IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

From Multi-Band Full Combination to Multi-Stream Full Combination Processing in Robust ASR

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

IDIAP–RR 00-20

July 2000

to appear in
ISCA Tutorial and Research Workshop ASR2000, Paris, France, 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,bourlard}@idiap.ch.

IDIAP Research Report 00-20

From Multi-Band Full Combination to Multi-Stream Full Combination Processing in Robust ASR

Astrid Hagen Andrew Morris Hervé Bourlard

July 2000

to appear in

ISCA Tutorial and Research Workshop ASR2000, Paris, France, 2000

Abstract. The multi-band processing paradigm for noise robust ASR was originally motivated by the observation that human recognition appears to be based on independent processing of separate frequency sub-bands, and also by "missing data" results which have shown that ASR can be made significantly more robust to band-limited noise if noisy sub-bands can be detected and then ignored. Of the different multi-band models which have been proposed, only the "Full Combination" or "all-wise" multi-band HMM/ANN hybrid approach allows us to consistently overcome the difficult problem of deciding which sub-bands are noisy, by integrating over all possible positions of noisy sub-bands. While this system has performed better than any other multi-band system which we have tested, we have also found that it only shows significantly improved robustness to noise when the noise is strongly band-limited. In real noise environments this is rarely the case. An alternative paradigm for noise robust ASR is multi-stream, as opposed to multi-band, ASR. In multi-stream processing the aim is to combine evidence from a number of different representations of the full speech signal, rather than from a number of frequency sub-bands. Several models for multi-stream ASR have recently reported significant performance improvements for speech with real noise. In this article we first present evidence to show how multi-*band* ASR has a strong advantage over the baseline system with band-limited noise, but no clear advantage with wide-band noise. We then show how the principled theoretical basis for Full Combination multi-band ASR can be directly transfered to multi-*stream* combination, and we show how this model can be used to combine data streams comprising three commonly used types of acoustic features. Preliminary results show significantly improved recognition with clean speech.

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

Multi-band automatic speech recognition was largely motivated by Fletcher's theory of human perception [1, 9], in which sub-bands of the auditory spectrum are processed independently and later recombined for making a global decision. Experiments in ASR with missing data theory have also shown that recognition can be improved when noisy sub-bands can be detected and ignored [5, 17].

If in automatic sub-band processing we process frequency sub-bands independently, a lot of correlation information between feature components is lost. As it is moreover not easy to decide which sub-band(s) are reliable and which are corrupted by noise, we should integrate over all possible positions of reliable data. For this we must process and then combine not just individual sub-bands, but every possible combination of sub-bands. This approach was recently introduced as the "Full Combination" (FC) approach [12, 19] to multi-band speech recognition.

Results show that multi-band probability combination usually yields significant improvements mainly in the case of band-limited noise [11, 16]. In more realistic noise conditions, such as real-environmental car or factory noises, a slight increase in robustness can sometimes be gained, but only when using sophisticated weighting strategies [13, 19].

In multi-*stream* ASR each input stream consists of features stemming from the whole frequency domain, as opposed to multi-*band* processing where each stream comprises features from several smaller spectral regions. Some variations of the multi-stream approach have shown promising results under various noise conditions [3, 16]. The power of multi-stream processing lies in its ability to constructively combine information from different sources. For maximum benefit, the expert processing each stream should have complementary error characteristics. The combination schemes so far used in multi-stream ASR have mainly used arbitrary expert combination techniques. In this article, we show how the theoretically well-founded Full Combination (or "all-wise") approach can and should also be used in the framework of multi-stream processing. An important point to note is that the underlying theory can easily be transferred from the multi-*band* approach to the multi-*stream* approach.

As with FC multi-band, in which it is not known which combination of sub-bands comprises the largest set of uncorrupted data features, so also with FC multi-stream, it is not known which combination of data streams will give the greatest recognition performance. Therefore, for the multi-stream Full Combination scheme, the feature vectors from each stream have, first, to be concatenated into all possible combinations of feature vectors. Second, each stream combination is processed independently by different experts. The phoneme probability outputs from all experts are then combined in a linear weighted sum, via the Full Combination formula, before being passed on to the decoder. The multi-stream Full Combination approach will be further discussed in Section 3.

Experiments were carried out on the Numbers95 database [4] of connected numbers recorded over the telephone line. Different kinds of noise were added to the clean test set at various signal-to-noise ratios (SNR), like real-environmental factory noise from the Noisex92 database [20] and an in-house car noise from Daimler Chrysler as well as artificial (constant and time-varying "siren") band-limited noise. The results are discussed in Sections 2 and 4.

2 Full Combination in Multi-Band ASR

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 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 Equation (1). Providing that the number of combinations is not too large, we can train a Multi-layer Perceptron (MLP) expert on the clean data from each combination x_{c_j} to output phoneme posteriors $P(q_k|x_{c_j})$.

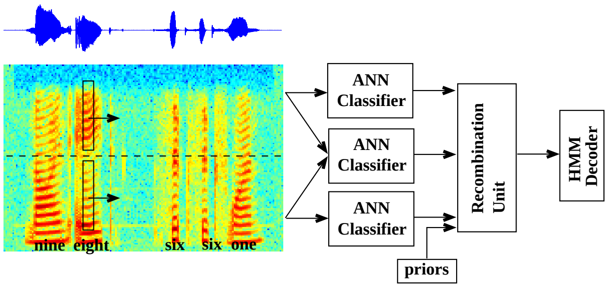

Figure 1: Illustration of Full Combination in multi-band ASR with two sub-bands.

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

Various methods for estimation of the weights w_j were described in [13, 19]. Here we would like to illustrate only one method for static weights estimation. These relative frequency (RF) based weights are estimated as follows:

$$w_j = \frac{n_j}{n} \tag{2}$$

where n_j is the number of frames of training data for which expert j has the largest posterior, across all experts, for the target phoneme, and n is the total number of frames of training data.

Tests were run in the framework of hybrid HMM/MLP systems, the MLP-part of which needed to be trained for each combination of 4 sub-bands. For the input to the MLPs, we used j-rasta features [15] accordingly extracted from the 4 spectral bands, covering the ranges of [115-629 Hz], [565-1370 Hz], [1262-2292 Hz] and [2122-3769 Hz]. The size of the hidden layer was changed according to the number of input features, between 660 and 1750 units. MLP output nodes corresponded to the 27 phonemes used in the database. The training set comprised 3233 utterances of clean speech from the Numbers95 database. For details see [11]. Experiments were carried out on a test set of 200 utterances from this database which were corrupted with different noises: band-limited pink noise in each of the 4 sub-bands (cf. Table 2), artificial siren noise (cf. Table 1 and Figure 3) and wide-band car and factory noises (cf. Table 3), at SNRs of 0 and 12 dB. The car noise stems from real in-car recordings by Daimler Chrysler. The factory noise is taken from the Noisex92 database [20], which also constitutes a real-environmental noise. Word entrance penalties for each system were fixed to give optimal performance on clean data.

In the experiments, the FC multi-band system was compared to the baseline full-band system and the early multi-band approach, which combines the posteriors from the 4 sub-bands by product rule. The results, which can be seen in Tables 2 and 1, show a significant advantage for the Full Combination approach with stationary, band-limited noise and an even stronger advantage with non-stationary band-limited "siren" noise (which j-rasta features are unable to eliminate). This is already true for equal weights and the RF weights gave no further significant improvement. However, with wide-band noise the multi-band approaches show no significant advantage, as shown in Table 3, although the FC approach yielded higher robustness than the original multi-band approach. RF-weighting in FC further improved robustness to this kind of noise but not to an extend where it outperforms the full-band system. RF weights rendered the FC system competitive to the full-band system in clean speech (cf. Table 3).

	Siren		Avg.
	12 dB	0 dB	WER
Baseline	36.9	89.4	63.2
Early MB	31.0	53.8	42.4
FC-MB equal	18.9	39.0	29.0
FC-MB RF	20.1	42.1	31.1

Table 1: WERs for Full Combination multi-band (FC-MB) system using equal and relative frequency (RF) weights compared to the baseline full-band system and the early multi-band (MB) approach on time varying band-limited noise.

	Narrow-Band Noise								
	Band 1		Band 2		Band 3		Band 4		Avg.
	12 dB	0 dB	12 dB	0 dB	12 dB	0 dB	12 dB	0 dB	WER
Baseline	14.0	31.4	16.6	44.6	18.9	35.0	17.4	23.9	25.2
Early MB	20.6	34.6	21.8	41.1	23.1	33.1	23.0	25.5	27.9
FC-MB equal	12.0	27.1	14.9	24.1	14.6	21.5	14.4	16.6	18.2
FC-MB RF	12.0	29.0	13.8	22.5	14.5	22.2	12.9	15.6	17.8

Table 2: Word error rates for Full Combination multi-band system using jrasta features with equal and relative frequency (RF) weights compared to the baseline full-band system and the early multi-band approach (Early MB) on band-limited noise in sub-bands 1 to 4.

	Car		Factory		clean	Avg.
	12 dB	0 dB	12 dB	0 dB	45 dB	WER
Baseline	10.6	32.8	11.4	34.6	8.0	17.6
Early MB	21.0	56.6	22.2	57.8	14.1	31.0
FC-MB equal	10.6	41.8	11.9	42.2	8.9	20.7
FC-MB RF	10.1	35.1	10.0	36.8	8.0	18.0

Table 3: Word error rates for Full Combination multi-band system using equal and relative frequency (RF) weights compared to the baseline full-band system and the early multi-band approach on wide-band car and factory noise and clean speech. The last column shows the average word error over all noise conditions of 0, 12 and 45 dB car and factory noise.

3 Full Combination in Multi-Stream ASR

As could be seen in the preceding section, the Full Combination multi-band approach shows competitive performance to that of the baseline full-band system in clean speech. For noise-corrupted speech, the advantage of FC depends on the respective noise condition. FC is very powerful in band-limited noise as has already been pointed out in other studies [11]. On the more realistic noise cases such as car and factory noise though, the Full Combination multi-band approach needs sophisticated weighting strategies in order to come close to the performance of the full-band system, still without gaining higher robustness to this kind of noise.

Considering the fact that the FC multi-band system never performs consistently better than the full-band system for all noise conditions, we should rather look into how to extract the advantages of this approach but model a more robust system for all noise conditions.

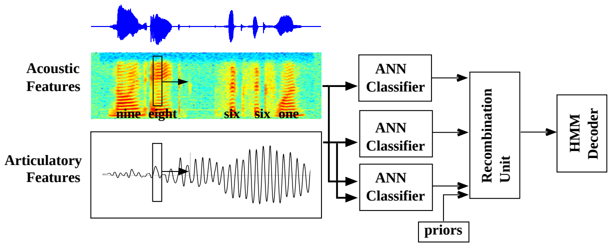

Figure 2: Illustration of Full Combination in multi-stream ASR on two streams using different time-scales.

With wide-band noise we need to appeal to another mechanism for robustness to noise which does not rely on a significant proportion of the data remaining completely uncorrupted. There are many possibilities in ASR for improving robustness to noise by combining evidence from multiple data streams [18]. Examples include combining vision (e.g. mouth shape) with acoustics [7], or acoustic features from different time scales [8, 10, 21]. These successful experiments in multi-stream combination have experimented with a number of linear and non-linear combination techniques, and the effectiveness of any given technique may vary according to the particular problem to which it is applied. However, it has been theoretically proven that when the outputs from a given number of experts are linearly combined (even as a simple average) then:

> "if we can increase the spread of the predictions of the committee members without increasing the errors of the individual members themselves, then the committee error will decrease" Bishop [2, p. 369].

This is the reason why we propose in this article to follow the theoretically consistent Full Combination approach as described in Section 2 and Equation (1), but now in the framework of a system for combining streams which comprises multiple complementary descriptions of the whole signal, rather than separate frequency sub-bands (cf. Fig. 2). In the present study we apply this technique to combine three different kinds of acoustic features: plp, j-rasta and mfcc features. In future we will also be looking at combining features with more widely varying error characteristics, such as acoustic features from different time scales [8, 10, 21] and acoustic features with articulatory, visual and phonetic features.

4 Test Results

For the multi-stream FC HMM/MLP hybrid system we used the same training and test conditions as for the multi-band system described in the Section 2. The goal was to test the multi-stream system on the real-environmental noises where the multi-band FC system had failed. Moreover, we included the noise case (siren) for which multi-band FC had worked best to see whether multi-stream FC will be as powerful. The different feature streams comprised plp [14], j-rasta [15] and mfcc [6] features. One MLP was trained for each combination of these features, which results in 7 MLPs (the combination which includes no features being excluded). For comparison, they were not only tested in the framework of multi-stream FC but also on their own. Results can be seen in Tables 4 and 5.

As compared to the baseline (which used jrasta features because these were easily the most robust to real fullband noise), multi-stream FC shows increased robustness on non-stationary band-limited noise (cf. Table 4), though to a lesser degree than for multi-band FC on this kind of noise (cf. Figure 3). On car and factory noise, the multi-stream FC results indicate that other than equal weights will be needed to render the system competitive, as it had been the case for multi-band FC. However, Table 5 shows that significant improvement was gained with multi-stream FC with equal weights on clean speech as compared to both the baseline system and to multi-band FC.

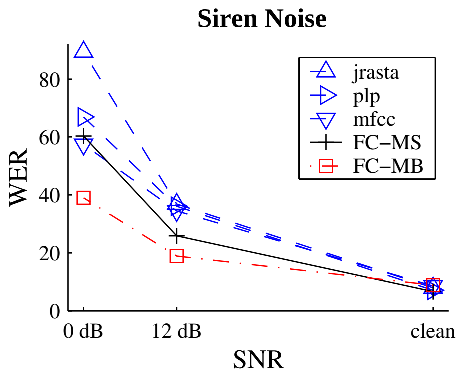

Figure 3: Word error rates (WER) for siren noise at 0 and 12 dB SNR for the 3 pure full-band systems using j-rasta, plp and mfcc features as well as the full combination multi-band (FC-MB) and the full-combination multi-stream (FC-MS) systems.

	Siren		Avg.
	12 dB	0 dB	WER
jrasta	36.9	89.4	63.2
plp	36.1	66.9	51.5
mfcc	34.5	57.4	46.0
plp-jrasta	32.9	68.2	50.6
plp-mfcc	29.9	61.6	45.8
mfcc-jrasta	30.2	74.6	52.4
plp-jrasta-mfcc	28.9	67.0	48.0
FC-MS	25.9	60.3	43.1

Table 4: WERs for Full Combination Multi-Stream (FC-MS) for siren noise. The FC-MS system (using equal weights) is compared to each one of the single or concatenated (referred to by "-") feature systems by itself. The "jrasta" system corresponds to the baseline in Section 2. In the last column the average word error rate over all noise conditions is given for each system.

	Car		Factory		Clean	Avg.
	12 dB	0 dB	12 dB	0 dB	45 dB	WER
jrasta	10.6	32.8	11.4	34.6	8.0	17.6
plp	13.8	50.5	14.6	52.6	7.1	24.3
mfcc	22.6	63.4	21.2	68.8	8.6	32.2
plp-jrasta	13.6	50.1	14.4	48.0	8.0	23.7
plp-mfcc	14.0	50.9	14.5	54.1	6.9	24.6
mfcc-jrasta	14.1	48.4	14.6	47.4	7.1	23.1
plp-jrasta-mfc	13.4	48.1	13.5	49.8	6.9	23.1
FC-MS	12.6	47.5	14.0	46.9	6.7	22.4

Table 5: Word error rates for Full Combination Multi-Stream on plp, j-rasta and mfcc features. The Full Combination Multi-Stream (FC-MS) system in the first row (using equal weights) is compared to each one of the single or concatenated (referred to by "-") feature systems by itself. Tests were run on wide-band car and factory noise. The 'jrasta' system corresponds to the baseline in Section 2. The last column shows the average word error over all noise conditions of 0, 12 and 45 dB car and factory noise.

5 Discussion

The Full Combination multi-band approach has previously been introduced as a model for noise robust ASR. In this article we have argued that the FC multi-band approach, though superior to any other multi-band method which we have tested, provides a significant advantage over the full-band HMM/ANN hybrid baseline only in the case where noise is strongly band-limited. We also showed how the theoretical basis of the FC multi-band approach can be applied directly to provide a principled basis for combining separate sources of speech data which may arise not from separate spectral sub-bands, but from two or more different representations of the speech signal as a whole. Experiments with combining multiple sources of speech information in the framework of HMM/ANN based ASR have recently shown that this approach can offer considerably increased robustness under several kinds of realistic noise conditions. However, the methods used for combining different data streams in these experiments [16], rather like the earlier experiments in multi-band ASR, did not have any consistent theoretical basis.

In this paper we have applied the Full Combination approach, previously used for sub-band evidence combination, to the combination of data arising from three different commonly used types of acoustic features. From the results which we have reported here it appears that this technique offers a significant advantage only for clean speech. However, these initial results were obtained simply using equal weights for each combination expert, and using data streams which, though different, were not perhaps as widely different as is necessary in order to more significantly benefit from this approach. We plan to test the full potential of this Full Combination multi-stream system in future using streams covering a wider range of different types of speech information, including acoustic data from different time scales [10], and non acoustic data such as articulatory, phonetic and visual features.

References

[1] J. B. Allen. How do humans process and recognize speech? *Transactions on Speech and Audio Processing*, 2(4):567–577, 1994.

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

[3] 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.

[4] 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.

[5] 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, Keele, UK, 1996.

[6] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 4:357, 1980.

[7] S. Dupont and J. Luettin. Using the multi-stream approach for continuous audio-visual speech recognition: experiments on the M2VTS database. *Int. Conf. on Spoken Language Processing*, 2:1283–1286, 1998.

[8] D. P.W. Ellis. Stream combination before and/or after the acoustic model. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 3:1635–1638, 2000.

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

[10] A. Hagen and H. Bourlard. Using multiple-time scales in the framework of full combination multi-stream speech recognition. *Int. Conf. on Spoken Language Processing*, (to appear), 2000.

[11] 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.

[12] A. Hagen, A. Morris, and H. Bourlard. Subband-based speech recognition in noisy conditions: The full combination approach. IDIAP-RR 15, IDIAP, 1998.

[13] 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.

[14] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.

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

[16] A. Janin, D. Ellis, and N. Morgan. Multi-stream speech recognition: Ready for prime time? *Proc. European Conf. on Speech Communication and Technology*, 2:591–594, 1999.

[17] R. P. Lippmann and B. A. Carlson. Using missing feature theory to actively select features for robust speech regonition with interruptions filtering and noise. *Proc. European Conf. on Speech Communication and Technology*, pages 37–40, 1997.

[18] N. Morgan, H. Bourlard, and H. Hermansky. Automatic speech recognition: an auditory perspective. IDIAP-RR 17, IDIAP, 1998.

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

[20] 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.

[21] K. Weber. Multiple time scale feature combination towards robust speech recognition. *Konvens, 5. Konferenz zur Verarbeitung natürlicher Sprache*, (to appear), 2000.