IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Adaptive ML-Weighting in Multi-Band Recombination of Gaussian Mixture ASR

Astrid Hagen [§] Hervé Bourlard [§]
Andrew Morris [§]

IDIAP–RR 01-05

February 2001

TO APPEAR IN
Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2001)

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
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

§ IDIAP—Dalle Molle Institute of Perceptual Artificial Intelligence, P.O. Box 592, CH-1920 Martigny, Switzerland, {hagen,bourlard,morris}@idiap.ch.

IDIAP Research Report 01-05

Adaptive ML-Weighting in Multi-Band Recombination of Gaussian Mixture ASR

Astrid Hagen Hervé Bourlard Andrew Morris

February 2001

to appear in
Proceedings of the International Conference on Acoustics, Speech and Signal Processing
(ICASSP 2001)

Abstract. Multi-band speech recognition is powerful in band-limited noise, when the recognizer of the noisy band, which is less reliable, can be given less weight in the recombination process. An accurate decision on which bands can be considered as reliable and which bands are less reliable due to corruption by noise is usually hard to take. In this article, we investigate a maximum-likelihood (ML) approach to adapting the combination weights of a multi-band system. The Gaussian Mixture Model parameters are kept constant, while the combination weights are iteratively updated to maximize the data likelihood. Unsupervised offline and online weights adaptation are compared to use of equal weights, and 'cheating' weights where the noisy band is known, as well as to the fullband system. Initial tests show that both ML-weighting strategies show a robustness gain on band-limited 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).

1 INTRODUCTION

In multi-band (MB) processing, the speech signal in the spectral domain is split into several (non-overlapping) frequency subbands. These frequency subbands are processed separately for feature extraction, orthogonalization (e.g. DCT) and, in our case, frame level phoneme probabilities estimation. The estimated probabilities from each subband recognizer are then recombined by a combination rule, such as the weighted sum or product.

The strength of MB systems lies in the fact that possibly occurring noise from one subband does not mix with neighbouring subbands, as is usually the case. In fullband processing, feature extraction and orthogonalization are both carried out only once for the whole frequency domain which results in a feature vector in which noise in any one subband is spread over all features.

Experiments in recognition with missing data [2] have shown that it is possible to significantly improve the recognition task when noisy feature coefficients are ignored. A similar finding, but this time based on human auditory processing, was obtained by Fletcher in 1953 [4]. He showed that humans are often able to extract sufficient residual information from clean frequency subbands when a considerable part of the frequency domain is corrupted by noise. MB processing permits us to process each frequency subband separately, and thus to discard noisy subbands – if they can be detected.

The extent to which a subband should be included in the recognition task is controlled by the weighting factors used in the recombination process. Each subband is normally given a certain weight according to its estimated reliability. These weights can be calculated in advance (i.e. in an offline manner) and/or during recognition (i.e. online).

Different weighting schemes, such as mean square error, mutual information [3] or signal-to-noise ratio (SNR) based weights [7] can already be found in the literature. In this article, we present a new maximum-likelihood (ML) based weighting strategy for MB processing, which can be used for unsupervised online adaptation. For this, in the framework of Gaussian Mixture Expert/HMM systems [5], the parameters of the Gaussian Mixture Model (GMM) experts from each subband combination are fixed. Only the combination weights for each subband model and recognition unit, i.e. in our case phonemes, are iteratively updated.

In Section 2, the (online and offline) ML-weights adaptation for a GMM-based MB system is presented. Experiments employing this new weighting function in clean and noise-corrupted data are discussed in Section 3.

2 UNSUPERVISED ML ADAPTATION

Given an acoustic vector x^t at time t and the whole set of model parameters Θ, b_j tells us the expert j for which x_j is clean[1]. We now decompose the state probability $p(x^t|q_k,\Theta)$ into a weighted sum of subband GMM distributions, summing over all $b_j, j = 1..J$, with $J = 2^d$ where d is the number of subbands:

$$p(x^t|q_k,\Theta) = \sum_{j=1}^{J} p(x^t|b_j,q_k,\Theta)P(b_j|q_k,\Theta) \tag{1}$$

The parameters in (1) are the parameters Θ^g of the GMMs (means, variances and mixture weights) and the combination weights for each subband model and each state, denoted by $w_{jk} = P(b_j|q_k,\Theta)$. The whole set of parameters is thus $\Theta = \{\Theta^g, \mathbf{w}\}$.

In the following, we will consider the possibility of fast adaptation of weights $\mathbf{w}$, while keeping all other parameters Θ^g fixed. The idea behind this is that all the subband classifiers have been optimized on clean speech. Thus, when narrow band noise is present the classifiers of the noise-contaminated bands should be downweighted to achieve optimal performance. The limited number of parameters

[1] Time index t is dropped for $x_j^t = x_j$ for sake of clarity.

to be adapted here theoretically allows for fast adaptation. The number of parameters to be adapted correspond roughly to 1% of the total number of fixed model parameters.

A separate GMM expert is trained on clean data for every combination x_j, $j = 1..J$, of data subbands. The probabilities from these experts are combined in the same way that each Gaussian component is usually combined into a GMM.

$$p(x|q_k,\Theta) = \sum_j p(x|m_j,q_k,\Theta)P(m_j|q_k) \tag{2}$$

The adaptive expert weights are formed by combining local weights estimated online during recognition with global weights estimated offline during training (or equal weights).

2.1 Offline ML Expert Weights Estimation

In the case of fixed mixture component parameters, the usual iterative EM estimation equations [1] for mixture weights $w_{jk}^{(m+1)} = P^{(m+1)}(b_j|q_k)$ at iteration $(m+1)$ are as follows:

$$w_{jk}^{(m+1)} = \frac{1}{T}\frac{1}{P(q_k)}\sum_{t=1}^{T} P^{(m)}(b_j,q_k|x^t,\Theta^g,w^{(m)}) \tag{3}$$

$$P^{(m)}(b_j,q_k|x^t) = \frac{p^{(m)}(x^t|b_j,q_k,\Theta^g)P^{(m)}(b_j,q_k)}{\sum_{j'k'} p^{(m)}(x^t|b_{j'},q_{k'},\Theta^g)P^{(m)}(b_{j'},q_{k'})} \tag{4}$$

The only difference here being that the mixture weights are now fixed, and b_j in place of the usual mixture index m_j tells us the expert j for which x_j is clean, and its complement $\overline{x_j}$ is noisy and should be ignored. This means we can factorize the probability $p(x|b_j,q_k)$ into reliable and unreliable parts as follows

$$p(x|b_j,q_k) = p(x_j|b_j,q_k,\Theta)p(\overline{x_j}|x_j,b_j,q_k) \tag{5}$$

The unreliable factor $p(\overline{x_j}|x_j,b_j,q_k)$ in (5) can be approximated [2] with minimum variance by its expected value

$$\alpha_{jk} = \mathcal{E}[p(\overline{x_j}|x_j,b_j,q_k,\Theta)] \tag{6}$$

In the initial experiments reported here we have made two simplifying assumptions. One is that α_{jk} is a constant independent of j and k, and therefore cancels out when (5) and (6) are substituted into (4). Another is that in place of all 2^J experts, we use just one expert per subband, plus one more for the fullband data.

2.2 Online ML Expert Weights Adaptation

We now only consider one iteration per time step and thus drop the *iteration* index m, introducing a *time* index n. In online adaptation, $N \ll T$ frames are used to obtain a local estimate $w_N^{(n)} = P^{(n)}(b_j|q_k)$ for the weights, which is then combined with the previous estimate $w^{(n-1)} = P^{(n-1)}(b_j|q_k)$ from the former time step $(n-1)$ in a weighted sum as follows:

$$\begin{aligned} w^{(0)} &= \text{offline weights} \\ w^{(n)} &= (1-\alpha)\ w^{(n-1)} + \alpha\ w_N^{(n)} \end{aligned} \tag{7}$$

with $w_N^{(n)} = \frac{1}{N}\frac{1}{P(q_k)}\sum_{n_i=n}^{n+N-1} P^{(n)}(b_j, q_k|x_{n_i})$ (cf. (3)).

The α-value determines how fast the weights change from one update to the next: $\alpha = 0$ actually results in no adaptation as the new weights are not taken into consideration, while $\alpha = 1$ is the other extreme when only the new local weights are used in the next update.

3 EXPERIMENTS

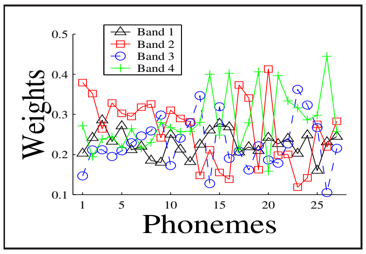

Figure 1: *Illustration of offline adapted weights for clean speech (MFCC features).*

Experiments were carried out on a test set of 100 utterances from the Numbers95 database of connected numbers recorded over the telephone line. For tests on noise-corrupted data, artificial band-limited (stationary) noise at SNRs of 12 and 0 dB was added to each frequency subband at a time, although due to filter characteristics there is a slight noise leakage between bands.

Our MB system comprises 4 subbands. Two sets of features were chosen: PLP (Perceptual Linear Prediction) and MFCC features. GMM classifiers were (unsupervised) trained on each set of features and for each frequency subband (as well as the fullband). The MB system (recombination by weighted sum and product) is tested in the different noise conditions employing the offline (3) and (4), and online (7) adaptive ML-weights. Results are compared to the same set-up using equal weights and 'cheating' weights, which were set to zero for the noisy subband and equal for the clean bands, as well as to the fullband GMM classifiers, which were trained on the whole frequency domain. The offline weights were calculated on a different set of 100 utterances, corrupted with the respective noise. Online weights were updated every $N = 100$ frames (1250 ms).

3.1 Offline ML Expert Weights Adaptation

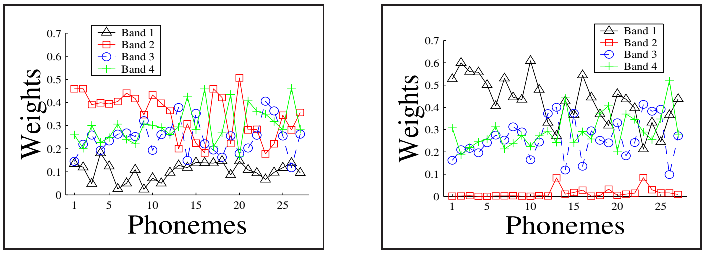

Figure 2: *Illustration of offline adapted weights for noise in* ***subband 1*** *(left) and subband 2 (right) (MFCC features).*

In a multi-stream system using 4 subbands as input, we would expect the new ML-weights to show a clear advantage over equal weights when one of the bands is totally corrupted by noise. Therefore,

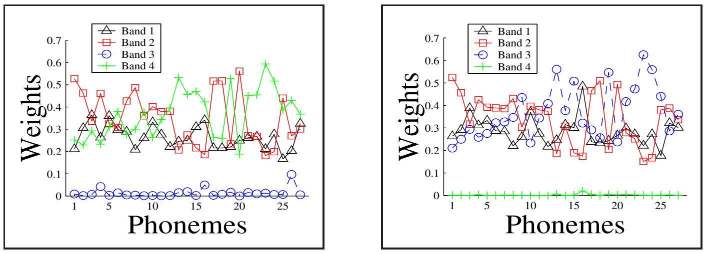

Figure 3: *Illustration of offline adapted weights for noise in subband 3 (left) and subband 4 (right) (MFCC features).*

noise **SNR 12 dB**	fullband WER	MB system w/ **sum rule** weights	WER	cheat
clean	13.5	equal	17.0	
		offline	15.2	-
band 1	64.4	equal	42.0	
		offline	42.5	34.9
band 2	23.8	equal	24.6	
		offline	21.6	21.9
band 3	25.6	equal	25.1	
		offline	24.8	23.8
band 4	21.4	equal	21.4	
		offline	23.8	22.4

Table 1: *Word error rates (WER) on clean and band-limited noise at 12 dB SNR on* ***MFCC*** *features for the fullband system and the MB system of 5 bands (i.e. the 4 subbands and the fullband) using equal weights and offline ML-weights.*

initial experiments were carried out on band-limited noise (in one subband at a time). We calculated the offline ML-weights for clean speech and each of the noises using MFCC and PLP features, the first of which can be seen in **Figures 1**, **2** and **3**. Clearly, for clean speech the weights depend on both the subband and the respective phoneme and thus change from phoneme to phoneme (**Fig. 1**). For noise-corrupted speech however, it can be seen how the noisy band gets consistently downweighted as compared to the clean subbands (**Fig. 2** and **3**). Results for MFCC features are given in **Tabs. 1**, **2**

noise **SNR 0 dB**	fullband WER	MB system w/ **sum rule** weights	WER	cheat
clean	13.5	equal	17.2	
		offline	15.2	
band 1	85.7	equal	49.6	
		offline	46.4	39.6
band 2	63.4	equal	32.4	
		offline	22.6	24.6
band 3	51.6	equal	30.7	
		offline	29.5	26.0
band 4	44.2	equal	29.7	
		offline	29.2	25.6

Table 2: *WER on clean and band-limited noises at 0 dB SNR on* ***MFCC*** *features for the fullband system and the MB system of 5 bands (i.e. the 4 subbands and the fullband) using equal weights and offline ML-weights in the* ***sum rule****.*

and **3**, for the PLP features in **Tab. 4**. As compared to the baseline fullband systems (2^{nd} columns), the MB systems already show higher noise robustness when using equal weights (4^{th} columns, upper number), with the exception of one noise c. Only in the case of clean speech give the MB systems weaker performance which is due to using 4 classifiers only (cf. PLP features). Including the fullband classifier, as was done for the MFCC features, already improves performance. It can be expected that more competitive performance of the MB system as compared to the fullband system in clean, would easily be achieved by extending the MB system to consider all possible combinations of subbands [8].

noise **SNR 0 dB**	fullband WER	MB system w/ **product rule** weights	WER	cheat
clean	13.5	equal	13.0	
		offline	13.8	-
band 2	63.4	equal	36.1	
		offline	19.9	15.0
band 3	51.6	equal	42.8	
		offline	23.6	17.4

Table 3: *WER on clean and band-limited noises at 0 dB SNR on* ***MFCC*** *features for the fullband system and the MB system of 5 bands (i.e. the 4 subbands and the fullband) using equal weights and offline ML-weights in the* ***product rule****.*

Comparing the new ML-weights to equal weighting in noise-corrupted speech, shows that the MB system using PLP features (**Tab. 4**), gains improved robustness for noise in subbands 1, 2 and 4 whereas results in band 3 worsened although control calculation of the data likelihood showed that also for this subband the likelihood was increased. For the MFCC features (**Tabs. 1**, **2** and **3**), using the ML-weights almost always gave higher noise robustness (with the exception of noise at 12 dB in

noise **SNR 0 dB**	fullband WER	MB system w/ **sum rule** weights	WER
clean	9.3	equal offline	21.1 22.9
band 1	64.4	equal offline	51.4 44.7
band 2	75.7	equal offline	50.9 44.7
band 3	89.2	equal offline	65.8 72.7
band 4	90.7	equal offline	57.2 56.5

Table 4: *WER on **PLP** features for the fullband system and the MB system of 4 subbands using equal weights and offline ML-weights in the **sum rule**.*

subband 4) than equal weights, in two cases outperforming the cheating experiment ('cheat'). (The results from the 'cheating' experiments indicate to what extent the MB system could be improved by an optimal weighting strategy). Also for clean speech the ML-weights increased performance on the MFCC features when using the sum rule. (For sake of performance in clean speech, for this set of features the MB system also included the fullband). Due to the fullband always being corrupted by noise and although the ML-weights were estimated to downweight the fullband in noise, the overall gain on each noise was less striking than for a pure MB system (cf. PLP features) consisting of the 4 subbands only (which though performs much worse in clean). Using the product rule, an improvement with ML-weights in clean speech was not observed, but considerable noise robustness was achieved with the ML-weights for all band-limited noise cases (cf. Table **3**), also when the fullband was included.

noise $\alpha =$	offline 0	online 0.2	0.5	1
band 1	51.4	45.2	45.9	46.7
band 2	50.9	43.7	42.8	44.7
band 3	65.8	71.3	71.3	71.7
band 4	57.2	59.5	57.7	58.2

Table 5: *WER on **PLP** features for the MB system using equal and online ML-weights on band-limited noise (sum rule).*

3.2 Online ML Expert Weights Adaptation

Next, we tested the online version of the ML-weights which were calculated as described in (7). $\alpha = 0$ in this case corresponds to initial (i.e. not updated) equal weights. $\alpha = 1$ only takes the newly calculated values from the last 100 frames into account and disregards former weight values.

It can be seen in **Table 5** that for this kind of stationary noise, lower α-values ($\alpha = 0.2$ or 0.5) usually give better results as they rely more on global weight estimates. It is however expected that this approach should give more benefit in case of non-stationary noise.

4 CONCLUSION

As could be seen in the experiments, MB speech recognition is usually more robust to band-limited noise than a fullband system. With an appropriate weighting strategy, such as the ML-weights introduced in this article, the MB system could be rendered even more competitive. For more realistic noise conditions and to further improve MB performance in clean speech, we have to either employ the so-called "full combination" approach, in which all possible combinations of subbands are considered [8], or change to the multi-stream domain [6].

For both alternatives, the ML-weights can be as easily used as in MB processing showed so far in this article. We thus plan on employing this new ML-weighting also in the before mentioned frameworks. Moreover, for more appropriately testing the online ML-weights adaptation we will extend our experiments to include *non*-stationary band-limited noise and different lengths N for the update window.

We usually work in the framework of HMM/MLP hybrid systems, which we found to be more powerful than Gaussian Mixture Expert/HMM systems. As in HMM/MLP hybrid systems the likelihoods which are needed for ML-weighting are not available, we need to consider how the ML-weights adaptation could be derived directly for use in the posterior-based approach of HMM/MLP hybrid systems.

References

[1] Ch.M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[2] M. Cooke, A. Morris, and P. Green. Recognising occluded speech. *Proceedings of the ESCA Workshop on the auditory basis of speech perception*, pages 297–300, 1996.

[3] D.P.W. Ellis and J.A. Bilmes. Using mutual information to design feature combinations. *Int. Conf. on Spoken Language Processing*, 3:79–82, 2000.

[4] H. Fletcher. *Speech and Hearing in Communication*. Krieger, New York, 1953.

[5] A. Hagen and A. Morris. Comparison of HMM experts with MLP experts in the full combination multi-band approach to robust ASR. *Int. Conf. on Spoken Language Processing*, 1:345–348, 2000.

[6] A. Hagen, A. Morris, and H. Bourlard. From multi-band full combination to multi-stream full combination processing in robust ASR. *ISCA ITRW ASR2000*, pages 175–180, 2000.

[7] H. Hermansky, S. Tibrewala, and M. Pavel. Towards asr on partially corrupted speech. *Int. Conf. on Spoken Language Processing*, pages 462–465, 1996.

[8] A. Morris, A. Hagen, H. Glotin, and H. Bourlard. Multi-stream adaptive evidence combination to noise robust ASR. *Speech Communication*, 34(1-2), 2001.