IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Using Multiple Time Scales in the Framework of Multi-Stream Speech Recognition

Astrid Hagen [§] Hervé Bourlard [§]

IDIAP–RR 00-22

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,bourlard}@idiap.ch.

IDIAP Research Report 00-22

Using Multiple Time Scales in the Framework of Multi-Stream Speech Recognition

Astrid Hagen Hervé Bourlard

July 2000

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

Abstract. In this paper, we present a new approach to incorporating multiple time scale information as independent streams in multi-stream processing. To illustrate the procedure, we take two different sets of multiple time scale features. In the first system, these are features extracted over variable sized windows of three and five times the original window size. In the second system, we take as separate input streams the commonly used difference features, i.e. the first and second order derivatives of the instantaneous features. In the same way, any other kinds of multiple time scale features could be employed. The approach is embedded in the recently introduced "full combination" approach to multi-stream processing in which, the phoneme probabilities from all possible combinations of streams are combined in a weighted sum. As an extension of this approach we have found that replacing the sum of probabilities by their product, in the same "all wise" context, can result in higher robustness. Capturing different information in each stream, and with the longer time scale features being more robust to noise, the multiple time scale multi-stream system gained a significant performance improvement in both clean speech and in real-environmental 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

Recently, the so-called "full combination" (FC) approach to multi-*band* speech recognition was introduced [10, 11]. This approach came into existence from the search for a way to overcome the usually necessary but restricting independence assumption in sub-band processing. By integrating over all possible combinations of sub-bands the independence assumption was no longer necessary.

Multi-band speech recognition using several distinct frequency sub-bands for feature extraction, and in this framework the theoretically well-based FC approach, was proven to work especially well on band-limited noise [5, 8]. Unfortunately, for more realistic wide-band noise the multi-band approach still does not show significantly increased noise robustness as compared to full-band processing.

Multi-*stream* processing is another approach to combining the evidence from different information streams which can be seen as a generalization of multi-band processing [6]. In multi-stream processing, each stream captures different features from the whole frequency domain. Each stream carries complementary information, and this can lead to improved performance when these streams are properly combined. For this reason, the feature extraction performed by each stream should be as different and thus as complementary as possible. This can be achieved by using either different feature extraction techniques (such as e.g. PLP and MFCC features as was done in [2, 6]) or different time-scales for feature extraction (as, e.g. for the J-Rasta and MSG features in [12]) or both approaches combined.

Multiple-time scale features seem especially promising in this framework as they extract different information in each time scale and usually show higher robustness to noise for the longer time scale features [9, 12, 14, 15]. One way of extracting multiple time scale information is by simply extending the window size for feature extraction from its regular size of, e.g., 25 ms to three or five times its size. Another possible source of more independent features from different time scales are the well-known difference features, i.e. the first and second order derivatives of the instantaneous features. In the following, we show for both methods how the full combination approach can be used to combine features from different time scales.

2 Full Combination in Multi-Stream

In the multi-band full combination (FC) approach we suppose that at each instant one combination of sub-bands x_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 [10]. As it is not known which combination of sub-bands comprises the largest set of uncorrupted data features we integrate over all possibilities. Likewise in multi-stream it is not known which combination of data streams will give the greatest recognition performance. It has been suggested that independent features are best processed as separate streams, while dependent features are best combined in the same stream. 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 (Step 1). Second, each stream combination is processed independently by different Artificial Neural Network (ANN) experts (Step 2). The phoneme probability outputs from all experts are combined in a linear weighted sum before being passed on to the decoder as scaled likelihoods.

Another approach for combining expert outcomes which has proven to be very efficient is combination by multiplication (the so-called "product rule") [2, 4]. While the product rule implies the assumption of independence between the probabilities from different experts, which is clearly violated, so the "sum rule" for probabilities assumes mutual exclusivity – which is not clearly defined here. The experimental results presented below show that this method seems to be the most effective way to combine multiple classifiers.

3 Variable Window Size Features

In the first case, acoustic features (in our case, Perceptual Linear Prediction (PLP) [7] feature vectors) are extracted from regular short-term segments of 25 ms windows, shifted every 12.5 ms. Similar features are then extracted on longer segments of three and five times the original window size, yielding window sizes of 75 ms and 125 ms (also shifted every 12.5 ms), as illustrated in Figure 1. (In Table 1, these features are respectively referred to as (1), (3) and (5)).

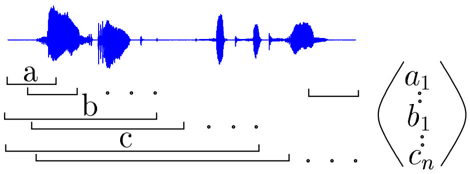

Figure 1: Illustration of multiple time scale features extracted from a regular-sized (a), triple- (b) and quintuple-sized (c) data window and concatenated to yield feature vector (1-3-5).

Since these multi-scale features are clearly correlated, and not very likely to function well by themselves, they were respectively concatenated to the instantaneous feature vector (1), resulting in feature vectors (1-3), (1-5) and (1-3-5). For each feature combination (augmented by its delta and delta-delta components), one ANN was trained on 9 frames of contextual input. The ANN (here a Multi-Layer Perceptron (MLP)), which is a feed-forward neural network, is able to model the correlation across input features.

Experimental Results

Experiments were carried out on a test set of 200 utterances from the Numbers95 database [3] of connected numbers recorded over the telephone line. Real-environmental car noise (from an in-house database) and factory noise (from the Noisex92 database [13]) were added at signal to noise ratios (SNR) of 12 and 0 dB to test the system's performance in the difficult task of speech corrupted by wide-band noise. For each system, the word entrance penalty used in decoding was optimized in clean speech and then kept constant for the experiments in noise.

Results for the "variable window size system" can be seen in Table 1. The baseline system – referred to as (1) – consisted of a regular HMM/MLP hybrid system extracting information from one, short-term window of 25 ms only. The multiple-time scale system consisted of 4 MLPs each trained on one of the different time scales (1), (1-3), (1-5), and (1-3-5). The posterior probabilities at the output of the MLPs were then combined via product rule (which was yielding the best performance). As we are working in the framework of HMM/MLP hybrid systems [1], the combined posterior probabilities are then passed on as scaled likelihoods to the HMM decoder.

Although quite disappointing, the results show some, though no consistent, improvement using the multiple time scale features as compared to the baseline system.

	Car		Factory		Clean
	12 dB	0 dB	12 dB	0 dB	100 dB
(1)	13.8	50.5	14.6	52.6	7.1
(1)*(1-3)* (1-5)*(1-3-5)	14.2	45.5	14.7	49.3	8.1

Table 1: Word error rates (WER) on clean speech and speech with car and factory noise for the combination of variable window size features extracted on window sizes of 25 ms (1), 75 ms (3) and 125 ms (5) (frame shift 12.5 ms) using HMM/MLP hybrid systems.

4 Difference Features as Multiple Time Scales in FC

The difference features used in the previous system are calculated by regression over 5 (delta features) and 7 (delta-delta features) frames of input. They thus cover respectively 75 ms and 100 ms of speech data. As these difference features are more or less independent of the instantaneous features they can more appropriately be treated as separate, higher time scale feature streams. In the following, we thus used these features as feature streams, which will be recombined according to the FC approach.

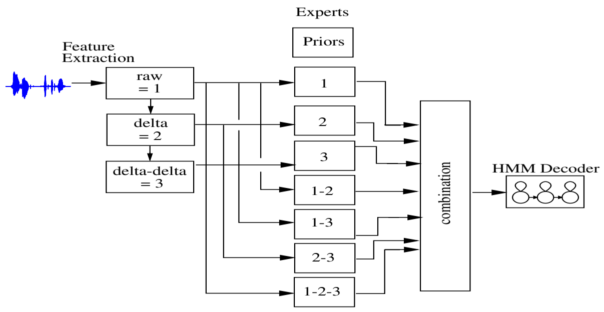

Figure 2: Illustration of recognizers combination according to the full combination approach, using raw, delta and delta-delta features as individual input streams as well as all possible combinations of feature streams in the framework of an HMM/MLP system.

These three feature streams were (according to Step 1 in Section 2) concatenated to give all possible combinations of feature streams. In our case, this amounts to a total of seven streams. For each of these seven streams an MLP recognizer was trained (Step 2). The posterior probabilities at the output of the MLPs are combined via product rule which gave increased performance as compared to recombination by an (equally weighted) sum. This procedure is illustrated in Figure 2.

Experimental Results

	Factory		Clean
	12 dB	0 dB	100 dB
raw	18.4	67.6	9.6
deltas	24.6	83.4	10.4
d-deltas	32.6	91.1	12.4
raw-d	15.0	60.8	7.9
raw-dd	14.2	58.6	7.0
d-dd	21.9	81.0	9.0
raw-d-dd (Baseline)	14.6	52.6	7.1
FC product	**11.9**	**49.2**	**6.6**

Table 2: WERs for clean speech and for speech with factory noise for each stream on its own and for the Full Combination system (combination by product rule).

In Table 2, these feature streams are respectively referred to as raw, deltas (d) and d-deltas (dd). For comparison, performance was also measured for each stream on its own (raw, deltas, d-deltas) (lines

1 to 3 in Table 2) as well as on each of their combinations (lines 4 to 7 in Table 2). The priors were excluded.

The results of the baseline system (referred to as raw-d-dd), which is the same as in the last section, are shown in line 7 of Table 2 and line 1 of Table 3.

Results of the multiple time scale FC approach using the sum rule to recombine the 7 experts gave no significant improvement and was thus neglected in the tables. Recombination using the product rule, on the other hand, gave a significant performance improvement both for clean speech (from 7.1 down to 6.6 WER) and for all noise conditions as can be seen in Table 2 (factory noise) and Table 3 (car noise).

Car Noise	12 dB	0 dB
raw-d-dd (Baseline)	13.8	50.5
FC product	**11.6**	**35.4**

Table 3: WERs for speech with car noise for the baseline full-band system and the Full Combination system (combination by product rule).

These results show an important improvement to the first set of multiple time scale features, and especially to the way the difference features are usually used, i.e. by concatenating them into one feature stream which is then used as the input to one recognizer. In HTK difference features can be treated as separate streams, but as the likelihoods from each stream are simply multiplied, this is – under the usual assumption of diagonal covariance – also equivalent to simple concatenation. The use of the difference features is especially appealing as these features, firstly, are more independent than the variable window size features and, secondly, only need to be extracted once which speeds up calculation time.

5 Conclusion

In this paper we employed the recently introduced "full combination" approach, which was originally developed for multi-band processing, in a multi-stream framework. Multi-stream processing benefits from the fact that different recognizers usually make different errors and thus potentially complement each other when combined. This effect depends on the diversity of the input feature streams.

Using streams with features extracted from different window sizes presents one way of establishing complementary information streams. These features though can not really be regarded as independent.

Another good source of more independent features capturing complementary information from multiple time scales are the time-difference features. These features are widely known and usually used in simple feature concatenation with the instantaneous feature components.

In this article, both sets of features were used as multiple time scale feature streams. The difference features were moreover consistently employed in the "full combination" scheme. This means that all possible combinations of feature streams have to be set up, and then the posterior probabilities from each stream must be combined. We found that in the present case, the product rule resulted in improved robustness to wide-band noise, while the sum rule was not performing as well. This still has to be investigated further. Using the difference features even gave significant improvement in both clean and noisy speech.

Combination of more divers feature streams extracted from different time scales as for example PLP, MSG (Modulation-Filtered Spectrogram) and TRAP (Temporal Pattern) features as done in [12] also led to significant performance improvement, and should also be considered in the framework of full combination.

Acknowledgments:

This work is supported by the Swiss Federal Office for Education and Science (FOES) through the European TMR SPHEAR (SPeech, HEAring and Recognition) and RESPITE (REcognition of Speech by Partial Information TEchniques) projects.

References

[1] H. Bourlard and N. Morgan. *Connectionist Speech Recognition. A Hybrid Approach.* Kluwer Academic Publishers, 101 Philip Drive, Assinippi Park, Norwell, Massachusetts 02061 USA, 1994.

[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] 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.

[5] 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.

[6] A. Hagen and A. Morris. From multi-band full combination to multi-stream full combination processing in robust ASR. *ISCA ITRW ASR2000*, (to appear), 2000.

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

[8] 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.

[9] P. McCourt, S. Vaseghi, and N. Harte. Multi-resolution cepstral features for phoneme recognition across speech sub-bands. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1:557–560, 1998.

[10] 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.

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

[12] A. Sharma, D. Eliis, S. Kajarekar, P. Jain, and H. Hermansky. Feature extraction using non-linear transformation for robust speech recognition on the aurora database, 2000.

[13] 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.

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

[15] S.-L. Wu, B.E.D. Kingsbury, N. Morgan, and S. Greenberg. Performance improvements through combining phone- and syllable-scale information in automatic speech recognition. *Int. Conf. on Spoken Language Processing*, 2:459–462, 1998.