IDIAP RESEARCH REPORT

Phase AutoCorrelation (PAC) features in Entropy based Multi-Stream for Robust Speech Recognition

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

IDIAP–RR 03-54

October 27, 2003

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] Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP), Martigny, Switzerland.
[b] Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.

IDIAP Research Report 03-54

Phase AutoCorrelation (PAC) features in Entropy based Multi-Stream for Robust Speech Recognition

Shajith Ikbal Hemant Misra Hervé Bourlard Hynek Hermansky

October 27, 2003

Abstract. Methods to improve noise robustness of speech recognition systems often result in degradation of recognition performance for clean speech. Recently proposed Phase AutoCorrelation (PAC) [1, 2] based features, showing noticeable improvement in noise robustness, also suffer from this draw back. In this paper, we try to alleviate this problem by using the PAC based features along with regular speech features in a multi-stream framework. The multi-stream system uses entropy of the posterior probability distribution, computed during recognition, as a confidence measure to adaptively combine evidences from different feature streams [3]. Experimental results obtained on OGI Numbers95 database and Noisex92 noise database show that such a system yields best possible recognition performance in all conditions. Actually, the combination always performs better than the best performing stream for all the 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

Traditional features used for speech recognition, typically derived from power spectrum, show excessive sensitivity to external additive noise and generally result in degradation of recognition 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. Several techniques such as spectral subtraction [4] for stationary noise and RASTA processing [5] for slow varying noise, have been developed to handle this sensitivity. Those techniques typically work at the spectral level, trying to alleviate the effect of noise on the spectrum.

Recently, this problem has been addressed at the autocorrelation level, trying to make the correlation coefficients less sensitive to external noise. A new measure of correlation called Phase AutoCorrelation (PAC) [1, 2], that uses angle between the time delayed speech vectors as a measure of correlation instead of the dot product as used in traditional autocorrelation, has been introduced. The motivation behind it is the fact that in the presence of external additive noise, angle gets less affected than the dot product [6]. As a result, PAC and the features derived from it are expected to be less sensitive to external noise than the traditional autocorrelation. This is confirmed by the experimental results reported in [1].

In spite of their improved robustness in noisy conditions, PAC based features in clean speech are inferior to the state-of-the-art features. This performance degradation in clean speech is typically observed in most of the noise robust techniques as they affect the inter-class discriminatory information to some extent. This makes the PAC based features less competitive for use in state-of-the-art speech recognition systems. In this paper, we try to alleviate this problem by using the PAC based features in a multi-stream framework along with the traditional features such as Perceptual Linear Prediction (PLP) cepstrum [7]. An entropy based multi-stream combination system as proposed in [3] is used for this purpose.

In the next section, we first explain the PAC and the PAC based features and then discuss their advantages and disadvantages. In Section 3, we explain the entropy based multi-stream system used to combine the PAC features with the regular features. In Section 4, we explain the experimental set up used to evaluate the system in clean and noisy conditions. In Section 5 we present and discuss the results of the experiments.

2 PAC based features

A short review of the Phase AutoCorrelation (PAC) which was initially introduced in [1] is as follows: If $s[n]$ represents a speech frame given by,

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

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,

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

From the above equation it is clear that compared to the traditional autocorrelation coefficients the computation of PAC coefficients involve two additional operations namely, energy normalization and 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. As angle gets less affected in noise than the dot product, PAC coefficients are more robust to noise than the regular autocorrelation coefficients [6, 2]. The Fourier equivalent of PAC coefficients in frequency domain is called PAC spectrum. Similar 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.

The PAC based features are expected to be more robust to noisy conditions than their spectral counterpart. Experimental results given in [1, 2] indeed confirm this. However, the energy normalization and inverse cosine operations performed to compute PAC coefficients affect to some extent the class discriminative information in the speech signal, and cause degradation of performance in the clean speech. As given in [2] incorporating energy as a separate component in the PAC based features would improve the clean speech performance. But still that is a partial remedial solution as inverse cosine operation also contributes to the degradation. In this paper, we address this problem by using PAC based features along with the traditional features in a multi-stream framework. The next section explains the entropy based full-combination multi-stream system used to perform the combination.

3 Entropy based Multi-Stream Full-Combination

In full combination multi-stream system, several classifiers, assigned one each to all possible combinations of the feature streams, are trained [8, 3]. For example, as shown in Figure 1, if there are 2 feature streams, F_1 and F_2, there are 4 possible combinations, denoted by $C_1 = \{NULL\}$, $C_2 = \{F_1\}$, $C_3 = \{F_2\}$, and $C_4 = \{F_1, F_2\}$. Let the parameter sets of 4 classifiers trained on each of these combinations be denoted by θ_1, θ_2, θ_3, and θ_4, respectively. During recognition, evidences given by each of these classifiers are combined at frame level to get a cumulative evidence. For example, if $P(q_k|C_1, \theta_1)$, $P(q_k|C_2, \theta_2)$, $P(q_k|C_3, \theta_3)$, and $P(q_k|C_4, \theta_4)$ denote the posterior probabilities obtained from the classifiers for C_1, C_2, C_3, and C_4 respectively for k^{th} class, then the combined posterior probability of the system for k^{th} class is given by equation

$$P(q_k|F1, F2) = \sum_{i=1}^{4} w_i P(q_k|C_i, \theta_i) \tag{4}$$

where w_i constitutes the confidence factor for i^{th} classifier. The value of w_i decides how much to rely upon the i^{th} combination to compute the combined posterior probability. The closer that value is to one, the more we rely upon the corresponding combination to compute the combined posterior probability.

Adaptively changing the weights w_i depending upon the reliability of the corresponding combination for each frame would make the system to yield the best possible performance in all kinds of scenarios. Entropy of the posterior probability distribution obtained from a particular classifier serves as a reasonable estimate for the reliability of the corresponding combination [3]. If there are K classes (which in our case corresponds to phonemes), entropy of the posterior probability distribution for i^{th}

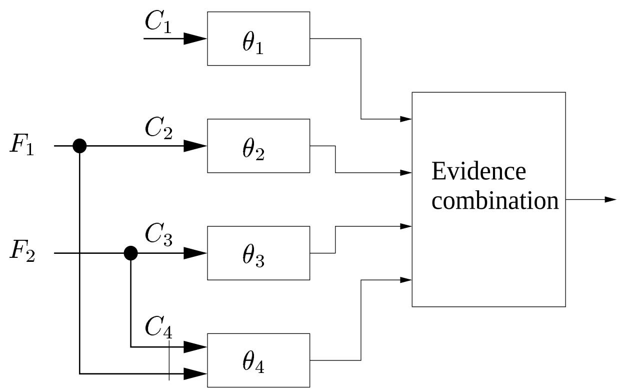

Figure 1: Full-combination multi-stream system.

classifier is computed by equation,

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

The closer the entropy value is to zero, the more reliable the combination is. Hence a normalized inverse value of the entropy is used as an adaptive weighting factor, w_i in (4), for the i^{th} stream, computed as:

$$w_i = \frac{1/h_i}{\sum_{j=1}^{4} 1/h_j}$$

4 Experimental Setup

To illustrate the performance of the proposed combination system, speech recognition experiments were conducted on both clean and additive noise conditions. The database used for these experiments is the OGI Numbers95 connected digits telephone speech database [10], described by a lexicon 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 such as 6dB and 12 dB SNR.

In all the experiments PAC-MFCC has been used as a representative of the PAC based features and PLP cepstrum has been used as a representative of the traditional features. Both PAC-MFCC and PLP cepstrum were of dimension 39, including 13 static coefficients, 13 delta coefficients, and 13 delta-delta coefficients.

The speech recognition system used for the experiments is a hybrid Hidden Markov Model-Artificial Neural Network (hybrid HMM-ANN) system, that uses Multi-Layer Perceptron (MLP) for computing the emission probabilities [9]. Hybrid HMM-ANN is preferred for this study, because it provides a nice framework for performing multi-stream combination. The MLP takes 9 frames of contextual input and has 1800 hidden units, and 27 output units corresponding to the number of mono phones.

5 Results and Discussion

Table 1 compares the individual performance of the PAC-MFCC and the PLP cepstrum with their multi-stream combination performance, for clean as well as noisy speech. Noise experiments were conducted with additive factory noise levels of 12dB SNR and 6dB SNR. From the table it is clear that the multi-stream combination system always picks up the best performing stream in all conditions, i.e., the performance of the PLP cepstrum is achieved in clean speech and the performance of the PAC-MFCC is achieved in noisy speech.

Category	WER for clean %	WER for SNR 12 dB %	WER for SNR 6dB %
PAC-MFCC	13.5	18.0	28.8
PLP cepstrum	10.3	19.0	32.5
Multi-stream	**9.9**	**15.8**	**27.4**

Table 1: Comparison of the speech recognition performances of PAC-MFCC, PLP cepstrum, and their multi-stream combination for clean speech and noisy speech with additive factory noise levels of 12 dB and 6dB SNRs. WER stands for Word Error Rate.

Moreover, interestingly the combination always performs better than the best performing individual streams in all the conditions[1]. This can be attributed to the fact that the PAC-MFCC may be emphasizing certain aspects of the spectrum that is complementary to those emphasized by the PLP cepstrum. Frame level entropy based combination of the feature streams makes better use of this complementary information to yield a better combined performance than the best individual performance.

6 Conclusion

In spite of the improved robustness achieved by PAC based features in noisy conditions, their recognition performance in clean speech is inferior to that of the state-of-the-art features. In this paper, we addressed this problem by combining the PAC based features with PLP cepstrum in a multi-stream framework. The recognition results obtained through entropy based multi-stream combination show that in all conditions the combination yields even better performance than the best performing feature. This can be attributed to the fact that the PAC based features emphasizes certain aspects of the spectrum that is complementary to what has been done by the PLP cepstrum, and entropy based multi-stream full-combination technique makes use of this to yield better performance.

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," to appear in *Proc. of IEEE ASRU 2003 Workshop*, Nov-Dec, 2003.

[3] 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.

[4] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," in *Proc. of IEEE ASSP-27*, Apr.1979, pp. 113-120.

[1] This may raise speculations about the use of more number of parameters in the multi-stream system. But it has been verified through experiments that the individual performance of the streams do not change significantly with the increase in number of parameters.

[5] H. Hermansky, and N. Morgan, "RASTA Processing of Speech," *IEEE Transactions on Speech and Audio Processing*, Oct. 1994, Vol.2, No:4, pp. 578-589.

[6] D. Mansour, and B. H. Juang, "A Family of Distortion Measures based upon Projection Operation for Robust Speech Recognition," *in Proc. of ICASSP-88*, 1988, pp. 36–39.

[7] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis for Speech," *Journal of Acoustic Society of America*, 1990, pp: 1738-1752.

[8] A. C. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Communication*, 2001, vol.34, pp. 25-40.

[9] 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.

[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.