuous Mapping by Neural Networks," *Neural Networks*, vol. 2, pp. 183-192, 1989.

S. Furui, (1986), "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Feb. 1986.

S. Furui, (1992), "Towards Robust Speech Recognition Under Adverse Conditions," in *Proc. of ESCA Workshop on Speech Processing in Adverse Conditions*, Cannes, France, 1992, pp. 31-41.

M. J. F. Gales, and S. J. Young, (1996), "Robust Continuous Speech Recognition Using Parallel Model Combination," *IEEE Trans. on Speech and Audio Processing*, 4, 1996, pp. 352-359.

M. J. F. Gales, (1996a), "Model Based Techniques for Noise Robust Speech Recognition," *Ph.D. Thesis*, Cambridge University, 1996.

R. Haeb-Umbach, and H. Ney, (1992), "Linear Discriminant Analysis for Improved Large Vocabulary Speech Recognition," in *Proc. of ICASSP-92*, vol. 1, pp. 13-16, Mar. 1992.

A. Hagen, A. Morris, and H. Bourlard, (1998), "Sub-Band Based Speech Recognition in Noisy Conditions: The Full Combination Approach," *IDIAP Research Report*, IDIAP, Martigny, Switzerland, IDIAP-RR 15, 1998.

A. Hagen, (2001), "Robust Speech Recognition Based on Multi-Stream Processing," *Thesis*, EPFL, Lausanne, Switzerland. Dec. 2001.

S. Haykin, (1994), "Neural Networks," *New York: MacMillan*, 1994.

H. Hermansky, (1990), "Perceptual Linear Predictive (PLP) Analysis for Speech," *Journal of Acoustic Society of America*, 1990, pp. 1738-1752.

H. Hermansky, and N. Morgan, (1994), "RASTA Processing of Speech," *IEEE Transactions on Speech and Audio Processing*, Oct. 1994, Vol.2, No:4, pp. 578-589.

H. Hermansky, D. Ellis, and S. Sharma, (2000), "Tandem Connectionist Feature Extraction for Conventional HMM Systems," *in Proc. of ICASSP-00*, Istanbul, June 2000.

H. Hermansky, (2004), "TRAP-TANDEM: Data Driven Extraction of Temporal Features from Speech," in *Proc. of IEEE ASRU 2003 Workshop*, St. Thomas, Virgin Islands, USA. Nov-Dec, 2003.

H. G. Hirsch, and D. Pearce, (2000), "The AURORA Experimental Framework for The Performance Evaluations of Speech Recognition Systems Under Noisy Conditions," *ISCA ITRW ASR2000 - Automatic Speech Recognition: Challenges for The Next Millennium*, Paris, France, Sep. 2000.

S. Ikbal, H. Misra, and B. Yegnanarayana, (1999), "Analysis of Autoassociative Mapping Neural Networks," in *Proc. of IJCNN-99*, Washington, Jul. 1999.

S. Ikbal, (1999a), "Autoassociative Neural Network Models for Speaker Verification," *M. S. Thesis*, Dept. of Comp. Sci. & Engg., IIT, Madras, India, May 1999.

S. Ikbal, H. Bourlard, S. Bengio, and K. Weber, (2001), "IDIAP HMM/HMM2 System: Theoretical Basis and Software Specifications", *IDIAP Research Report*, IDIAP, Martigny, Switzerland. IDIAP-RR 01-27, Oct. 2001. ftp://ftp.idiap.ch/pub/reports/2001/rr01-27.ps.gz

S. Ikbal, K. Weber, H. Bourlard, (2002), "Speaker Normalization Using HMM2," in *Proc. of IEEE NNSP-02 Workshop*, Martigny, Switzerland, pp. 647-656, Sep. 2002.

S. Ikbal, H. Misra, and H. Bourlard, (2003), "Phase AutoCorrelation (PAC) derived robust speech features," in *Proc. of ICASSP-03*, Hong Kong, Apr. 2003, II-133–II-136.

S. Ikbal, H. Hermansky, and H. Bourlard, (2003a), "Nonlinear Spectral Transformations for Robust Speech Recognition," in *Proc. of IEEE ASRU 2003 Workshop*, St. Thomas, Virgin Islands, USA. Nov-Dec, 2003.

S. Ikbal, H. Misra, H. Bourlard, and H. Hermansky, (2004), "Phase AutoCorrelation (PAC) Features in Entropy Based Multi-Stream for Robust Speech Recognition," in *Proc. of ICASSP-04*, Montreal, Canada, May. 2004.

S. Ikbal, H. Misra, S. Sivadas, H. Hermansky, and H. Bourlard, (2004a), "Entropy Based Combination of Tandem Representations for Noise Robust ASR," to appear in *Proc. of INTERSPEECH-ICSLP-04*, Jeju Island, Korea, Oct. 2004.

S. Ikbal, M. Magimai.-Doss, H. Misra, and H. Bourlard, (2004b), "Spectro-Temporal Activity Pattern (STAP) Features for Robust ASR," to appear in *Proc. of INTERSPEECH-ICSLP-04*, Jeju Island, Korea, Oct. 2004.

S. Ikbal, H. Misra, H. Bourlard, and H. Hermansky, (2004c), "Phase AutoCorrelation (PAC) Features for Noise Robust ASR," *IDIAP Research Report*, IDIAP-RR 04-40, (also submitted to *IEEE Trans. on Speech and Audio Processing*), IDIAP Research Institute, Martigny, Switzerland, Jul. 2004.

S. Ikbal, H. Bourlard, and M. Magimai.-Doss, (2004d), "HMM/ANN Based Spectral Peak Location Estimation for Noise Robust Speech Recognition," *IDIAP Research Report*, IDIAP-RR 04-50, (also submitted to *Proc. of ICASSP-05*), IDIAP Research Institute, Martigny, Switzerland, Sep. 2004.

J. C. Junqua, and J. P. Haton, (1996), "Robustness in Automatic Speech Recognition: Fundamentals and Applications," *Kluwer Academic Publishers*, Boston, USA, 1996.

D. Klatt, (1976), "A Digital Filter-bank for Spectral Masking," in *Proc. of ICASSP-86*, 1986, pp. 741-744.

G. E. Kopec, (1986), "Formant Tracking Using Hidden Markov Models and Vector Quantization," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, no. 4, pp. 709-729, 1986.

C. J. Legetter, and P. C. Woodland, (1995), "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression," in *Proc. of ARPA Workshop on Spoken Language Systems Technology*, 1995, pp. 110-115.

R. G. Leonard, (1984), "A Database for Speaker Independent Digit Recognition," in *ICASSP-84*, vol. 3, pp. 42.11 1984.

J. S. Lim, (1979), "Spectral Root Homomorphic Deconvolution System," *IEEE Trans. on ASSP*, 27, 1979, pp. 223-233.

R. P. Lippmann, (1997), "An Introduction to Computing with Neural Nets," *IEEE ASSP Magazine*, pp. 4-22, Apr. 1987.

P. Lockwood, and J. Boudy, (1992), "Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the Projection of Robust Speech Recognition in Cars," *Speech Communication*, 11, pp. 215-228, 1992.

P. Lockwood, J. Boudy, and M. Blanchet, (1992a), "Nonlinear Spectral Subtraction (NSS) and Hidden Markov Models for Robust Speech Recognition in Car Noise Environments," in *Proc. of ICASSP-92*, 1992, pp. 265-268.

P. Lockwood, and P. Alexandre, (1994), "Root Adaptive Homomorphic Deconvolution Schemes for Speech Recognition in Noise," in *Proc. of ICASSP-94*, vol. 1, pp.441-444, 1994.

D. Macho, L. Mauuary, B. Noe, Y. M. Cheng, D. Ealey, D. Jouver, H. Kelleher, D. Pearce, F. Saadoun, (2002), "Evaluation of A Noise-Robust DSR Front-End on Aurora Databases," in *Proc. of ICSLP-02*, Denver, USA, 2002.

D. Macho and Y. M. Cheng, (2001), "SNR-Dependent Waveform Processing for Robust Speech Recognition," in *Proc. of ICASSP-01*, 2001.

J. Makhoul, (1975), "Linear Prediction: A Tutorial Review," *Proc. IEEE*, vol.63, pp. 561-580, Apr. 1975.

D. Mansour, and B. H. Juang, (1988), "A Family of Distortion Measures based upon Projection Operation for Robust Speech Recognition," in *Proc. of ICASSP-88*, 1988, pp. 36-39.

L. Mauuary, (1998), "Blind Equalization in the Cepstral Domain for Robust Telephone based Speech Recognition," in *Prof. of EUSIPCO-98*, vol. 1, pp. 359-363, 1998.

S. McCandless, (1974), "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-22, pp. 135-141, newblock 1974.

B. Mellor, and A. Varga, (1993), "Noise-Masking in a Transform Domain," in *Proc. of ICASSP-93*, pp. II 87-90, 1993.

H. Misra, H. Bourlard, and V. Tyagi, (2003), "New Entropy based Combination Rules in HMM/ANN Multi-Stream ASR," in *Proc. of ICASSP-03*, Hong Kong, Vol.II, Apr. 2003, 741-744.

H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky, (2004), "Spectral Entropy Based Features for Robust ASR," in *Proc. of ICASSP-04*, Montreal, Canada, May 2004.

B. C. J. Moore, (1997), "An Introduction to The Psychology of Hearing," *Academic Press*, New York, 1997.

A. Morris, A. Hagen, H. Glotin, and H. Bourlard, (2001), "Multi-Stream Adaptive Evidence Combination for Noise Robust ASR," *Speech Communication*, 34, pp. 25-40, 2001.

H. Ney, (1984), "The Use of A One-Stage Dynamic Programming Algorithm for Connected Word Recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-32, pp. 263-271, 1984.

B. Noe *et al.*, (2001), "Noise Reduction for Noise Robust Feature Extraction for Distributed Speech Recognition," in *Proc. of Eurospeech-01*, 2001.

J. Nolazco-Flores, and S. Young, (1994), "Continuous Speech Recognition in Noise Using Spectral Subtraction and HMM Adaptation," in *Proc. of ICASSP-94*, 1994, pp. 409-412.

H. Nyquist, (1928), "Certain-Topics in Telegraph Transmission Theory," *AIEE Transactions*, pp. 617, 1928.

S. Okawa, E. Bocchieri, and A. Potamianos, (1998), "Multi-Band Speech Recognition in Noisy Environments," in *Proc. of ICASSP-98*, Seattle, Washington, USA, May 1998.

A. V. Oppenheim, and R. W. Schafer, (1975), "Digital Signal Processing," *PTR Prentice-Hall, Inc.*, Englewood Cliffs, NJ, USA, 1975.

D. O'Shaughnessy, (1987), "Speech Communication - Human and Machine," *Addison-Wesley*, 1987.

L. R. Rabiner, and R. W. Schafer, (1978), "Digital Processing of Speech Signals," *PTR Prentice-Hall, Inc.*, Englewood Cliffs, NJ, USA, 1978.

L. R. Rabiner, and B. H. Juang, (1993), "Fundamentals of Speech Recognition," *PTR Prentice-Hall, Inc.*, Englewood Cliffs, NJ, USA, 1993.

B. Raj, R. Singh, and R. M. Stern, (1998) "Inference of Missing Spectrographic Features for Robust Speech Recognition," in *Proc. of ICSLP-98*, Sydney, Australia, Nov. 1998.

B. Raj, M. L. Seltzer, R. M. Stern, (2001) "Robust Speech Recognition: The case for Restoring Missing Features," in *Proc. of the Workshop on Consistent and Reliable Acoustic Cues*, Aalborg, Denmark, Sep. 2001.

D. E. Rumelhart, G. E. Hinton, and R. T. Williams, (1988), "Learning Internal Representations by Error Propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, Eds., Cambridgg, MA: MIT Press, vol. 1, pp. 318-362, Reprinted in Anderson and Rosenfield, 1988.

C. Shannon, and W. Weaver, (1949), "The Mathematical Theory of Communication,", *University of Illinois Press*, Urbana, 1949.

S. Sharma, (1999), "Multi-Stream Approach to Robust Speech Recognition," *Ph.D. Thesis*, OGI, Portland, USA, Apr. 1999.

K. Shikano, (1987), "Improvement of Word Recognition Results by Trigram Models," in *Proc. of ICASSP-87*, Dallas, Texas, USA, pp. 1261-1264, Apr. 1987.

S. Sivadas, and H. Hermansky, (2004), "On the Use of Task Independent Training Data in Tandem Feature Extraction," in *Proc. of ICASSP-04*, Montreal, Canada, May 2004.

T. A. Stephenson, M. Mathew, and H. Bourlard, (2003), "Speech Recognition with Auxiliary Information," *IEEE Transactions on Speech and Audio Processing*, vol 12, no:3, pp. 189-203, May 2003.

R. M. Stern, A. Acero, F. H. Liu, and Y. Ohshima, (1996), "Signal Processing for Robust Speech Recognition," *Chapter in Speech Recognition*, C. H. Lee, F. Soong, Eds., Kluwer Academic Publishers, Boston, USA, pp. 351-378, 1996.

R. M. Stern, B. Raj, P. J. Moreno, (1997), "Compensation for Environmental Degradation in Automatic Speech Recognition," in *Proc. of the ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-au-Monsson, France, pp. 33-42, Apr. 1997.

B. Strope, and A. Alwan, (1998), "Robust Word Recognition Using Threaded Spectral Peaks," in *Proc. of ICASSP-98*, Seattle, vol. II, pp. 625-629, May 1998.

V. Tyagi, I. McCowan, H. Misra, and H. Bourlard, (2003), "Mel-Cepstrum Modulation Spectrum (MCMS) Features for Robust ASR." in *Proc. of IEEE ASRU 2003 Workshop*, St. Thomas, Virgin Islands, USA. Nov-Dec, 2003.

A. Varga, and K. Ponting, (1989), "Control Experiments on Noise Compensation in Hidden Markov Model Based Continuous Word Recognition," in *Proc. of Eurospeech-89*, 1989, pp. 167-170.

A. P. Varga, and R. K. Moore, (1990), "Hidden Markov Model Decomposition of Speech and Noise," in *Proc. of ICASSP-90*, 1990, pp. 845-848.

A. Varga, H. Steeneken, M. Tomlinson, and D. Jones, (1992), "The NOISEX-92 study on the affect of additive noise on automatic speech recognition," *Technical report*, DRA Speech Research Unit, Malvern, England, 1992.

K. Weber, S. Bengio, H. Bourlard, (2000), "HMM2 - A Novel Approach to HMM Emission Probability Estimation," in *Proc. of ICSLP-00*, Beijing, China, vol. III, pp. 147-150, 2000.

K. Weber, S. Bengio, and H. Bourlard, (2001), "A Pragmatic View of the Application of HMM2 for ASR," *IDIAP Research Report*, IDIAP-RR 00-11, IDIAP Research Institute, Martigny, Switzerland, 2001.

K. Weber, S. Bengio, and H. Bourlard, (2002), "Increasing Speech Recognition Noise Robustness with HMM2," in *Proc. of ICASSP-02*, Orlando, Florida, USA, vol. 1, pp. 929-932, 2002.

K. Weber, (2003), "HMM Mixtures (HMM2) for Robust Speech Recognition," *Ph.D. Thesis*, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, 2003.

K. Weber, S. Ikbal, S. Bengio, and H. Bourlard, (2003a), "Robust Speech Recognition and Feature Extraction Using HMM2," *Computer, Speech, & Language*, vol. 17/2-3, pp.195-221, 2003.

L. Welling, and H. Ney, (1998), "Formant Estimation for Speech Recognition, *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 1, pp. 36-48, newblock Jan. 1998.

S. Young, et. al, (1992), "The HTK book," version 3.2, Cambridge University Engineering Department, England, 1992.

Curriculum Vitae

Shajith Ikbal

Permanent address:	No:10, Ezhil Nagar,	Phone:	+91-44-26493331/1443
	Poonamallee, Madras,	email:	ikbal@idiap.ch
	PIN-600056, Tamil Nadu, India.	Citizenship:	Indian

Education

2000– Docteur ès Sciences (anticipated November 2004).
The Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.
Thesis title: *Nonlinear Feature Transformations for Noise Robust Speech Recognition*.

1997–1999 Master of Science (by Research) in Computer Science.
Department of Computer Science & Engineering., Indian Institute of Technology (IIT), Madras, India.
Thesis title: *Autoassociative Neural Network Models for Speaker Verification*.

1993–1996 Bachelor of Technology in Instrumentation Engineering,
Madras Institute of Technology, Anna University, Madras, India.

1990–1993 Bachelor of Science in Physics.
University of Madras, Madras, India.

Professional Experience

2000–	IDIAP Research Institute, Martigny, Switzerland. Speech Group, Research Assistant.
1999-2000	Speech and Software Technology (India) Pvt. Ltd., Madras, India. Speech Research & Development Group, System Analyst.
1999-1999	Hughes Software Systems Gurgaon, India. DSP Division, Software Engineer.
1997-1999	Indian Institute of Tech., (IIT), Madras, India. Dept. of Comp. Sci. & Engg., Teaching Assistant.
1996-1996	Larsen & Toubro Ltd., Madras, India. Electrical & Instrumentation Division, Graduate Engineering Trainee.

Computer Experience

Operating Systems:	Linux, UNIX, Windows
Languages:	C, C++
Packages:	Matlab, HTK, Tcl/Tk.

Publications

S. Ikbal, M. Magimai.-Doss, H. Misra, and H. Bourlard, (2004b), "Spectro-Temporal Activity Pattern (STAP) Features for Robust ASR," to appear in *Proc. of INTERSPEECH-ICSLP-04*, Jeju Island, Korea, Oct. 2004.

S. Ikbal, H. Misra, S. Sivadas, H. Hermansky, and H. Bourlard, (2004a), "Entropy Based Combination of Tandem Representations for Noise Robust ASR," to appear in *Proc. of INTERSPEECH-ICSLP-04*, Jeju Island, Korea, Oct. 2004.

M. Magimai.-Doss, T. Stephenson, S. Ikbal, and H. Bourlard, (2004), "Modeling Auxiliary Features in Tandem Systems," to appear in *Proc. of INTERSPEECH-ICSLP-04*, Jeju Island, Korea, Oct. 2004.

S. Ikbal, H. Misra, and H. Bourlard, (2003), "Phase AutoCorrelation (PAC) Features in Entropy Based Multi-Stream for Robust Speech Recognition," in *Proc. of ICASSP-04*, Montreal, Canada, May. 2004.

S. Ikbal, H. Hermansky, and H. Bourlard, (2003a), "Nonlinear Spectral Transformations for Robust Speech Recognition," in *Proc. of IEEE ASRU 2003 Workshop*, St. Thomas, Virgin Islands, USA. Nov-Dec, 2003.

S. Ikbal, H. Misra, and H. Bourlard, (2003), "Phase AutoCorrelation (PAC) derived robust speech features," in *Proc. of ICASSP-03*, Hong Kong, Apr. 2003, II-133–II-136.

K. Weber, S. Ikbal, S. Bengio, and H. Bourlard, (2003a), "Robust Speech Recognition and Feature Extraction Using HMM2," *Computer, Speech, & Language*, vol. 17/2-3, pp.195-221, 2003.

S. Ikbal, K. Weber, H. Bourlard, (2002), "Speaker Normalization Using HMM2," in *Proc. of IEEE NNSP-02 Workshop*, Martigny, Switzerland, pp. 647-656, Sep. 2002.

S. Ikbal, H. Bourlard, S. Bengio, and K. Weber, (2001), "IDIAP HMM/HMM2 System: Theoretical Basis and Software Specifications", *IDIAP Research Report*, IDIAP, Martigny, Switzerland. IDIAP-RR 01-27, Oct. 2001. ftp://ftp.idiap.ch/pub/reports/2001/rr01-27.ps.gz

S. Ikbal, (1999a), "Autoassociative Neural Network Models for Speaker Verification," *M. S. Thesis*, Dept. of Comp. Sci. & Engg., IIT, Madras, India, May 1999.

S. Ikbal, H. Misra, and B. Yegnanarayana, (1999), "Analysis of Autoassociative Mapping Neural Networks," in *Proc. of IJCNN-99*, Washington, Jul. 1999.

H. Misra, S. Ikbal, and B. Yegnanarayana, (2003), "Speaker-Specific Mapping for Text-Independent Speaker Recognition," *Speech Communication*, vol. 39, pp. 301-310, Feb. 2003.

H. Misra, S. Ikbal, and B. Yegnanarayana, (1999), "Spectral Mapping as a Feature for Speaker Recognition," in *Proc. of National Conference on Communication*, Kharagpur, India, pp. 151-156, Jan. 1999.

Unpublished Technical Reports

S. Ikbal, H. Misra, H. Bourlard, and H. Hermansky, (2004c), "Phase AutoCorrelation (PAC) Features for Noise Robust ASR," *IDIAP Research Report*, IDIAP-RR 04-40, (also submitted to *IEEE Trans. on Speech and Audio Processing*), IDIAP Research Institute, Martigny, Switzerland, Jul. 2004.

S. Ikbal, H. Bourlard, and M. Magimai.-Doss, (2004d), "HMM/ANN Based Spectral Peak Location Estimation for Noise Robust Speech Recognition," *IDIAP Research Report*, IDIAP-RR 04-50, (also submitted to *Proc. of ICASSP-05*), IDIAP Research Institute, Martigny, Switzerland, Sep. 2004.

S. Ikbal, H. Bourlard, S. Bengio, and K. Weber, (2001), "IDIAP HMM/HMM2 System: Theoretical Basis and Software Specifications", *IDIAP Research Report*, IDIAP, Martigny, Switzerland. IDIAP-RR 01-27, Oct. 2001. ftp://ftp.idiap.ch/pub/reports/2001/rr01-27.ps.gz