IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

MULTI-STREAM ADAPTIVE EVIDENCE COMBINATION FOR NOISE ROBUST ASR

Andrew Morris, Astrid Hagen, Hervé Glotin, Hervé Bourlard

IDIAP-RR 99-26

January 2000

ACCEPTED FOR PUBLICATION in Speech Communication, 2000

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 - 27 - 721 77 11
fax +41 - 27 - 721 77 12
email secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP-RR 99-26

MULTI-STREAM ADAPTIVE EVIDENCE COMBINATION FOR NOISE ROBUST ASR

Andrew Morris, Astrid Hagen, Hervé Glotin, Hervé Bourlard

January 2000

Abstract

In this paper we develop different mathematical models in the framework of the multi-stream paradigm for noise robust ASR, and discuss their close relationship with human speech perception. Largely inspired by Fletcher's "product-of-errors" rule in psychoacoustics, multi-band ASR aims for robustness to data mismatch through the exploitation of spectral redundancy, while making minimum assumptions about noise type. Previous ASR tests have shown that independent sub-band processing can lead to decreased recognition performance with clean speech. We have overcome this problem by considering every combination of data sub-bands as an independent data stream. After introducing the background to multi-band ASR, we show how this "full combination" approach can be formalised, in the context of HMM/ANN based ASR, by introducing a latent variable to specify which data sub-bands in each data frame are free from data mismatch. This enables us to decompose the posterior probability for each phoneme into a reliability weighted integral over all possible positions of clean data. This approach offers great potential for adaptation to rapidly changing and unpredictable noise.

Acknowledgements: This work was supported by the Swiss Federal Office for Education and Science (OFES) in the framework of both the EC/OFES SPHEAR (SPeech, HEAring and Recognition) project and the EC/OFES RESPITE project (REcognition of Speech by Partial Information TEchniques).

Contents

1. Introduction[1]

Even low levels of mismatch between field data and data used in system training, due to noise or channel distortion, can still cause a sharp loss in ASR (Automatic Speech Recognition) performance. As a result, present ASR technology is still inadequate for the majority of applications in which human-like performance is required in an ambient noisy environment[2]. Conventional processing can deal with some kinds of data mismatch quite successfully. Slowly varying noise or channel mismatch effects can be removed by conventional robust preprocessing techniques, such as spectral subtraction for additive noise, or cepstral mean subtraction for convolutive noise. Stationary noise can also be reduced by noise modelling techniques, such as PMC [12, 44], and speaker variation can be effectively accommodated through a combination of training with large vocabulary multi-speaker databases, and speaker adaptation. However, the mismatch which remains after these techniques have been applied often causes an unacceptable loss in recognition performance.

One effect of the frequency analysis which is applied in the first stages of both natural and artificial speech processing, is that the level of redundancy in the resulting two dimensional spectro-temporal representation is greatly increased. Another effect is that the data from different sources in the resulting "auditory scene" is largely separated. Conventional robust preprocessing can further increase this separation. Recognition experiments with band limited noise in both psychoacoustics [1, 11] and ASR [32, 25] have shown that narrow sub-bands of clean data can often be sufficient for speech recognition. There is therefore great potential for improved noise robustness with any system which is able to detect local data mismatch and focus recognition on the clean speech data which remains.

In Section 2 we introduce the advantages of multi-stream processing in general. In Section 3 we discuss evidence from auditory physiology and psychoacoustics for multi-channel processing in ASR. In Section 4 we briefly discuss the main historical approaches to multi-band ASR, and their associated limitations. One of these limitations is that independent sub-band processing reduces recognition performance in clean speech. In Section 5 we introduce the "full combination" approach to multi-band ASR, in which the assumption of sub-band independence is overcome by introducing a latent variable which permits us to integrate over all possible positions of missing (i.e. strongly mismatching) data. In this way the full-band discriminant is decomposed into a mismatch weighted sum of clean-data discriminants over all sub-band combinations. The effectiveness of this approach is however dependent on the weights given to each sub-band combination discriminant, and in Section 6 we present a number of different techniques for estimating these weights. In Section 7 we test some variations of this model on a free-format numbers recognition task, under different noise conditions. These results, and the problems which they raise, are discussed in Section .

2. Multi-stream processing

Multi-expert systems arise in many different fields of data classification and function approximation in general. These systems have a number of proven theoretical and practical advantages, of which the following are of particular relevance to ASR:

---

1. *This paper is a combined and extended version of two papers [4, 16] which were presented in the Tampere workshop "Robust Methods for Speech Recognition in Adverse Conditions".*

2. *This is contrary to the surprisingly prevalent idea that "robust speech recognition is now a solved problem", which the speech technology industry would of course like its clients to believe.*

- **Hierarchical systems of experts reduce problem perplexity:** Unsupervised-training can be used to train a hierarchical system of experts together with a gating network for expert selection. The gating network may be trained to use large scale features for expert selection, so that each expert is trained on a subregion of the input data space, with correspondingly reduced perplexity [23].
- **Linear combination of multiple experts can improve generalisation:** Linear combinations of estimators have been studied and used by the statistics community for a long time. When expert outputs are linearly combined (even as a simple average), the expected committee error will decrease, both in theory [3], and in practice [38]. This error will also decrease further if the spread of the experts' predictions can be increased without increasing the expected errors of the individual members[1]. Different experts can be obtained by varying the parametric function used to model each expert, and/or by varying the data which is used to train each expert, either by using different features from the same data, or different noise added to the same data.

3. Multi-stream processing in human speech recognition

In any recognition process it is advantageous to constructively combine as many sources of information[2] as are available [30]. An example from human perception is that the auditory system is "wired" to combine visual with acoustic information even at the subconscious level, so that perceived phoneme category is directly influenced by lip movements [26, 29]. Experiments in ASR have demonstrated that combining mouth shape with acoustic data can strongly improve recognition performance with noisy speech [10, 13, 41, 46].

Further evidence for the use of multiple experts in the mammalian auditory system is seen at the first stage of central auditory processing. In the cochlear nucleus each fibre in the auditory nerve splits and carries the same data through about seven different types of specialised nerve cell, each having a very different characteristic response. The outputs from these cells are recombined at higher levels of processing [36].

While investigating the effects of band limited noise on human hearing, Fletcher [11] found that the error rate for human phoneme perception using the full frequency range was approximately equal to the product of the error rate using high-pass filtered speech, with the error rate using low-pass filtered speech at the same cutoff frequency. Furthermore, this error rate was independent of the cutoff frequency used. A generalisation of this rule to more than two sub-bands was more recently popularised by Allen in [1], and is now commonly known as the Fletcher-Allen principle, or "product-of-errors" rule (PoE rule):

In human perception, the error rate for full-band perception is equal to the product of the sub-band error rates obtained through perception of each sub-band on its own.

$$P(error) = \prod_i P(error_i) \tag{1}$$

Under the assumption of sub-band error independence, it follows from this rule that:

Full-band classification is incorrect if and only if classification is incorrect in every sub-band

1. *The optimal linear expert combination can be obtained as a function of the error covariance matrix.*
2. *Sometimes also known as "multiple cues" [29].*

or equivalently:

Full-band classification is correct if and only if classification is correct in any sub-band

Figure 1 shows how the probability of correct classification, under the product-of-error rule, is distributed as a function of the probability of correct classification in each of two sub-bands. The product-of-errors rule serves as proof of existence for a system which combines multiple guesses at the speech information with an infallible mechanism for selecting the correct guess when it is present. Although the accuracy of the product-of-errors rule has more recently been questioned [40], this powerful recognition paradigm has strongly motivated the development of multi-band ASR. Of course, as the number of sub-bands is increased each sub-band gets narrower and the error rate for recognition within each isolated sub-band will increase. However, in Table 4 it is shown that, if a correct answer can always be spotted when present, then the increased number of guesses can easily offset the penalty of increased sub-band error rate, and the optimal number of sub-bands may be considerably greater than the four which we use throughout this article. It is interesting to note here that this trend has recently been followed in [20].

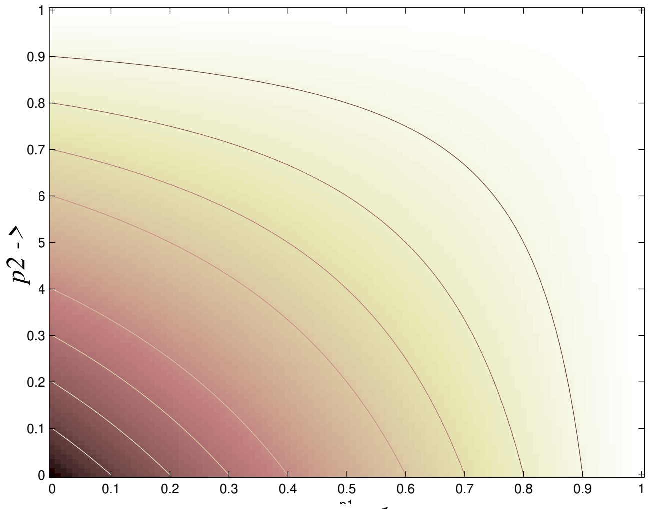


Figure 1. Probability of correct classification when classification error follows the product-of-errors rule, for two (independent) observation channels (white corresponding to maximum probability). Classifiers yield probabilities $p1 = \hat{P}(q_k|x_1)$ *and* $p2 = \hat{P}(q_k|x_2)$. $P(correct) = 1-(1-p1)(1-p2) = p1+p2-p1.p2$. *Contours show lines of equal correct recognition probability (contour height is value on contour-axis intercept).* $E[P(correct)]$ *is maximum when the expected values of* $p1$ *and* $p2$ *are maximum, and their covariance is zero.*

4. Multi-band processing in artificial speech recognition

While there are many possibilities in ASR for combining evidence from different data streams, such as vision with acoustics, or acoustic features from different time scales [15, 24, 47], in the present work we are primarily concerned with the combination of experts operating on different subdivisions of the acoustic frequency spectrum.

While the main motivation behind the multi-band approach is to exploit spectral data redundancy in a way which reflects the PoE rule for human speech perception, other potential advantages of the multi-band approach include:

WER	band 1	PoE WER	band 1	band 2	PoE WER	band 1	band 2	band 3	band 4	PoE WER
clean	7.1	7.1	16.1	32.3	5.20	33.1	29.7	38.9	55.1	2.11
SNR 12	15.6	15.6	22.6	46.7	10.55	39.6	41.1	50.7	70.8	9.58
SNR 0	50.0	50.0	54.4	81.6	44.39	70.7	75.4	80.2	88.9	38.01

Table 1: Sub-band word error rates for the Numbers95 database of connected digits are shown above, from left to right, for when the full frequency range is divided into 1, 2 and 4 sub-bands, respectively. Thick outlined boxes show the WER that would theoretically result if sub-band error rates combined according to the product-of-errors rule. In this case, although WER increases in each sub-band as the number of sub-bands is increased, the larger number of sub-bands would show considerable advantage. (See Section 7 for more details on sub-band frequency ranges and databases).

- **Channel specific processing**: Different recognition strategies might ultimately be applied in each sub-band. For example, higher frequencies could use greater time resolution, and lower frequencies greater frequency resolution. It would also be possible to use sub-band specific speech subunits.
- **Subunit specific expert combination**: It is sometimes possible to weight each expert according to the speech subunit which is being distinguished. Consonants, for example, might give more weight to high frequency sub-bands.
- **Channel asynchrony**: Models discussed here use the same phoneme set for each expert, and force synchrony between experts, but it would be possible to permit some level of sub-band asynchrony[1] [*5, 21, 28, 42*], (Figure 2), and to use speech subunits specific to each frequency subrange [28].

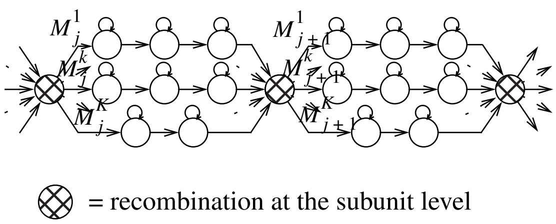


Figure 2. General form of a K-streams recogniser with anchor points between speech units (to force synchrony between different streams). Note that the model topology is not necessarily the same for the different subsystems.

The first multi-band ASR systems were based on the HMM/ANN (Hidden Markov Model/Artificial Neural Network) model [6, 7]. In standard full-band ASR an MLP (Multi-layer Perceptron) is first used to transform the acoustic data into posterior phoneme probabilities[2,3], $P(q_k|x^n;\Theta)$, for each word subunit, q_k, and data frame, x^n. The posterior

1. *When streams are not frame synchronous the complexity of the decoding algorithm required may be considerably greater than for a standard recogniser. Results to date have indicated that allowing asynchrony between streams does not give any significant performance improvement [28].*

probabilities from the MLP are then passed as scaled likelihoods [17] into an HMM for decoding. In early multi-band processing [5, 28, 35] the aim was to divide the frequency range into a number of sub-bands, x_i, and to process each of these independently, so one MLP was trained for each frequency sub-band, x_i. The posterior probabilities, $P(q_k|x_i^n;\Theta_i)$, were then combined, usually at the frame level, before decoding as for the full-band HMM/ANN. A number of different combination strategies have been proposed:

1. the linear weighted average:

$$P(q_k|x) \cong \sum_{i=1}^{d} w_i P(q_k|x_i;\Theta_i) \tag{2}$$

2. the geometric weighted average:

$$P(q_k|x) \cong \prod_{i=1}^{d} P(q_k|x_i;\Theta_i)^{w_i} \tag{3}$$

3. MLP combination.

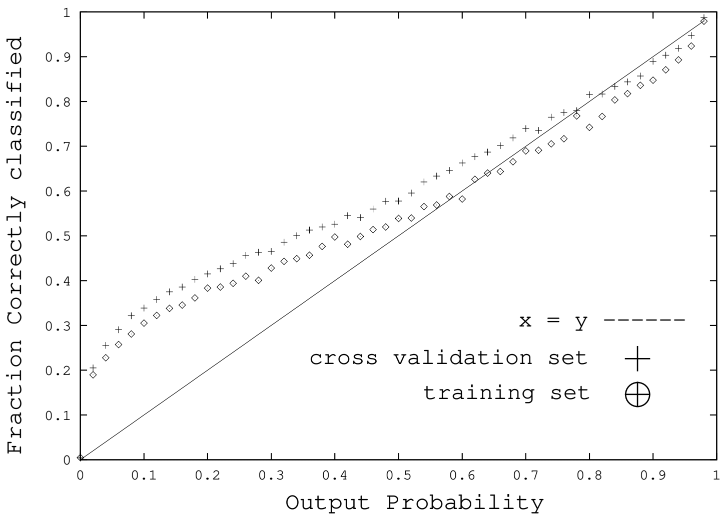


Figure 3. It is possible to generate "good" posterior probabilities from a neural network, and these are indeed good measures of the probability being correct. This plot was generated from real speech data by collecting statistics over the acoustic parameters from 1750 Resource Management speaker-independent training sentences and 500 cross-validation sentences (not used for training, but for which correct classification was known) [from 6].

---

2. *Any sufficiently flexible parametric function which is trained to minimise the sum of square errors against a target classification function (with one output per class) will generate output class posteriors which are optimal against the distribution of the training data [39]. As the number of hidden units increases from zero, MLP performance increases rapidly to the Bayesian optimum, when the posterior probability for data belonging to each class is equivalent to the probability that classification is correct if this class is selected [9]. Figure 3 shows that this equivalence is very close for a typical MLP used for phoneme classification in continuous speech.*

3. *See Nomenclature section for full definition of all mathematical symbols used.*

In this context $w_i \in [0, 1]$ is a fixed weight for each sub-band expert which should reflect the extent to which it contains features which are useful for phoneme discrimination. Best results were obtained using the more flexible of these functions, MLP combination [28].

This approach has a number of limitations. If it was required to adapt weights to changing noise conditions then supervised weight training could not be used. This problem could be overcome using methods similar to those described in Section 6. A more important limitation of this sub-band approach is that *independent sub-band processing looses all joint spectral information*, such as the shape of the spectral envelope. The PoE rule tells us that intelligent combination of a number of narrow band recognisers should be able to equal or outperform the full-band expert, but experimental results have repeatedly shown that independent sub-band processing leads to a significant decrease in performance with clean speech.

5. Multi-band ASR with latent variables

> *The first (rule of the method which leads to the truth) is to accept nothing as true which you do not clearly recognise to be so.* (Descartes, 1619)

It is now known that the PoE rule is only approximately true [40]. Even in human perception, where the error rate in each sub-band is far lower than with sub-band ASR (especially in noise), some use is made of joint sub-band information. One of the first methods which was attempted to overcome this problem was to combine the four sub-band experts with a full-band expert. This led to performance improvement towards, but not significantly surpassing, that of the full-band expert on its own [28]. In [21] tests were made with a seven sub-band system in which

> *an expert was used for each possible sub-band combination.*

Several combination strategies were tested. Simple linear combination showed satisfactory performance in clean speech, while hand selection of the correct expert (when present) showed the potential of this system for very strong robustness to band limited noise. This confirmed the following result, which had previously been demonstrated in the context of missing feature theory (MFT) [25, 32]: an effective strategy for reducing the effects of data mismatch is to detect and simply ignore strongly mismatching data.

> *At any point in time there are a number of sub-bands whose mismatch level is such that recognition will improve if data in these sub-bands is ignored.*

An obvious problem with the strategy of ignoring unreliable data is that estimating the local data mismatch level in each sub-band is not easy. At best we can assign a probability to each sub-band being "clean" (i.e. having mismatch below some threshold level) at time step n. In the case of d sub-bands there are 2^d possible sub-band combinations (if we include both the full and empty sets). Let s_i denote the i$^{\text{th}}$ combination of sub-bands from x. In estimating the full-band posterior $P(q_k|x)$ we would like to select the expert i which gives the "best" estimate for this phoneme. In the presence of band limited noise we can reasonably assume that the best combination will be the largest set of clean sub-bands. Define an indicator latent variable c_i to represent the event that s_i is the largest set of clean sub-bands. Providing events s_i includes the empty set, the set of events c_i are mutually exclusive and exhaustive. Therefore:

$$P(q_k|x) = P\left(q_k \cap \bigcup_i c_i \middle| x\right), \text{ because the set of events } c_i \text{ are exhaustive} \quad (4)$$

$$= \sum_i P(q_k \cap c_i | x), \text{ because } c_i \text{ are mutually exclusive} \quad (5)$$

$$= \sum_i P(c_i|x)P(q_k|c_i \cap x) \tag{6}$$

The condition "c_i true" tells us that x can be partitioned into certain and uncertain parts (s_i, s_i'). If nothing is known about the uncertain data, then this data is simply "missing", or "not given", so that:

$$P(q_k|c_i \cap x) = P(q_k|s_i) \tag{7}$$

Providing the number of combinations is not too large (d not greater than about 7), we can train an MLP expert on clean data from each combination[1] to output phoneme posterior estimates $P(q_k|s_i;\Theta_i)$ [2]. It is given that s_i in Eq. 7 is clean, so $P(q_k|s_i) \cong P(q_k|s_i;\Theta_i)$. Therefore:

$$P(q_k|x) \cong \sum_i P(c_i|x)P(q_k|s_i;\Theta_i) \tag{8}$$

Eq. 8 gives us a factorisation of the full-band posterior into a composition of combination weights and *clean* combination posteriors[3]. We call this the "full combination" (FC) multi-band approach. The utility of this method depends directly on the accuracy with which combination weights $P(c_i|x)$ can be estimated and adapted to changing noise conditions. Various methods for combination weight estimation are described in Section 6. The full-combination ASR system used in the experiments reported here were implemented in an HMM/ANN based system with four sub-bands. A two sub-band FCM system is illustrated in Figure 4.

Full-combination approximation. For d much greater than about 7 it becomes impractical to train a separate MLP for all 2^d sub-band combinations. One way to alleviate this problem may be to prune out combinations which have a-priori negligible weight for a particular application. Alternatively, although we must avoid the assumption $p(x_i|x_j) \cong p(x_i)$ of full sub-band independence, it is easy to show [33] that under the weaker assumption of *conditional* sub-band independence, $p(x_i|x_j \cap q_k) \cong p(x_i|q_k)$, all 2^d combination posteriors can be expressed in terms of the d single sub-band posteriors. If $|s_i|$ is the number of sub-bands in combination s_i, then the clean combination posteriors $P(q_k|s_i)$ in Eq. 8 can be approximated as follows:

$$P(q_k|s_i) \cong \bar{P}_{ki} / \sum_m \bar{P}_{mi}, \text{ where } \bar{P}_{ki} = P^{1-|s_i|}(q_k) \prod_{x_j \in s_i} P(q_k|x_j) \tag{9}$$

It is shown in Section 7 that the full-combination approximation based on Eq. 9 consistently outperforms the assumption of full independence that is implicit in the simple sub-band combination of Eq. 2.

1. *It is important to note that data should be orthogonalised within each sub-band [5, 35]. Orthogonalisation across the full data vector would usually spread noise between sub-bands, after which the noise would not be band limited and the advantage of sub-band processing would be lost. Alternative means of spectral data orthogonalisation have been proposed [34] which mix features across time rather than frequency, but this would also have a tendency to mix clean with noisy data.*

2. *We distinguish here between the true probabilities $P(q_k|s_i)$ and their estimates $P(q_k|s_i;\Theta_i)$, which are parametric functions (MLPs) trained on clean speech, and therefore can only be assumed to be accurate when it is known that the speech data in sub-band subset s_i is clean.*

3. *Likelihood based models (such as HMMs) can be decomposed in a similar way, but some complications can arise. More details on both posteriors and likelihood based decomposition is discussed in [31].*

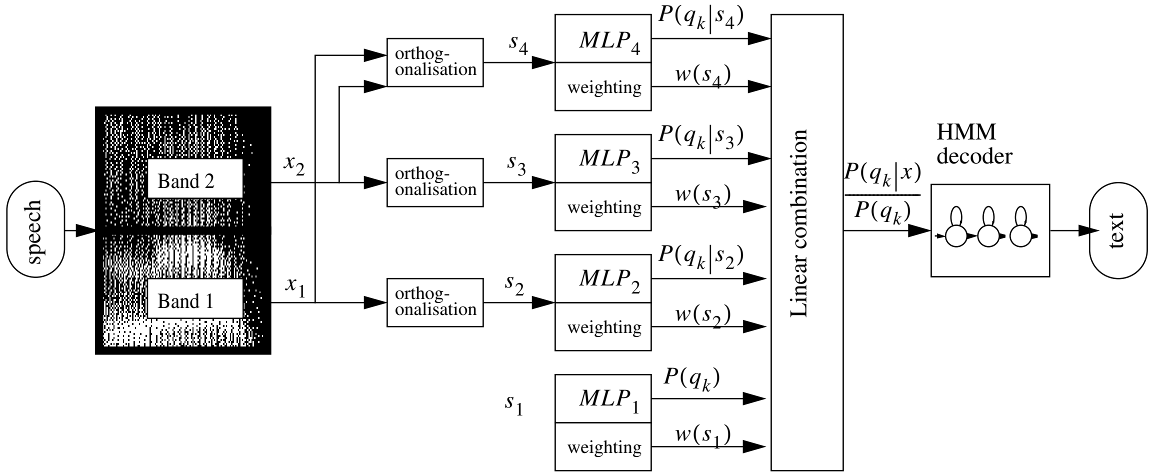


Figure 4. Full Combination multi-band ASR HMM/ANN system with 2 sub-bands. Each possible combination of frequency sub-bands is first orthogonalised and then input to an MLP which has been trained for this sub-band combination, and to a weighting function. Posterior phoneme probabilities from each MLP expert are then combined in a linear weighted sum to give a single posterior for each phoneme. These posteriors are then converted to scaled likelihoods before input to a fixed parameter HMM for Viterbi decoding.

6. Full-combination expert weighting

We describe here a number of approaches to both fixed and adaptive weighting.

If the latent variable c in Eq. 8 is not an indicator for each sub-band combination being the largest clean combination, then we are no longer justified in assuming that $P(q_k|s_i) \cong P(q_k|s_i;\Theta_i)$. On the other hand, if we can assume that the data is clean, and we are interested in exploiting the possibility that data in some sub-bands should carry more weight than data in others, then we can replace c by the latent variable b, which indicates whether each sub-band combination is “the best” or most useful. As b_i are also mutually exclusive and exhaustive, the same working used to derive Eq. 8 also gives us[1]:

$$P(q_k|x) \cong \sum_i P(b_i|x)P(q_k|s_i;\Theta_i) \tag{10}$$

Although the weights $P(b_i|x)$ in Eq. 10 depend on x, if we are interested in using fixed weights, then we must ignore x and assume that $P(b_i|x) \cong P(b_i)$.

6.1 Fixed weight estimation using linear & non-linear LMSE

Fixed weights $w_i = P(b_i)$, $i = 1 \ldots 2^d$, can be estimated using the (supervised) LMSE (least mean square error) criterion:

1. *It is also possible to combine the latent variables b_i and c_j by using a double summation [31].*

$$w_{LMSE} = \arg min_w \sum_{n=1}^{N} \sum_{k=1}^{2^d} (y_k^n(w) - t_k^n)^2 \tag{11}$$

where $y_k^n(w)$ is the combined posterior for q_k at frame n, given fixed weights w, and $t_k^n = P(q_k|x^n)$ is the target posterior, $= 1$ if target class for frame n is class q_k, else $= 0$. For *linear* LMSE:

$$y_k^n(w) = P(q_k|x^n;w, \Theta) = \sum_{i=1}^{2^d} w_i P(q_k|s_i^n;\Theta_i) \tag{12}$$

In this case the resulting LMSE “normal equations” are linear and can be solved directly, but in this case the weights cannot be constrained to be positive or sum to 1 across all experts. For non linear LMSE, when an MLP is trained using “back error propagation” (a particular case of gradient descent), weights can be constrained to sum to 1, if required, by using the softmax activation function in the output layer.

6.2 Fixed weight estimation using maximum likelihood

The fixed weights, $w_i = P(b_i)$, can be estimated using relative-frequencies[1] as follows:

$$w_i = P(b_i) \cong n_i / n \tag{13}$$

where n_i is the number of frames of training data for which expert i has the largest posterior, across all experts, for the target phoneme (and therefore has the smallest Kullback-Leibler distance from the target probability distribution), and n is the number of frames of training data.

6.3 Adaptive weighting using estimated local data mismatch

It is usually reasonable to assume that sub-band combination reliability can be estimated from the reliability of each of its component sub-bands, and that sub-band reliabilities are independent. The local data mismatch level in each sub-band can be estimated from measures $\gamma(x_j)$ of *likeness to speech data*, such as local SNR[2] [22, 33], harmonicity index [2] or, in the case of stereo data, interaural time delay [14]. By definition of c_i the adaptive weights $P(c_i|x)$ can then be obtained from these sub-band reliabilities as follows:

$$c_i \Leftrightarrow (reliable(x_j) \forall x_j \in s_i) \cap (\neg reliable(x_j) \forall x_j \notin s_i)$$

$$\Rightarrow P(c_i|x) = \prod_{x_j \in s_i} P(reliable(x_j)) \prod_{x_j \notin s_i} P(\neg reliable(x_j)) \tag{14}$$

If the measure $\gamma(x_j)$ is increasing with level of mismatch, then we can model $P(reliable(x_j))$ as $P(\gamma(x_j) < \varepsilon_j)$, where the threshold ε_j for each sub-band is estimated as the average level of $\gamma(x_j)$, over the training set, above which recognition improves when band x_j is ignored. If we assume that the estimated mismatch value $\hat{\gamma}(x_j)$ is its true value plus a zero mean Gaussian error, with variance σ_j^2, then we obtain:

1. *Relative-frequency is the ML (Maximum Likelihood) estimate for a Bernoulli probability.*
2. *SNR is not directly related to mismatch level and should be used with care. Low noise energy implies low data mismatch, but low noise energy combined with zero speech energy gives infinitely low SNR.*

$$P(reliable(x_j)) \cong P(\gamma(x_j) < \varepsilon_j) = \Phi\left(\frac{\varepsilon_j - \hat{\gamma}(x_j)}{\sigma_j}\right) \tag{15}$$

where σ_j^2 are parameters which reflect our confidence in the estimation precision for each sub-band, which could be tuned, for example, to maximise performance in clean speech.

7. Experimentation

Databases. Speech was taken from the Numbers95 database of multi-speaker US English free-format numbers[1] telephone speech, with 30 words and 33 phonemes [8]. Factory noise from the Noisex92 database [43], and car noise from an in-house database, were added at varying SNR levels relative to the average signal energy in each utterance (excluding non-speech periods). Multi-band ASR is best suited to band limited noise, but the noises used in these initial tests were not band limited.

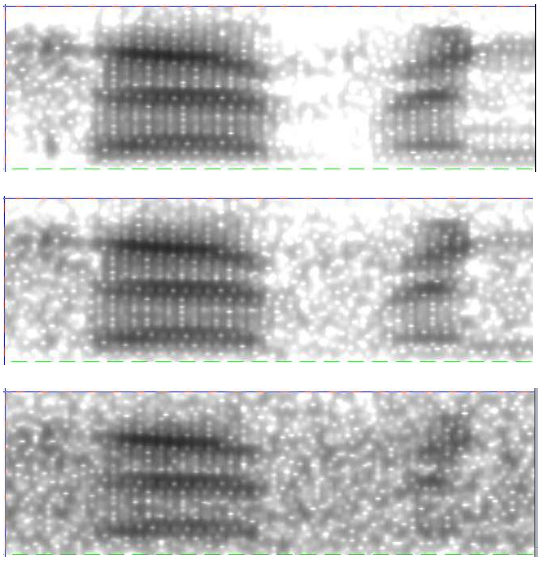


Figure 6. Signal "seven" from Numbers95 (top) mixed with Daimler-Chrysler car noise at SNR = 12 (middle) & 0 dB (bottom).

Acoustic features. Tests were made using PLP coefficients [18, 30, 37], with an analysis window size of 25 ms, and 12.5 ms shift. Tests were also made using J-Rasta-PLP features [19, 30], which result from PLP features after an adaptive cancellation of slowly changing additive and convolutive noises. In order to compete with the state-of-the-art baseline system used here, data features had to be near orthogonal. Orthogonalisation was applied within each data stream, so as not to spread noise between streams.

Recognition systems. Baseline tests were made with a full-band HMM/ANN hybrid [7] in which the ANN was an MLP with one hidden layer of 1000 units, trained to input 9 consecutive data frames and output the posterior probabilities $P(q_k|x)$ for each of 33 phonemes q_k and each data frame, x. Training used the full Numbers95 training set [8]. In recognition, posterior probabilities from the MLP, divided by their priors[2], were passed as scaled

1. *e.g. "two hundred eleven".*

likelihoods to a fixed parameter HMM for decoding. The HMM for each phoneme used a 1 to 3 repeated-state model. No language model was used.

For the full-combination multi-band system a separate MLP (of identical design to the full-band MLP) was trained on clean data for each sub-band combination. Multiple MLP outputs were then merged at the frame level (which here was also the state level), using Eq. 8, to give a single posterior probability for each class, before these were passed as scaled likelihoods to the same HMM as used by the full-band system

Sub-band definition. The number of sub-bands used in sub-band ASR, and their frequency ranges, are somewhat arbitrary. For these experiments we have chosen to work with four sub-bands. The original reason for this choice was based on having one sub-band for each formant. Frequency ranges were chosen accordingly and are displayed in Table 2. Although it would be instructive to test with different sub-band configurations, we kept the sub-band specification constant so that test results could be compared over a number of different experimental configurations.

Sub-bands Def.	band 1	band 2	band 3	band 4
frequency range /Hz	216-778	707-1632	1506-2709	2122-3769
# coeffs, lpc order	5, 3	5, 3	3, 2	3, 2

Table 2: Sub-bands definition

Recognition tests. Recognition tests were made with the two different noise types as described, at SNR levels clean (45 dB SNR), 12 dB and 0 dB SNR. Tests used the full Numbers95 development test set. Results are summarised in Figure 7 for PLP features, and Figure 8 for J-Rasta-PLP features.

Top figures show results using different weighting measures, car noise (left) and factory noise (right), for:

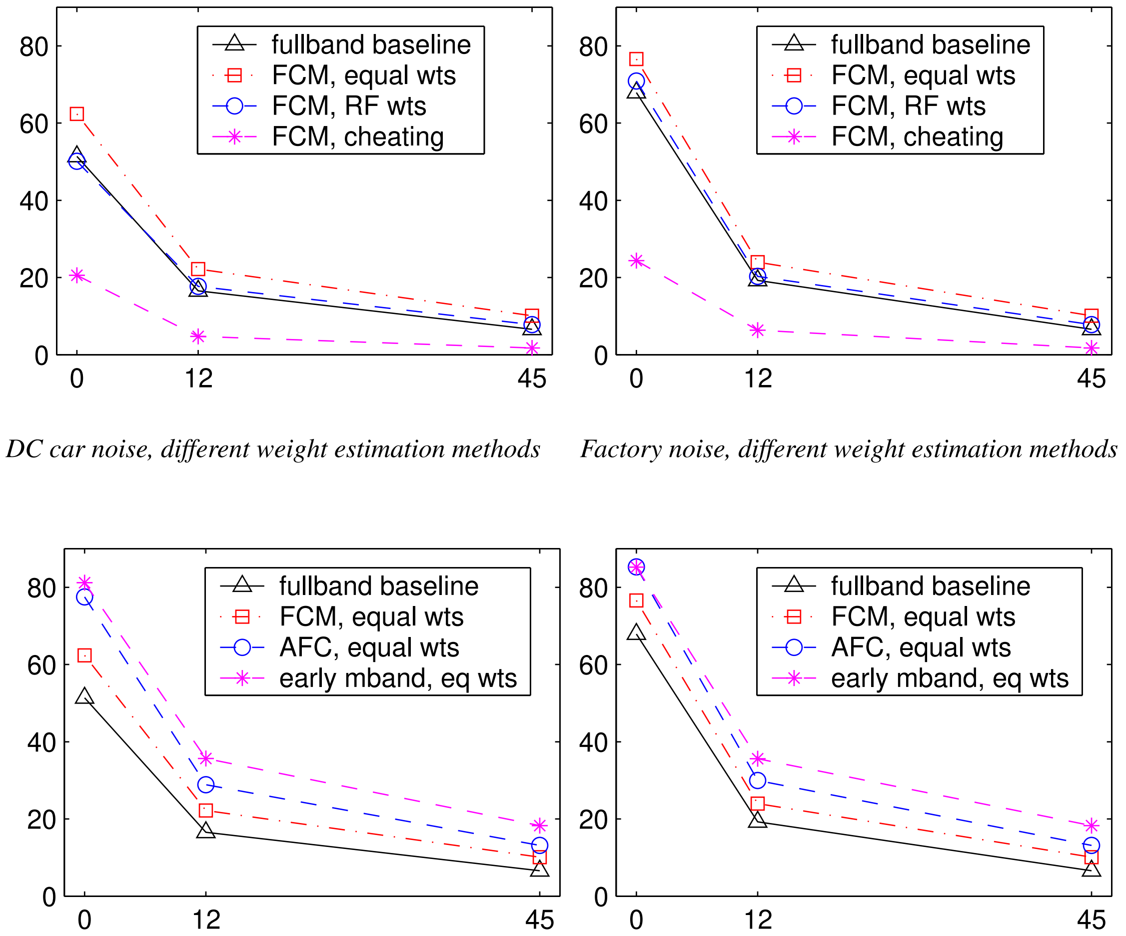
1. full-band baseline HMM/ANN hybrid ASR system
2. full combination multi-band hybrid, with 4 sub-bands, using equal weights, “FCM, equal wts”
3. FCM using relative-frequency, “FCM, RF wts”
4. FCM with cheating, “FCM, cheating”, i.e. correct phoneme selected whenever selected by any expert.

Bottom figures show results using different multi-band models, all using equal weights, for the following:

1. full-band baseline HMM/ANN hybrid ASR system
2. full combination multi-band hybrid, using equal weights, “FCM, equal wts”
3. full combination approximation, using Eq. 9, using equal weights, “AFC, equal wts”
4. early multi-band approach (one expert per sub-band), with equal weights, “early mband, eq wts”

---

2. *Although we should theoretically divide posteriors by priors to get the (scaled) likelihoods which are required by the HMM for decoding, in practice the division by estimated priors is only helpful if these estimates accurately represent the priors in the test data. Inaccurate priors can result if some of the phonemes in the training set were undersampled, or if the true priors in the training set do not match the priors in the test set. Better results were obtained in the experiments reported here when the posteriors were not divided by the priors. This is equivalent to assuming that all priors are equal.*

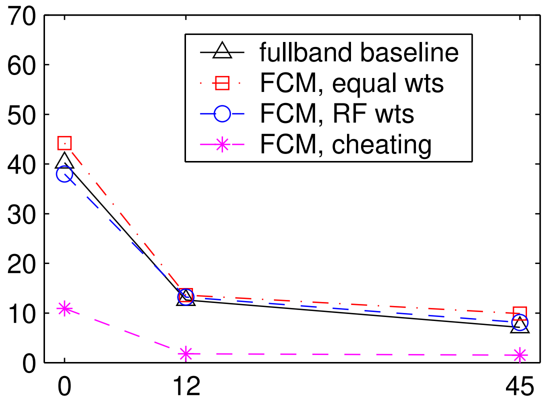


DC car noise, different weight estimation methods *Factory noise, different weight estimation methods*

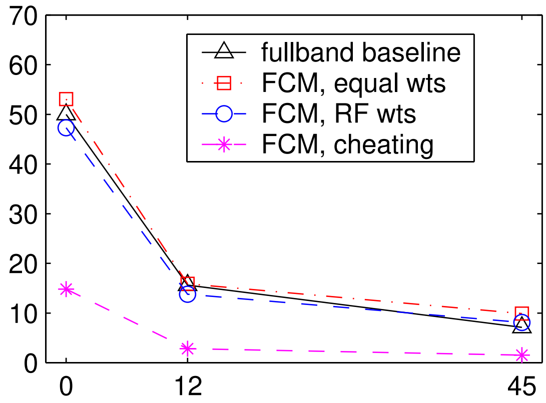
DC car noise, different multi-band ASR models *Factory noise, different multi-band ASR models*

Figure 7. WER (vertical axis) against SNR with PLP features

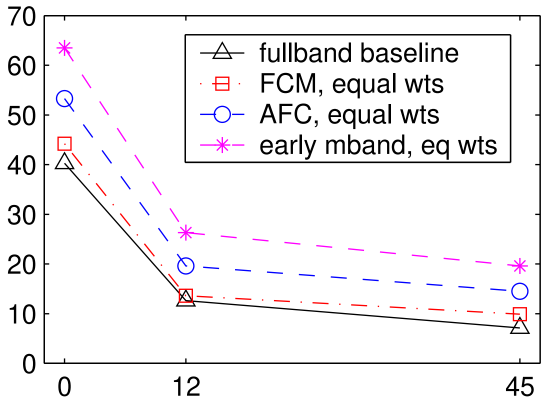
Results. As usual, J-Rasta-PLP features show a strong improvement over PLP features (which give similar performance to MFCCs). Note that the benefits of robust preprocessing are largely complementary to those offered by multiband processing. Of the two noise types tested, the more impulsive factory noise has a worse effect on performance than car noise, which is near stationary. The “cheating” performance, which would obtain if the PoE rule was in effect (i.e. if there was correct phoneme recognition whenever any sub-band combination expert had correct recognition), is very robust down to about 12 dB SNR. None of the methods tested has performance anywhere near as good as cheating, and only one (FCM with fixed RF weights and Rasta features) has performance which improves over the baseline in noise. Equal weights perform almost as well as FC weights, especially with Rasta features. The failure of any of the multi-band methods tested to significantly improve over the full-band baseline may be because both noise types also affected every sub-band, while the advantage of multiband processing is strongest when one or



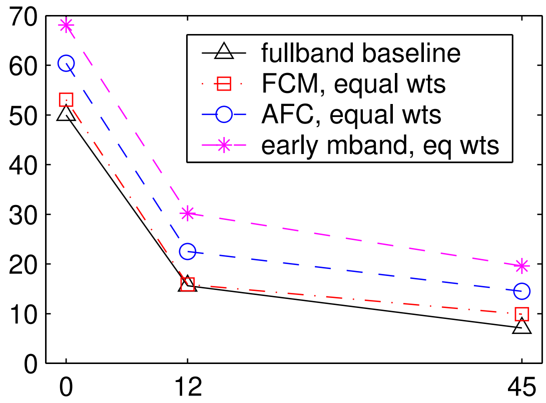
DC car noise, different weight estimation methods



Factory noise, different weight estimation methods



DC car noise, different multi-band ASR models



Factory noise, different multi-band ASR models

Figure 8. WER (vertical axis) against SNR with J-Rasta-PLP features

more sub-bands are often clean. However, the full-combination method consistently outperforms the full-combination approximation method (using just one expert per sub-band), which in turn significantly outperforms the "early sub-band" approach (which also used one expert per sub-band, but combines them in a less principled way).

8. Summary and conclusion

After an introduction to the general advantages of multi-stream processing, and to the evidence for its use in human perception, we recounted how Fletcher's product-of-errors rule for independent sub-band processing in speech perception, as well as a number of other factors, motivated the development of multi-band processing for the purpose of robustness to band-limited noise in ASR. We then argued that the original formulation of multi-band ASR was disadvantaged by independent sub-band processing, and described how this problem could be theoretically overcome by performing separate recognition on every possible combination of sub-bands, which enables us to integrate over all possible positions of noisy or mismatching data. A derivation of this “full combination” multi-bands approach was presented in the context of HMM/ANN based ASR in which an MLP outputs phoneme posterior probabilities for each time frame and for each sub-band combination. An indicator latent variable was used to specify whether each sub-band combination is the largest set of clean sub-bands. In this way the full-band phoneme posterior for each phoneme was decomposed into a reliability weighted sum of corresponding clean sub-band-combination posteriors. While theoretically attractive, especially in the context of time varying and band limited noise, the effectiveness of this approach depends strongly on the method used for reliability weighting. A number of weighting methods were presented. Some of these methods are only suitable for estimating static weights, while others can be used in the more general context where mismatch level varies over time and frequency.

The full-combination technique was tested on a speaker independent continuous speech free-format numbers recognition task, for clean speech and for speech plus car or factory noise. Tests compared the full-combination technique with the early sub-band approach, and also with the full-combination approximation. Results showed that, in the case of four sub-bands:

- both the full-combination method and its approximation outperform the early sub-band approach, even when equal weights are used for each combination expert.
- relative frequency based FCM weighting marginally outperforms the full-band baseline in noise when using J-Rasta-PLP features.

These results, for static weighting, have not shown a clear advantage for multi-band ASR over the full-band baseline. However, the main advantage of adaptive weighting in multi-band ASR is with mismatch which is band limited and/or changing, while the initial experiments which we report here use only noises which are not band limited and are near constant. Tests using a simple form of SNR based adaptive weighting [22] were made, but the results were not shown, because they were not significantly better than for equal weights.

The full advantage of the full-combination approach will not come into play unless mismatch is unequally distributed over time and/or frequency, with some proportion of data remaining clean. It is therefore imperative that effective noise robust preprocessing is used to remove as much constant or slowly changing noise as possible prior to multi-band processing. The FCM results reported here do not much improve over the baseline system, but they do show some increased advantage when noise robust J-Rasta-PLP features are used in place of simple PLP features, even when the noise is not band limited. It is also clear that expert weighting (mismatch estimation) should be as accurate and rapidly adapting as possible

Future directions. The static weighting in Section 6 is clearly of little use in the presence of noise. The techniques for adaptive mismatch estimation outlined in Section 6.3 assume that training data is “speech like”, and mismatching data is non speech like. However, mismatch concerns only how field data differs from the data population used in training, and not how it differs from “clean speech”. As the training population is defined by its pdf or “likelihood function”, it can be argued that mismatch estimation should be based, directly or indirectly, on use of the training data pdf. One such approach would be to use mismatch detection based on the assumption that clean and mismatching data have different (distributions of) likelihood values. Another would be to use weights which maximise the local clean-data likelihood (although this method could be applied only in the context of likelihood based FCM decomposition). We have developed both of these approaches in more detail in [31], although they have not yet been fully tested. Other approaches making more or less direct use of the training data pdf to downweight mismatching sub-bands are described in [27, 45].

Finally, it should be noted that although the “early” multiband approach tested here shows significantly reduced performance in relation to the full-combination approach, a lot of variants of independent sub-band processing have reported positive performance results [5, 24, 28]. With the present full-combination approach we are theoretically limited to linear expert combination. However, some of the best results for independent sub-band processing used MLP combination, which is highly non-linear. There are also a number of other combination strategies which worked well for independent sub-band processing which we have not yet tested with FCM, such as weighting by some function of the entropy in the posteriors output from each expert [2, 21, 28].

Acknowledgements

This work was supported by the Swiss Federal Office for Education and Science (OFES) in the framework of both the EC/OFES SPHEAR (SPeech, HEAring and Recognition) project and the EC/OFES RESPITE project (REcognition of Speech by Partial Information TEchniques).

Nomenclature and commonly used abbreviations

ASR	Automatic Speech Recognition
ANN	Artificial Neural Network
MLP	Multi-layer perceptron
HMM	Hidden Markov Model
ML	Maximum Likelihood
SNR	Signal to Noise Ratio
WER	Word Error Rate
FCM	Full-Combination Multiband model
AFC	Approximate Full Combination multiband model
$P(x)$	probability of “event x” occurring
$p(x)$	probability density at x of a continuous value
$P(x;\Theta)$	function with parameters Θ used to estimate $P(x)$
$\hat{P}(x)$	function (parameters unspecified) used to estimate $P(x)$
$P(a \cap b)$	probability of events ‘a’ and ‘b’ occurring (same as $P(a, b)$)
$P(a \cup b)$	probability of events ‘a’ or ‘b’ occurring
$\neg a$	not a , the negation of condition a
q_k	speech unit whose presence is being estimated, or event that data x is from this class
x , x^n	data window vector at time step n
d	number of spectral sub-bands
x_i	i[th] sub-band of x , for $i = 1 \dots d$
s_i, s_i'	i[th] sub-band combination of x , and its complement, for $i = 1 \dots 2^d$
x clean	x is from data population used in model training (i.e. x has no data mismatch)
b_i	s_i is best subset of x for estimating speech unit posteriors
c_i	s_i is largest clean subset of x
b_{ik}	s_i is best subset of x for estimating speech unit posteriors when true unit is q_k
w	vector of sub-band expert weights
$\Phi(x)$	cumulative density function for the standard Gaussian pdf

References

[1] Allen, J. B. (1994) “How do humans process and recognise speech?”, IEEE Trans. on Speech and Signal Processing, Vol.2, No.4, pp.567-576.

[2] Berthommier, F. & Glotin, H. (1999) “A new SNR-feature mapping for robust multi-stream speech recognition”, Proc. ICPhS’99, pp.711-715.

[3] Bishop, C. (1995) Neural Networks for Pattern Recognition, Clarendon Press, Oxford, pp.365-368.

[4] Bourlard, "Non-stationary multi-channel (multi-stream) processing towards robust and adaptive ASR", Proc. Tampere Workshop on Robust Methods for Speech Recognition in Adverse Conditions, pp. 1-10.

[5] Bourlard, H. & Dupont, S. (1996) “A new ASR approach based on independent processing and recombination of partial frequency bands”, Proc. ICSLP’96, Philadelphia, pp. 422-425.

[6] Bourlard, H. & Morgan, N. (1994) Connectionist speech recognition - a hybrid approach, Kluwer Academic Publishers.

[7] Bourlard, H. & Morgan, N. (1997) “Hybrid HMM/ANN Systems for Speech Recognition: Overview and New Research Directions”, Proc. “International School on Neural Nets: Adaptive Processing of Temporal Information”

[8] Cole, R. A., Noel, T., Lander, L. & Durham, T. (1995) “New telephone speech corpora at CSLU”, Proc. European Conf. on Sp. Comm. and Tech., 1, pp. 821-824.

[9] Duda, R. O. & Hart, P. E. (1993) Pattern classification and scene analysis, John Wiley.

[10] Dupont, S. & Luettin, J. (1998) “Using the multi-stream approach for continuous audio-visual speech recognition: experiments on the M2VTS database”, Proc. ICSLP’98, pp. 1283-1286.

[11] Fletcher, H. (1922) “The nature of speech and its interpretation”, J. Franklin Inst., 193(6), pp.729-747.

[12] Gales, M. J. F. & Young, S. J. (1993) “HMM recognition in noise using parallel model combination”, Proc. Eurospeecl’93, pp.837-840.

[13] Girin, L., Feng, G. & Schwartz, J.-L. (1998) “Fusion of auditory and visual information for noisy speech enhancement: a preliminary study of vowel transitions”, Proc. ICASSP’98, pp.1005-1008.

[14] Glotin, H., Berthommier, F. & Tessier, E. (1999) “A CASA-labelling model using the localisation cue for robust cocktail-party speech recognition”, Proc. Eurospeech’99, pp.2351-2354.

[15] Greenberg, S. (1997) “On the origins of speech intelligibility in the real world”, Proc. ESCA workshop on robust speech recognition for unknown communication channels, pp.23-32.

[16] Hagen, A., Morris, A. C. & Bourlard, H. (1999) “Different weighting schemes in the full combination sub-bands approach for noise robust ASR”, Proc. Tampere Workshop on Robust Methods for Speech Recognition in Adverse Conditions, pp. 199-202.

[17] Hennebert, J., Ris, C., Bourlard, H., Renals, S. and Morgan, N. (1997) “Estimation of global posteriors and forward-backward training of hybrid systems”, Proc Eurospeech’97, pp.1951-1954.

[18] Hermansky, H. (1990) “Perceptual linear predictive (PLP) analysis of speech”, J. Acoust. Soc. Am., 87(4), pp.1738-1752.

[19] Hermansky, H. & Morgan, N. (1994) “RASTA processing of speech”, IEEE Trans. on Speech and Audio Processing, 2(4), pp.578-589.

[20] Hermansky, H. & Sharma, S. (1999) “Temporal patterns (TRAPS) in ASR noisy speech”, Proc ICASSP’99, pp. 298-292.

[21] Hermansky, H., Tibrewela, S. & Pavel, M. (1996) “Towards ASR on partially corrupted speech”, Proc ICSLP’96, pp. 462-465.

[22] Hirsch, H. G. and C. Ehrlicher (1995) "Noise estimation techniques for robust speech recognition", ICASSP95 , 1995, pp. 153-156.

[23] Jordan, M. I., & Jacobs, R. A. (1994) "Hierarchical mixtures of experts and the EM algorithm", Neural Computation, 6, pp.181--214.

[24] Kingsbury, B. Morgan, N. & Greenberg, S. (1998) "Robust speech recognition using the modulation spectrogram", Speech Communication, Vol.25, Nos.1-3, pp.117-132.

[25] Lippmann, R. P. & Carlson, B. A. (1997) "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise", Proc. Eurospeech'97, pp. 37-40

[26] McGurk, H. & McDonald, J. (1976) "Hearing lips and seeing voices", Nature, No.264, pp.746-748.

[27] Ming, J. & Smith, F. J. (1999) "Union: a new approach for combining sub-band observations for noisy speech recognition", Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions, pp. 175-178.

[28] Mirghafori, N., (1999) "A multi-band approach to automatic speech recognition", PhD Dissertation, University of California at Berkeley, Dec 1998. Reprinted as ICSI Technical Report, ICSI TR-99-04.

[29] Moore, B. C. J. (1997) An introduction to the psychology of hearing (4th edition), Acedemic Press.

[30] Morgan, N., Bourlard, H. & Hermansky, H. (1998) "Automatic speech recognition: an auditory perspective", Research Report IDIAP-RR 98-17.

[31] Morris, A. C. (1999) "Latent variable decomposition for posteriors or likelihood based sub-band ASR", Research Report IDIAP-Com 99-04.

[32] Morris, A. C., Cooke, M. & Green, P. (1998) "Some solutions to the missing feature problem in data classification, with application to noise robust ASR", Proc. ICASSP'98, pp.737-740.

[33] Morris, A. C., Hagen, A. & Bourlard, H. (1999) "The full-combination sub-bands approach to noise robust HMM/ANN based ASR", Proc. Eurospeech'99, pp. 599-602.

[34] Nadeu, C., Hernando, J. & Gorricho, M. (1995) "On the decorrelation of filterbank energies in speech recognition", Proc. Eurospeech'95, pp.1381-1384.

[35] Okawa, S. Boccieri, E. & Potamianos, A. (1998) "Multi-band speech recognition in noisy environment", Proc. ICASSP'98, pp. 641-644.

[36] Pickles, J. O. (1988) An introduction to the physiology of hearing, Academic Press.

[37] Rao, S. & Pearlman, W.A. (1996) "Analysis of linear prediction, coding and spectral estimation from sub-bands", IEEE Trans. on Information Theory, Vol.42, pp.1160-1178.

[38] Raviv, Y., Intrator, N. (1996) "Bootstraping with noise: an effective regularisation technique", Connection Science, Special issue on Combining Estimators (8), pp.356-372.

[39] Richard, M. D. & Lippmann, R. P. (1991) "Neural network classifiers estimate Bayesian a-posteriori probabilities", J. Neural Computation, 3(4) MIT Press, pp.461-483.

[40] Steeneken, H. J. M. & Houtgast, T. (1999) "Mutual dependence of the octave-band weights in predicting speech intelligibility", Speech Communication, Vol.28, No.2, pp.109-123.

[41] Tomlinson, J., Russel, M. J. & Brooke, N. M. (1996) "Integrating audio and visual information to provide highly robust speech recognition", Proc. ICASSP'96, pp. 821-824.

[42] Tomlinson, J., Russel, M. J., Moore, R. K., Bucklan, A. P. & Fawley, M. A.(1997) "Modelling asynchrony in speech using elementary single-signal decomposition", Proc. ICASSP'97, pp.1247-1250.

[43] Varga, A., Steeneken, H. J. M., Tomlinson, M. & Jones, D. (1992) "The Noisex-92 study on the effect of additive noise on automatic speech recognition", Tech. Rep. DRA Speech Research Unit.

[44] Varga, A. & Moore, R. (1990) “Hidden markov model decomposition of speech and noise”, Proc. ICASSP’90, pp.845-848.

[45] de Veth, J., de Wet, F., Cranen, B. & Boves, L. (1999) “Missing feature theory in ASR: make sure you missing the right type of features”, Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions, pp.231-234.

[46] Westphal, M. & Waibel, A. (1999) “Towards spontaneous speech recognition for on-board car navigation and information systems”, Proc. Eurospeech’99, pp.1955-1958.

[47] Wu, S.-L., Kingsbury, B. E. Morgan, N. & Greenberg, S. (1998) “Performance improvements through combining phone and syllable scale information in automatic speech recognition”, Proc. ICASSP’98, pp.459-462.