IDIAP RESEARCH REPORT

Entropy Based Combination of Tandem Representations for Noise Robust ASR

Shajith Ikbal [a,b] Hemant Misra [a,b]
Sunil Sivadas [a,c] Hynek Hermansky [a]
Hervé Bourlard [a,b]

IDIAP–RR 04-19

April 27, 2004

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

[a] IDIAP Research Institute, Martigny, Switzerland.
[b] Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.
[c] Oregon Graduate Institute (OGI), Portland, Oregon, USA.

Entropy Based Combination of Tandem Representations for Noise Robust ASR

Shajith Ikbal Hemant Misra Sunil Sivadas Hynek Hermansky
Hervé Bourlard

April 27, 2004

Abstract. In this paper, we present an entropy based method to combine tandem representations of the recently proposed Phase AutoCorrelation (PAC) based features and Mel-Frequency Cepstral Coefficients (MFCC) features. PAC based features, derived from a nonlinear transformation of autocorrelation coefficients and shown to be noise robust, improve their robustness to additive noise in their tandem representation. On the other hand, MFCC features in their tandem representation show a significant improvement in recognition performance on clean speech. An entropy based combination method investigated in this paper adaptively gives a higher weighting to the representation of MFCC features in clean speech and to the representation of PAC based features in noisy speech, thus yielding a robust recognition performance in all conditions.

Acknowledgements: The authors thank the Swiss National Science Foundation for the support of their work through grant MULTI: FN 2000-068231.02/1 and through National Center of Competence in Research (NCCR) on "Interactive Multimodel Information Management (IM2)", as well as DARPA for supporting through the EARS (Effective, Affordable, Reusable Speech-to-Text).

1 Introduction

Methods to improve the noise robustness of speech recognition systems often result in degradation of recognition performance for clean speech. Phase AutoCorrelation (PAC) [1, 2] based features, showing noticeable improvement in noise robustness, compared to the standard features such as Mel-Frequency Cepstral Coefficients (MFCC), also suffer from this draw-back. In [3] an attempt has been made to alleviate this problem by using the PAC based features along with regular speech features in a multi-stream framework. The entropy-based posterior probability combination method in the hybrid ANN-HMM system [4] improved recognition performance in all conditions.

However, the recently proposed speech recognition technique, the tandem technique [5, 6], which combines the two major approaches, namely the hybrid HMM approach and GMM-HMM approach, improves recognition performance in both the clean as well as in the noisy conditions. In tandem approach, transformations of the outputs of a Multi-Layer Perceptron (MLP), which is trained discriminatively with the regular features to estimate the subword posterior probabilities, are used as features to the GMM-HMM systems. As will be shown in the later sections of the paper, the use of PAC based features in tandem system improve their robustness to additive noise while tandem with MFCC features improves the recognition performance in clean speech.

In this paper, we propose an entropy based algorithm to combine the tandem representations of PAC based features and MFCC features. Such a combined representation is expected to perform well in all the conditions, if the combination technique can automatically give higher weighting to MFCCs in clean speech and to PAC features in noisy speech. This indeed happens in experiments on OGI Numbers95 database with artificially added factory noise from Noisex92 database.

In the next section, we first give a small introduction of PAC based features and then discuss their noise robustness. In Section 3, we give a short explanation of tandem representation of feature vectors. In Section 4, we explain the entropy based combination method for tandem features and discuss the results.

2 PAC based features

Traditional features used for speech recognition, typically derived from a power spectrum, show excessive sensitivity to additive noise and their use generally results in degradation of performance in noisy conditions. This is because the autocorrelation coefficients (that are time domain Fourier equivalent of the power spectrum) are highly sensitive to the noise. In an attempt to make the correlation coefficients less sensitive to external noise, a new measure of correlation called Phase AutoCorrelation (PAC) [1, 2] has been introduced. PAC uses angle between the time delayed speech vectors as a measure of correlation instead of the dot product as used in traditional autocorrelation. If $\mathbf{s}$ represents a speech frame given by,

$$\mathbf{s} = \{s[0], s[1], ..., s[N-1]\}$$

where N is the frame length, and

$$\mathbf{x}_0 = \{s[0], s[1], ..., s[N-1]\}$$

$$\mathbf{x}_k = \{s[k], ..., s[N-1], s[0], ..., s[k-1]\}$$

then the autocorrelation coefficients, from which traditional features are extracted, is computed using dot product given by,

$$R[k] = \mathbf{x}_0^T \mathbf{x}_k \tag{1}$$

Alternatively,

$$R[k] = \|\mathbf{x}\|^2 \cos(\theta_k) \tag{2}$$

where $\|\mathbf{x}\|^2$ represents the energy of the frame and θ_k represents the angle between the vectors $\mathbf{x_0}$ and $\mathbf{x_k}$ in N dimensional space.

PAC coefficients, $P[k]$, are derived from the autocorrelation coefficients, $R[k]$, using equation,

$$P[k] = \theta_k = \cos^{-1}\left(\frac{R[k]}{\|\mathbf{x}\|^2}\right) \tag{3}$$

It is clear from the above equation that the computation of PAC coefficients involve two additional operations, namely 1) the energy normalization and 2) the inverse cosine. These two operations effectively convert the dot product of the speech vectors, as done during the computation of the autocorrelation coefficients, into angle between the vectors in N dimensional space. The Fourier equivalent of PAC coefficients in frequency domain is called PAC spectrum. Similarly to the features extracted from the regular spectrum, a class of features called PAC based features can be extracted from the PAC spectrum. Mel Frequency Cepstral Coefficients extracted from PAC spectrum gives PAC-MFCC.

Experimental results given in [1, 2] show that the PAC-MFCC is robust to noise than the MFCC derived from regular spectrum. However, their performance is inferior in clean speech.

A comparison of recognition performances of the PAC-MFCC and MFCC is given in Table 1. Experimental setup[1] for these experiments is as follows: Experiments were conducted on both clean and additive noise conditions. The database used for these experiments is the OGI Numbers95 connected digit telephone speech database [10], having a lexicon size of 30 words, and 27 different phonemes. For additive noise experiments, factory noise from Noisex92 database [11] has been added with Numbers95 database at noise levels of 12 dB, 6 dB and 0 dB Signal-to-Noise Ratio (SNR). The speech recognition system is a Hidden Markov Model (HMM) system consisting of 80 triphones, 3 left-to-right states per triphone, and 12 mixture Gaussian Mixture Model (GMM) to estimate emission probability within each state. HMMs are trained using the HTK package. The PAC-MFCC and MFCC used for the experiments are of dimension 39, consisting of 13 static coefficients, 13 delta coefficients, and 13 delta-delta coefficients.

	% WER for SNR			
Feature	clean	12 dB	6 dB	0 dB
PAC-MFCC	12.2	14.1	24.5	49.2
MFCC	5.6	15.8	38.3	83.3

Table 1: Comparison of the speech recognition performances of PAC-MFCC, and MFCC in a GMM-HMM system for clean speech and noisy speech with additive factory noise levels of 12 dB, 6 dB, 0 dB SNRs. WER stands for Word Error Rate.

From the table it is clear that on the clean speech, the PAC-MFCC is inferior to MFCC, while it is comparatively robust in noisy speech. In the next section, we give a brief review of tandem representation of the features, where the recognition performance improves significantly in all conditions.

3 Tandem Acoustic Representations

Introduced as a combination of two approaches for speech recognition, 1) the hybrid ANN-HMM approach and 2) the GMM-HMM approach, the tandem approach uses the transformed outputs of a Multi-Layer Perceptron (MLP) classifier as the input features to GMMs in conventional speech

[1] The same experimental set up is used for all the experiments reported in this paper.

recognizer [5]. We call these transformed outputs of the MLP the *tandem representation* of the input features. Figure 1 shows the tandem system. The transformations performed to the MLP outputs are 1) to modify the skewness of the probability distributions through the logarithmic transformation of the posterior probability distributions of the softmax outputs (or directly taking the pre-nonlinearity outputs of the MLP), and 2) decorrelate the features through the Karhunen-Loeve (KL) transform, derived from the training data. Tandem system has been shown to perform significantly better than both the hybrid HMM and GMM-HMM approaches. Interestingly, performance in noisy conditions also improves with tandem systems.

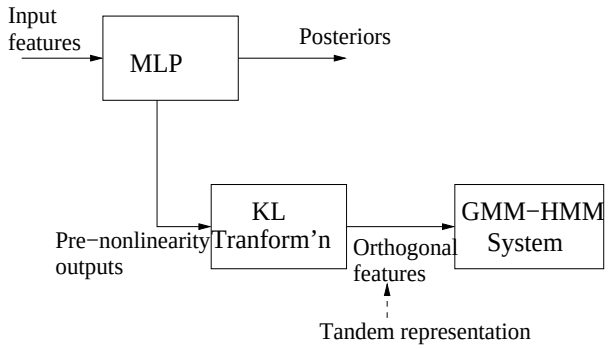

Figure 1: Illustration of a tandem system. Transformed pre-nonlinearity output of the MLP constitute the tandem representation of the input feature.

PAC based features stands in as an interesting candidate for the tandem system. Table 2 gives a performance comparison of tandem representations of the PAC-MFCC and MFCC. Tandem representations are extracted from a MLP that takes 9 frames of contextual input, and has 500 hidden units and 27 output units, corresponding to the number of context-independent phones.

	%WER for SNR			
Feature	clean	12 dB	6 dB	0 dB
T-PAC-MFCC	10.0	14.2	22.8	44.3
T-MFCC	4.7	12.9	25.8	52.4

Table 2: Comparison of the speech recognition performances of tandem representations of PAC-MFCC (T-PAC-MFCC) and MFCC (T-MFCC) for clean speech and noisy speech with additive factory noise levels of 12 dB, 6 dB, and 0 dB SNRs. WER stands for Word Error Rate.

From the table it is clear that the PAC-MFCC is more robust to additive noise in its tandem representation. However, its performance in clean speech is still inferior to that of the MFCC. Hence an appropriate combination of these two representations would yield a robust performance all the conditions. An entropy based combination method yielding a reasonably robust performance in all conditions is explained in the next section.

4 Entropy Based Combination Of Tandem Representations

Entropy of the posterior probability distribution estimated at the output of the MLP serves as a reasonable estimate for the reliability of the estimates. This fact has been proposed in [7] and utilized

and verified in [8], where weights based on entropy values have been assigned to different feature streams in a multi-stream frame work for combining their posterior probabilities. The effectiveness of the entropy based combination of posterior probabilities has been further demonstrated in [3].

In this paper, we propose the use of a similar method to combine the tandem representations of the feature streams. We have seen in Section 3 that the tandem representations are obtained by performing KL transformation on either the pre-nonlinearity outputs of the MLP or logarithm of the posterior probabilities. However, entropy based combination is performed either at the posterior probability level or at logarithmic posterior probability level [9]. In our case, it is done at the logarithmic posterior probability level. Hence the combined tandem representation is obtained by first finding out the probability of the entropy based combination and then performing the logarithm operation followed by the KL transformation. An illustration of entropy based combinations of tandem representations is given in Figure 2. If the MLP parameter set for the two streams, one for PAC-MFCC, $\mathbf{x_1}$, and the

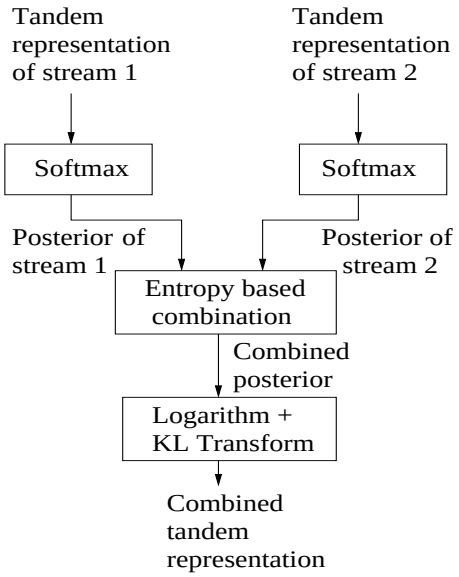

Figure 2: Entropy based combination of tandem representations.

other for MFCC, $\mathbf{x_2}$, are denoted respectively by θ_1 and θ_2, and if posterior probability outputs are denoted respectively by $P(q_k|\mathbf{x_1}, \theta_1)$ and $P(q_k|\mathbf{x_2}, \theta_2)$, then the combined log posterior probability[2] is given by equation

$$\log P(q_k|\mathbf{x_1}, \mathbf{x_2}) = \sum_{i-1}^{2} w_i \log P(q_k|\mathbf{x_i}, \theta_i) \tag{4}$$

where w_1 and w_2 constitute the adaptive weights computed using entropy as follows: If there are K classes (which in our case corresponds to the number of context-independent phonemes), the entropy of the posterior probability output distribution of i^{th} feature stream is computed by equation,

$$h_i = -\sum_{k=1}^{K} P(q_k|\mathbf{x_i}, \theta_i) log_2 P(q_k|\mathbf{x_i}, \theta_i) \tag{5}$$

[2]Log posterior probability combination is chosen over the posterior probability combination, as it gives better recognition performance.

The closer the entropy value is to zero, more reliable the posterior probability distribution is. Hence a normalized inverse value of the entropy is used as an adaptive weighting factor, w_i, computed as:

$$w_i = \frac{1/h_i}{\sum_{j=1}^{2} 1/h_j}$$

Speech recognition performance of the combined tandem representation of PAC-MFCC and MFCC is given in Table 3. From the table it is clear that the combined tandem representation gives a reasonably good recognition performance in all the conditions, when compared to the performance of individual tandem representations as given in Table 2. Interestingly, the combination always performs better than the best performing feature in noisy speech.

% WER for SNR			
clean	12 dB	6 dB	0 dB
5.4	11.7	20.8	44.9

Table 3: Speech recognition performances of entropy based combination of tandem representations of PAC-MFCC and MFCC for clean speech and noisy speech with additive factory noise levels of 12 dB, 6 dB, and 0 dB SNRs. WER stands for Word Error Rate.

5 Conclusion

In this paper, we have presented an entropy based method to combine tandem representations of the PAC-MFCC and MFCC. The main motivation behind this work is the fact that PAC-MFCC improves its robustness in its tandem representation and the tandem representation of the MFCC show a significant performance improvement in clean speech. The results of the experiments performed on OGI Numbers95 database added with factory noise from Noisex92 database show that the entropy based combination does an appropriate combination of the individual tandem representations to yield a reasonably robust recognition performance in all conditions.

References

[1] S. Ikbal, H. Misra, and H. Bourlard, "Phase AutoCorrelation (PAC) derived robust speech features," in *Proc. of ICASSP-03*, Hong Kong, Apr. 2003, II-133–II-136.

[2] S. Ikbal, H. Hermansky, and H. Bourlard, "Nonlinear Spectral Transformations for Robust Speech Recognition," in *Proc. of IEEE ASRU 2003 Workshop*, St. Thomas, Virgin Islands, USA. Nov-Dec, 2003.

[3] S. Ikbal, H. Misra, H. Bourlard, and H. Hermansky, "Phase AutoCorrelation (PAC) Features in Entropy Based Multi-Stream for Robust Speech Recognition," to appear in *Proc. of ICASSP-04*, Montreal, May. 2004.

[4] H. Bourlard, and N. Morgan, "Connectionist Speech Recognition: A Hybrid Approach," *The Kluwer International Series in Engineering and Computer Science*, Kluwer Academic Publishers, Boston, USA, 1993, Vol. 247.

[5] H. Hermansky, D. Ellis, and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems," *in Proc. of ICASSP-00*, Istanbul, June 2000.

[6] D. Ellis, R. Singh, and S. Sivadas, "Tandem Acoustic Modeling in Large-Vocabulary Recognition," *in Proc. of ICASSP-01*, Salt Lake City, Utah, USA, May 2001.

[7] S. Okawa, E. Bocchieri, and A. Potamianos, "Multi-Band Speech Recognition in Noisy Environments," in *Proc. of ICASSP-98*, Seattle, Washington, USA, May 1998.

[8] H. Misra, H. Bourlard, and V. Tyagi, "New Entropy based Combination Rules in HMM/ANN Multi-Stream ASR," in *Proc. of ICASSP-03*, Hong Kong, Apr. 2003, II-741–II-744.

[9] A. Hagen, "Robust Speech Recognition Based on Multi-Stream Processing," *Thesis*, EPFL, Lausanne, Switzerland. Dec. 2001.

[10] R. Cole, M. Noel, T. Lander, and T. Durham, "New telephone speech corpora at CSLU," *in Proceedings of European Conference on Speech Communication and Technology*, 1995, vol. 1, pp. 821–824.

[11] A. Varga, H. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the affect of additive noise on automatic speech recognition," *Technical report*, DRA Speech Research Unit, Malvern, England, 1992.