IDIAP RESEARCH REPORT

IDIAP
Martigny - Valais - Suisse

Comparison of HMM experts with MLP experts in the Full Combination Multi-Band Approach to Robust ASR

Astrid Hagen § Andrew Morris §

IDIAP–RR 00-21

July 2000

to appear in
Int. Conf. on Spoken Language Processing, Beijing 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
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,morris}@idiap.ch.

IDIAP Research Report 00-21

Comparison of HMM experts with MLP experts in the Full Combination Multi-Band Approach to Robust ASR

Astrid Hagen Andrew Morris

July 2000

to appear in
Int. Conf. on Spoken Language Processing, Beijing 2000

Abstract. In this paper we apply the Full Combination (FC) multi-band approach, which has originally been introduced in the framework of posterior-based HMM/ANN (Hidden Markov Model/Artificial Neural Network) hybrid systems, to systems in which the ANN (or Multilayer Perceptron (MLP)) is itself replaced by a Multi Gaussian HMM (MGM). Both systems represent the most widely used statistical models for robust ASR (automatic speech recognition). It is shown how the FC formula for the likelihood-based MGMs can easily be derived from the posterior-based approach by simply applying Bayes' Rule. The experiments show that the Full Combination multi-band system with MGM experts performs better, in all noise conditions tested, than the simple sum and product rules which are normally used. As compared to the baseline full-band system, the FC system shows increased robustness mainly on band-limited noise. The goal of this article is not a performance comparison between Multilayer Perceptrons and Multi Gaussian Models but between the *theory* of the two approaches, *posterior-based* vs. *likelihood-based* FC approach, so results are only given for the MGMs.

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

The Full Combination multi-band approach to robust ASR was recently proposed [13]. This model, based on HMM/ ANN hybrid systems, trains an ANN expert for all possible combinations of sub-bands, thereby avoiding the otherwise necessary assumption of independence between sub-bands. Since this model was introduced for posterior-based systems, it has been shown to perform better than both the full-band hybrid system and the former sub-band recombination approaches, in both narrow- and wide-band noise [9, 12].

As in other domains of ASR, likewise in multi-band ASR, Multi Gaussian Models [1, 6, 15] and HMM/ANN hybrid systems [5, 11] are the most widely used statistical models. Both approaches have their advantages and disadvantages, especially when applied to the multi-band paradigm. In favor of the HMM/ANN hybrid, the ANN (or Multilayer Perceptron (MLP)) is able to capture discriminate, dynamic features and reduces their dimensionality while preserving contextual information, whereas MGMs are trained using the maximum-likelihood criterion, which is not discriminate.

However, there are two important reasons why it is interesting to look at using MGM experts in place of the MLP (Multi-Layer Perceptron) experts which are usually used in the multi-band HMM/ANN hybrid. One reason is that to make the best use of full-combination multi-band processing, a separate MLP expert must be trained for each sub-band combination, which renders this approach almost infeasible if a higher number of sub-bands is chosen. For Multi Gaussian Models, on the other hand, all MGM experts can be derived directly from a single full-band MGM, using marginalization (missing data theory [4]). The other reason is that MGMs can be trained without the necessity of labeled data, so permitting the use of larger unlabeled data sets. MGMs also permit hidden state modeling and offer potential for likelihood based noise and speaker adaptation.

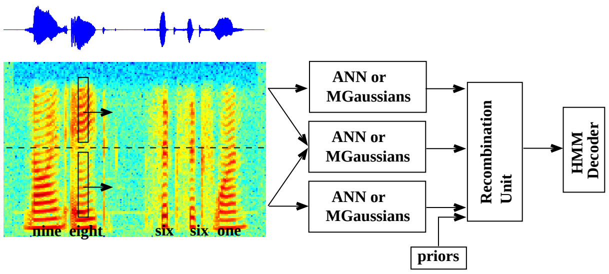

Figure 1: Illustration of Full Combination with ANN or Multi-Gaussian (MGaussian) Classifiers for two sub-bands.

The FC multi-band scheme has so far only been developed and tested using MLP experts. These output posterior probabilities which are recombined and then passed on to the HMM decoder as scaled likelihoods (cf. Section 2). In this paper we show how the FC approach can also be used with MGM experts (cf. Section 3). The underlying theory for likelihood combination is derived directly from the theory of posteriors combination, using Bayes' Rule. Both approaches, with MGM or MLP experts, use the same expert reliability weighting. It has been found that using equal weights already gives improved performance over baseline systems, such as the full-band expert or simple recombination schemes via sum or product of sub-bands. Experiments, incorporating equal weights, were carried out on the Numbers95 database with different kinds of noise added. They will be discussed in Sections 5 and 6.

2 FC in posterior-based systems

In the Full Combination approach we suppose that at each instant one combination of sub-bands x_{c_j} (with $j = 1..2^d$ combinations for d spectral sub-bands) is the largest combination which is free from noise, and thus carries the most useful data for identifying the current phoneme. Considering all 2^d possible combinations of sub-bands (as illustrated in Figure 1 for the case of 2 sub-bands) and using the fact that the events " $best_j$ = combination x_{c_j} is largest clean combination" are exhaustive and mutually exclusive, we can decompose the full-band posterior probability for each phoneme q_k into a composition of weights w_j and clean combination posteriors $P(q_k|x_{c_j})$ as illustrated in (1). Providing that the number of combinations is not too large, we can train an MLP expert on the clean data from each combination x_{c_j} to output phoneme posteriors $P(q_k|x_{c_j})$.

$$w_j = P(best_j|x)$$
$$P(q_k|x) \simeq \sum_{j=1}^{2^d} w_j P(q_k|x_{c_j}) \tag{1}$$

To avoid training of 2^d ANNs the combination posteriors $P(q_k|x_{c_j})$ can be approximated by only using the $P(q_k|x_i)$ issued from the observations x_i, $i \in \{1..d\}$ of the d sub-bands. Supposing class-conditional independence [8], we can thus approximate (1) by:

$$\begin{aligned} P(q_k|x) &\simeq \sum_{j=1}^{2^d} w_j \hat{P}(q_k|x_{c_j}) \qquad (2) \\ \text{where} \quad \hat{P}(q_k|x_{c_j}) &= \alpha_j P^{1-|c_j|}(q_k) \prod_{i \in c_j} P(q_k|x_i) \end{aligned}$$

α_j chosen such that $\prod_K \hat{P}(q_k|x_{c_j}) = 1$.

This approximation yields improved noise robustness to (artificial) band-limited noise but cannot compete with the real FC system when used in more realistic, wide-band noise conditions [8]. For the likelihood-based system, a corresponding approximation is also possible as will be pointed out in the next section. An approximation for the likelihood-based system though is not as crucial as for the posterior-based system as the 2^d combinations could be gained from one single MGM by using marginalization, i.e. no training of 2^d MGMs would be necessary as it is the case (up to now) for the posterior-based (MLP) experts. A way to also circumvent the extensive training for the posterior-based system would be possible by substituting the MLP experts by a Gaussian Radial Basis Function Network as pointed out in [14].

3 FC in likelihood-based systems

When using Multi Gaussian Models in multi-band ASR we receive at each time step from each sub-band MGM a local stream likelihood for each phoneme. These local stream likelihoods have to be recombined, just as it was the case with posterior probabilities, to acquire a common local likelihood. These final likelihoods are then used by the HMM decoder to classify the current phoneme.

As is well known, posterior probabilities can be converted to likelihoods (and v.v.) by applying Bayes' Rule:

$$P(q_k|x) = \frac{p(x|q_k)P(q_k)}{p(x)} \tag{3}$$

Utilizing (3) to convert (1) we obtain:

$$\begin{aligned} \frac{p(x|q_k)}{p(x)} &\simeq \sum_{j=1}^{2^d} w_j \frac{p(x_{c_j}|q_k)}{p(x_{c_j})} \qquad (4) \\ \text{with} \quad p(x_{c_j}) &= \sum_{k'=1}^{K} p(x_{c_j}|q_{k'})p(q_{k'}) \end{aligned}$$

$w_j = P(best_j|x)$ same as for posterior-based systems.

Though training of MGMs is usually faster than MLP training and the sub-band MGMs could be derived directly from the full-band MGM as pointed out before, we could theoretically also approximate the (likelihood-based) FC approach in the same way as for the posterior-based FC system. As before, the approximation assumes conditional independence:

$$\frac{p(x|q_k)}{p(x)} \simeq \sum_{j=1}^{2^d} w_j \frac{\prod_{i \in c_j} p(x_i|q_k)}{\sum_{k'=1}^{K} \prod_{i \in c_j} p(x_i|q_{k'})p(q_{k'})} \tag{5}$$

4 Databases and Systems Setup

Experiments were carried out on a test set of 200 utterances from the Numbers95 database [3] of connected numbers recorded over the telephone line. For robustness tests, different kinds of noise were added to the clean test set at various signal-to-noise ratios (SNR). We found rather divers behavior of our posterior-based FC system to different noise conditions [13], and for the reasons outlined in Section 1, we were interested in employing the same kind of noises for the likelihood-based FC system. We thus chose artificial narrow-band noise in sub-bands 1 and 2 as well as real-environmental wide-band noises. These included factory noise from the Noisex92 database [16] and an in-house database of car noise from Daimler Chrysler.

The multi-band systems were set up as follows. The frequency domain was split into 4 sub-bands, covering the ranges of [115-629 Hz], [565-1370 Hz], [1262-2292 Hz] and [2122-3769 Hz]. J-rasta features [10] were extracted from each of the sub-bands as well as from each of the combinations, resulting in a set of 15 feature stream vectors[1] (The stream belonging to combination 0 which does not have any features was neglected). The number of parameters (i.e. LP analysis order and coefficients count) for each stream was increased proportional to the size of the stream in such a way that the number of parameters in a combination corresponds to the sum of the parameters from each of its constituents. Each feature vector was then used to train the Multi Gaussian Model for the respective stream, after having been augmented by all delta and delta-delta components. The Multi Gaussian Models were 78 3-state triphone models with 64 Gaussian mixtures per triphone, using diagonal covariance matrices. Training was carried out on clean data only.

	Wide Band Noise				Clean
	Car		Factory		45 dB
	12 dB	0 dB	12 dB	0 dB	
Full-band	14.2	37.1	13.4	36.4	8.5
simple sum	22.5	56.2	23.9	61.4	12.5
simple product	20.1	55.9	19.2	57.5	11.8
FC-sum	16.0	49.9	16.5	52.0	10.6
FC-product	11.0	39.0	11.8	41.9	8.1

Table 1: Word error rates (WER) for clean speech and for speech corrupted with band-limited noise in sub-bands 1 and 2 for the Full Combination system using MGM experts according to (4).

Besides the FC sum in (4), another approach for combining expert outcomes which has proven to be very efficient is combination by multiplication (the so-called “product rule”) [2, 7]. While the product rule implies the assumption of independence between the probabilities from different experts, which is clearly not the case for FC experts, so the “sum rule” for probabilities assumes mutual exclusivity

[1]The feature stream comprising all 4 sub-bands and with that the full frequency domain constitutes the input to our (full-band) baseline system.

– which is apparently not always the case. Experimental results have demonstrated that this method can be a very effective way to combine multiple classifiers.

5 Experimental Results

The results for the experiments on wide-band car and factory noise can be seen in Table 1 (as well as Figure 2 for factory noise). The FC approach for both the sum (FC-sum) and product rule (FC-product) did significantly improve performance as compared to the original methods [2, 11] of combining the outputs from only the d experts by a simple sum or product. When comparing to the baseline full-band system in line 1, on the other hand, none of the proposed sub-band methods showed any significantly increased noise robustness. On clean speech, only the FC-product method was able to compete with the baseline system.

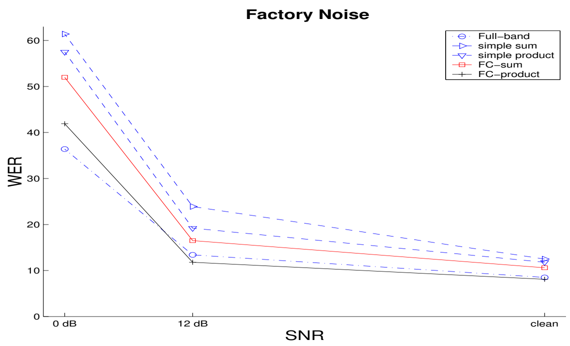

Figure 2: Comparison of WERs for the baseline full-band system, the simple sum and product rules as well as the FC-sum and FC-product systems for speech with factory noise.

	Narrow Band Noise			
	Band 1		Band 2	
	12 dB	0 dB	12 dB	0 dB
Full-band	16.4	31.1	17.9	42.9
simple sum	15.8	32.5	26.8	52.5
simple product	15.8	31.6	26.8	47.8
FC-sum	14.0	30.9	20.1	43.8
FC-product	10.6	24.6	15.5	33.4

Table 2: Word error rates (WER) for speech corrupted with artificial narrow-band noise in sub-bands 1 and 2 for the baseline full-band system, the simple sum and product rules as well as the FC-sum and FC-product systems.

As with the posterior-based FC system, the likelihood-based FC system showed its main advantage on band-limited noise (cf. Table 2). Whereas the simple product and sum rules deteriorate also on band-limited noise, as compared to the baseline full-band system, the FC-product rule resulted in significantly improved noise robustness on this kind of noise.

6 Conclusion

Results presented show that the Full Combination multi-band system with MGM experts performs better, in all noise conditions tested, than both the simple sum and product rules. The same result has also been found for the posterior-based FC system. When compared to the baseline full-band system, on the other hand, the likelihood-based FC approach (using equal weights) was only competitive on clean speech and high SNR conditions and only when using the product rule. The multi-band experiments conducted in this article used equal weights only. Although these were often found to work well, especially for band-limited noise, performance improvements are still possible when more appropriate weighting strategies are used. Different weighting schemes as employed for the posterior-based approach in [9], will be investigated also for the likelihood-based FC approach in the future.

As pointed out in the introduction, all MGM experts could directly be derived from a single full-band MGM by using marginalization. We plan to investigate this approach in order to compare its performance to the FC approach where all 2^d experts were explicitly trained as done in this article.

Acknowledgments:

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).

References

[1] Ch. Cerisara, D. Fohr, and J. P. Haton. Robust behavior of multi-band paradigm. *Proceedings of the ESCA Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pages 187–190, 1999.

[2] H. Christensen, B. Lindberg, and O. Andersen. Employing heterogeneous information in a multi-stream framework. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, III:1571–1574, 2000.

[3] R.A. Cole, M. Noel, T. Lander, and T. Durham. New telephone speech corpora at CSLU. *Proc. European Conf. on Speech Communication and Technology*, 1:821–824, 1995.

[4] M. Cooke, A. Morris, and P. Green. Missing data techniques for robust speech recognition. *Int. Conf. on Spoken Language Processing*, pages 863–866, 1997.

[5] S. Dupont. Études et développement de nouveaux paradigmes pour la reconnaissance robuste de la parole. *Ph.D. Thesis*, 2000.

[6] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J. M. Boite. Hybrid HMM/ANN Systems for training independent tasks: Experiments on PHONEBOOK and related improvements. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 3:1767–1770, 1997.

[7] A. K. Haberstadt and J. R.Glass. Heterongeneous measurements and multiple classifiers for speech recognition. *Int. Conf. on Spoken Language Processing*, 3:995–998, 1998.

[8] A. Hagen and H. Glotin. Études comparatives de l'approche 'full combination' et de son approximation sur bruits roses. *Journée d'Études sur la Parole, Aussois*, pages 317–320, 2000.

[9] A. Hagen, A. Morris, and H. Bourlard. Different weighting schemes in the full combination subbands approach in noise robust ASR. *Proceedings of the ESCA Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pages 199–202, 1999.

[10] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[11] N. Mirghafori. A multi-band approach to automatic speech recognition. *Ph.D. Thesis, ICSI*, 1999.

[12] A. Morris, A. Hagen, and H. Bourlard. The full combination sub-bands approach to noise robust HMM/ANN-based ASR. *Proc. European Conf. on Speech Communication and Technology*, 2:599–602, 1999.

[13] A. Morris, A. Hagen, H. Glotin, and H. Bourlard. Multi-stream adaptive evidence combination to noise robust ASR. *Speech Communication*, (to appear), 2000.

[14] A. Morris, L. Josifovski, H. Bourlard, M. Cooke, and P. Green. A neural network for classification with incomplete data: Application to robust asr. *Int. Conf. on Spoken Language Processing*, 2000.

[15] S. Okawa, E. Bocchieri, and A. Potamianos. Multi–band speech recognition in noisy environment. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2:641–644, 1998.

[16] A. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. *Technical Report, DRA Speech Research Unit*, 1992.