IDIAP RESEARCH REPORT

NONLINEAR SPECTRAL TRANSFORMATIONS FOR ROBUST SPEECH RECOGNITION

Shajith Ikbal [a,b]   Hynek Hermansky [a]

Hervé Bourlard [a,b]

IDIAP–RR 03-36

August 11, 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 03-36

NONLINEAR SPECTRAL TRANSFORMATIONS FOR ROBUST SPEECH RECOGNITION

Shajith Ikbal   Hynek Hermansky   Hervé Bourlard

August 11, 2003

Abstract. Recently, a nonlinear transformation of autocorrelation coefficients named Phase AutoCorrelation (PAC) coefficients has been considered for feature extraction [1]. PAC based features show improved robustness to additive noise as a result of two operations, performed during the computation of PAC, namely energy normalization and inverse cosine transformation. In spite of the improved robustness achieved for noisy speech, these two operations lead to some degradation in recognition performance for clean speech. In this paper, we try to alleviate this problem, first by introducing the energy information back into the PAC based features, and second by studying alternatives to inverse cosine function. Simply appending the frame energy as an additional coefficient in the PAC features has resulted in noticeable improvement in the performance for clean speech. Study of alternatives to inverse cosine transformation leads to a conclusion that linear transformation is the best for clean speech, while nonlinear functions help to improve robustness in noise.

Acknowledgements: The authors thank Swiss National Science Foundation for the support of their work through grant FN 2001-061325.00/1 and through National Center of Competence in Research (NCCR) on 'Interactive Multimodel Information Management (IM2)'.

1 Introduction

Traditional features used for speech recognition, typically derived from power spectrum, show excessive sensitivity to additive noise present in the signal and generally result in poor performance under noisy conditions. This is because the autocorrelation coefficients, that are the time domain Fourier equivalent of the power spectrum, are highly sensitive to the noise. Several techniques, such as spectral subtraction [2] for stationary noise and RASTA processing [3] for slow varying noise, have been developed to handle this sensitivity. Those techniques typically work at the spectral level, trying to get rid off the effect of noise on the spectrum.

Recently, this problem has been addressed at the autocorrelation level, trying to make the correlation coefficient less sensitive to the external noise, so that the power spectrum derived from it and the features derived further would be more robust. A new measure of autocorrelation called Phase AutoCorrelation (PAC) [1] that uses angle between the time delayed speech vectors as a measure of correlation instead of the dot product as used in the 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 [4]. As a result, PAC and the features derived from it are expected to be less sensitive to external noise than the traditional autocorrelation. Experimental results demonstrate that this is indeed the case [1].

The improvements in speech recognition performance while using PAC derived features in noise is achieved as a result of two operations performed during the computation of PAC namely, energy normalization followed by inverse cosine transformation. These two operations effectively convert the dot product of speech vectors into angle between the vectors. Energy normalization removes out the variation in energy that results from the presence of the noise and inverse cosine enhances a few aspects of the spectrum such as spectral peaks, that are more robust to noise.

Although PAC derived features show significant performance improvement in noise, they have a major drawback that their performance in clean condition is noticeably lower when compared with state of the art features. Both the energy normalization and inverse cosine operations contributes to this degradation. In this paper, we further analyze the PAC for both clean and noisy conditions, and try to improve their recognition performance for the clean speech. We expect the performance to improve if the energy information is introduced in the PAC derived features. In fact, improvement in recognition performance has been achieved by using energy as an additional coefficient with the PAC derived features. As the inverse cosine may not be the optimal nonlinear function to transform the energy normalized autocorrelation coefficients, we have also considered a few alternatives to it.

In the next section we analyze the PAC, to illustrate its robustness in noisy conditions. In section 3 we explain the experimental setup and give performance of PAC for clean as well as noisy conditions. We end that section with a discussion on drawbacks of PAC for clean speech. In section 4 we study the effects of energy normalization on clean speech and show through experimental results that introducing energy information as an additional coefficient in the PAC derived feature results in performance improvement for clean speech. Section 5 studies the effects of nonlinear transformation and discusses alternatives to inverse cosine function.

2 PAC - Analysis

If $s[n]$ represents a speech frame given by,

$$s[n] = \{s_0[n], s_1[n], ..., s_t[n], ..., s_{T-1}[n]\}$$

where T 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 equation for autocorrelation coefficients, from which traditional features are extracted, is given as follows:

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

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.

Phase AutoCorrelation (PAC) coefficients, $P[k]$, are derived from the autocorrelation coefficients, $R[k]$, using equation [1],

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

From the above equation it can be seen that the computation of PAC coefficients involve two operations namely,

1. Energy Normalization, to compute energy normalized autocorrelation.
2. Cosine Inverse, to nonlinearly transform the energy normalized autocorrelation coefficients into PAC coefficients.

These two operations effectively convert the dot product of speech vectors, that is done during the computation of the autocorrelation coefficients, into angle between the vectors.

In the presence of an additive noise $r[n]$ the resultant speech frame, $s^n[n] = s[n] + r[n]$, results in vectors $\mathbf{x}_0^n$ and $\mathbf{x}_k^n$. Now the dot product of these two vectors constitute the autocorrelation coefficient $R[k]$ of the noise corrupted speech, and angle between them constitute the PAC coefficient $P[k]$. As can be seen from 2-D illustration given in Figure 1, both the angle and the energy undergo change in the presence of noise. $R[k]$ depends both on the frame energy and angle between the vectors, where as $P[k]$ depends just on the angle between the vectors. Consequently, $P[k]$ are expected to be less susceptible to the external noise than the $R[k]$.

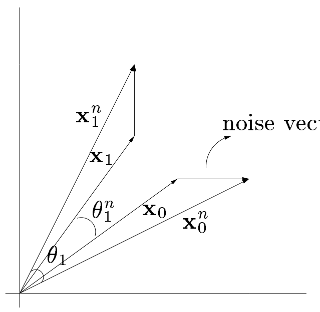

Figure 1: 2-D illustration of how additive noise affects the energy of the speech frame and angle between the time delayed speech vectors.

Ideally, if we go by the argument given above, even the use of energy normalizedautocorrelation coefficients should result in performance improvement in the presence of noise. i.e., the use of $\cos(\theta_k)$ as correlation coefficientsshould result in noise robustness, since it also depends just on the θ_k.This is indeed the case and experimental results given in the later section of this paper confirm this. But the inverse cosine performed to compute the angle θ_k also turns out to be an important operation, since better performance improvements are achieved in noise while using θ_k as correlation coefficients. The nonlinear transformation of the energy normalized autocorrelation coefficients into PAC coefficients using inverse cosine function enhances the peaks in the PAC spectrum. This is visually illustrated in figures 2 and 3. Figure 2 shows the regular spectrum and Figure 3 the PAC spectrum. The enhancement of PAC spectral peaks makes the PAC features more robust to noise, as spectral peaks are less sensitive to the noise.

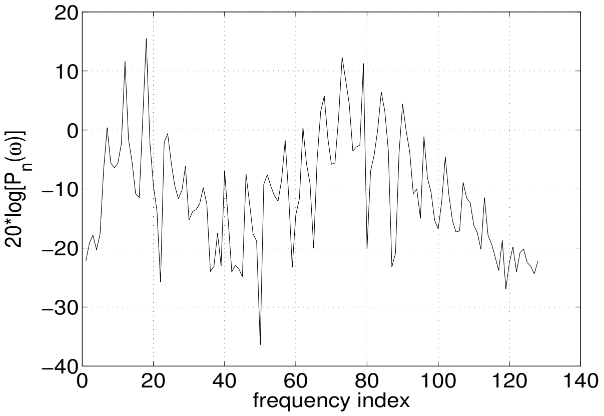

Figure 2: Logarithm of energy normalized power spectrum for a frame of phoneme 'ih'.

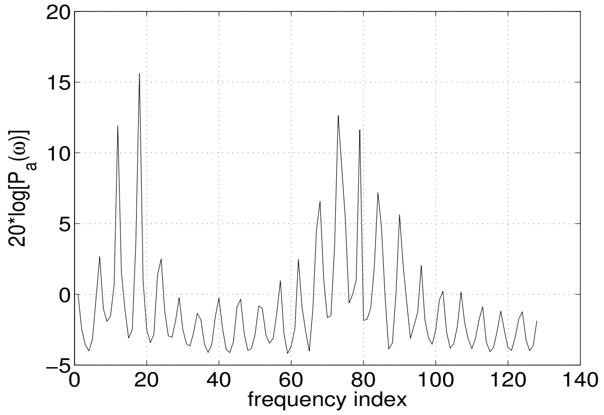

Figure 3: Logarithm of PAC power spectrum for a frame of phoneme 'ih'.

The explanation for the enhancement of the PAC spectral peak is as follows: Figure 4 shows a plot of $\cos^{-1}(x)$ for x values in the range -1 to $+1$. The values of the function are transformed according to the equation given in the y-axis of the figure to fit in range -1 to $+1$. From this figure it can be seen that the slope of the curve becomes larger as the magnitude of the x value becomes larger. This means variations in the values of x near $+/-1$ are magnified in the y-axis. Typically, a initial few coefficients of autocorrelation are high in magnitude. Hence, any variation across these coefficients is enhanced. These initial few coefficients of autocorrelation decide the shape of the spectral envelope, as they constitute the slow varying part in the corresponding spectral domain. Since the variation across these coefficients are enhanced, the shape of the spectral envelope, and hence the spectral peaks, are better enhanced in the PAC spectrum. On the other hand, when the autocorrelation coefficients are close to zero, which is typically the case in noisy vectors, the inverse cosine do not enhance the variation across them.

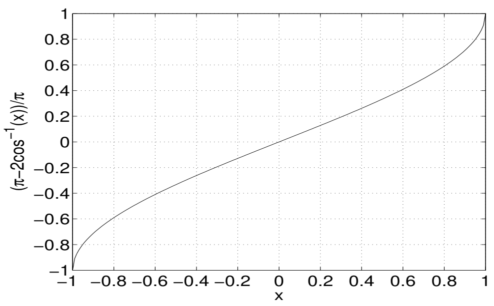

Figure 4: Cosine inverse function.

The above fact is further illustrated by Figure 5, showing the distribution of the PAC spectral power against the energy normalizedspectral power for an utterance. Each point in the figure corresponds to a particular frequency, with their x and y coordinates corresponding to spectral powers of energy normalized and PAC spectra respectively. From the figure it is clear that as the power values of the energy normalized spectrum gets larger, the relationship between energy normalized and PAC spectra becomes linear. Where as for the lower power values, the variations in regular spectrum is diminished in the corresponding PAC spectrum.

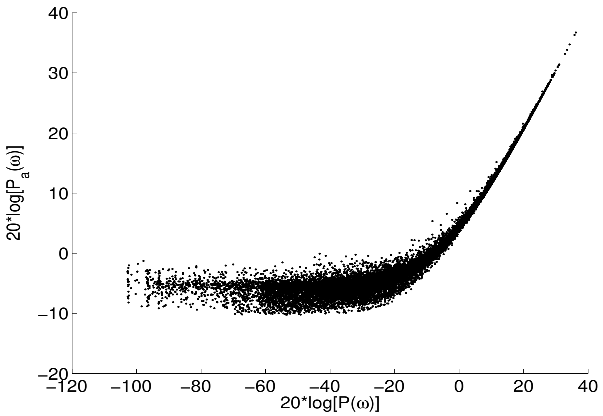

Figure 5: Distribution of the energy normalized spectral power against the PAC spectral power for an utterance.

Noise robustness of the PAC spectrum is illustrated by Figures 6 and 7. Figure 6 shows a plot of Euclidean distance between spectra of clean speech and spectra of speech corrupted by additive noise at 6dB SNR, over an utterance. Figure 7 shows similar plot for the PAC spectra. In order to have a fair comparison, the magnitudes of both the spectra are normalized to same range of values by mean removal and variance normalization. It is clear from the figure that the PAC spectra of noisy speech is closer to the PAC spectra of the clean speech, when compared to the regular spectra.

3 PAC - Performance

Experimental results shown in Figure 8 confirm the noise robustness of the PAC derived features. These experiments are conducted with regular MFCC and PAC MFCC features. These features are of

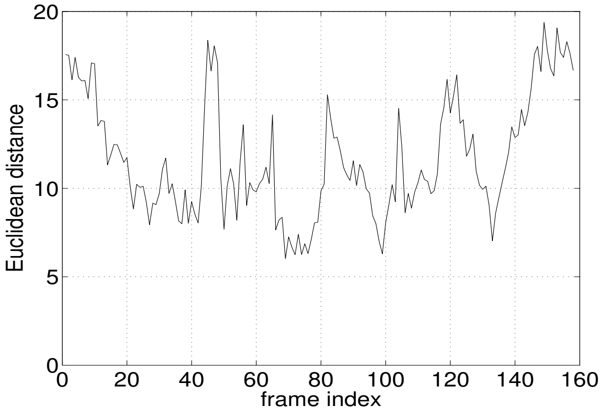

Figure 6: Euclidean distance between regular spectra of clean speech and 6 dB additive noise corrupted speech for an utterance.

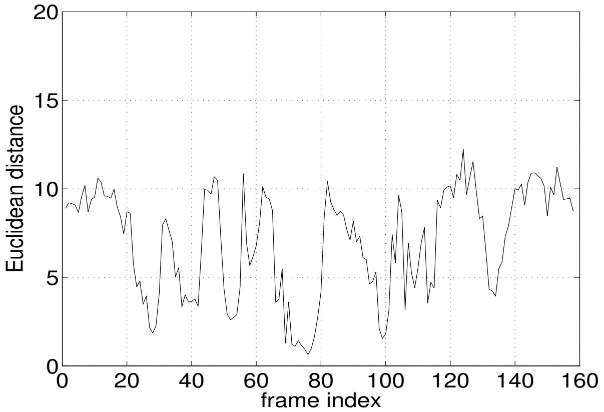

Figure 7: Euclidean distance between PAC spectra of clean speech and 6 dB additive noise corrupted speech for an utterance.

dimension 39, including 13 static coefficients, 13 delta coefficients, and 13 delta-delta coefficients. The Hidden Markov Model (HMM) system used for the experiment consists 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 HTK. Database used for the experiment is OGI Numbers95 connected digits telephone database [5], described by a lexicon of 30 words, and 80 different triphone. For additive noise, factory noise from Noisex91 database [6] has been used[1]. From Figure 8, it is clear that in the presence of the noise the performance of the PAC MFCC is significantly better as compared to the regular MFCC features. In [1] it is also shown that PAC MFCC was yielding performances comparable to RASTA-PLP which is a well known approach for noise robust speech feature extraction.

Though PAC derived features show better noise robustness, they have a major drawback that their performance in clean speech is noticeably lower than that of the state of the art features. Table 1 gives performance comparison of the PAC MFCC against the regular MFCC for clean speech.

The energy normalization and the inverse cosine transformation performed during the computation of the PAC cause performance degradation in clean speech, though they help to improve their noise

[1]All the experiments reported in this paper are conducted with similar settings.

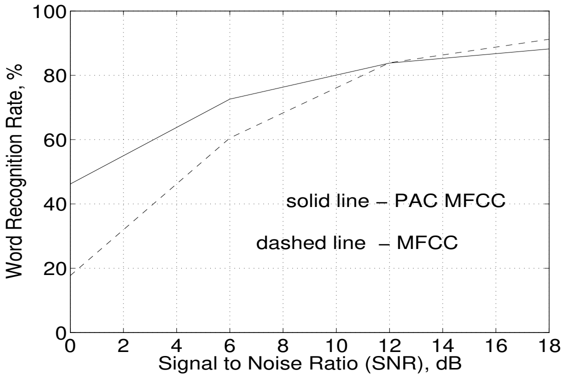

Figure 8: Performance comparison of PAC MFCC with regular MFCC for additive Factory noise.

Feature	Word Recognition Rate, % acc.
MFCC	93.7
PAC MFCC	88.7

Table 1: Comparison of the speech recognition performances for the clean speech.

robustness. In the next two sections we study the effect of energy normalization and inverse cosine transformation on the PAC spectrum, and try to alleviate the performance degradation of PAC in clean speech.

4 Energy Normalization

Energy normalization performed during the computation of PAC is important from two aspects. First, the inverse cosine transformation requires the autocorrelation values to be in the range $+/-1$. Second, energy normalization also contributes to the robustness of the feature vector in the presence of noise, as energy changes with the addition of the noise. This is evident from Figure 9, which shows performance comparison of energy normalized MFCC against regular MFCC for various additive factory noise levels.

However, energy normalization degrades the performance in clean speech as energyalso constitute an important source of information for recognition of clean speech. This is illustrated by the performance comparison given in the first two rows of Table 2 for energy normalized MFCC and regular MFCC. Hence to improve the performance of PAC derived features, for clean speech, energy information should be incorporated into the feature. Row 3 of Table 2 show performance of the PAC MFCC when energy is appended as an additional coefficient. Comparing this with the performance of PAC MFCC given in Table 1, it can be seen that PAC MFCC gains a significant improvement of 3.6% for clean speech just by incorporating the energy information back into the feature.

Figure 10 shows performance comparison of energy appended PAC MFCC and regular PAC MFCC for various noise levels of additive factory noise. From the figure it is clear that energy appended PAC MFCC performs better than PAC MFCC even in noise. This is interesting because in case of MFCC where energy information is already present the performance degrades drastically in noise. The improvements in present case can be attributed to the fact that energy information is completely decoupled from the other coefficients of the PAC MFCC and introduced as a single coefficient in the

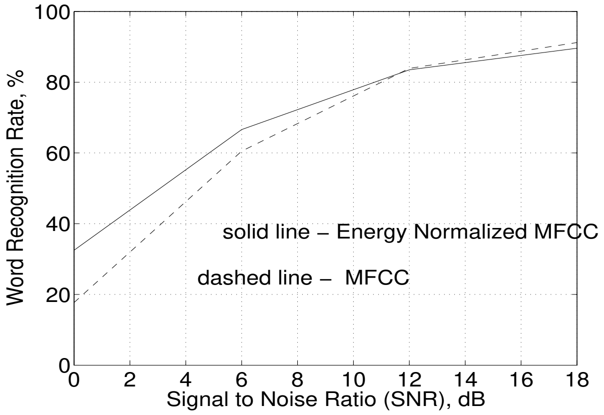

Figure 9: Performance comparison of energy normalized MFCC with regular MFCC for additive Factory noise.

Feature	Word Recognition Rate, % acc.
MFCC	93.7
Energy normalized MFCC	91.7
Energy appended PAC MFCC	92.3

Table 2: Comparison of the speech recognition performances for the clean speech.

feature. A behavior similar to this can be found in [7] where performance improvement is achieved while energy is used as an auxiliary variable.

5 Inverse Cosine

As explained in Section 2, inverse cosine function enhances the PAC spectral peaks. This results in improved noise robustness as the spectral peaks are less sensitive to noise. However, unfortunately in the clean speech, this results in degradation of the recognition performance. This is evident from the recognition results given in Table 2. The regular MFCC features and the energy appended PAC MFCC features carry the same information except for the fact that inverse cosine operation is performed additionally in the later case. This causes 1.4% drop in recognition rate for clean speech. This raises questions about optimality of the inverse cosine function for PAC computation. In this section we study alternatives to inverse cosine function.

Figure 11 shows a few examples of alternate functions we consider. In the figure, functions plotted by solid lines are linear and inverse cosine. Those plotted by dotted and dashed lines are alternate functions that yet have the shape of the inverse cosine but differ in the magnitudes. The family of dashed curves are specified by the values of variable f from -1 to $+1$. When $f = -1$ the function is linear and when $f = +1$ function is inverse cosine. All the functions in between are specified by f values between -1 to $+1$.

The function plotted with dotted line looks interesting for our current investigation because its slope is larger than inverse cosine for larger values of x. Hence, according to the argument in Section 2, this function should enhance the spectral peaks even better. Unfortunately, this function do not yield better performance both for clean and noisy speech. The recognition performances obtained are 87.0% for clean speech and 71.8% for 6dB noise corrupted speech. This turns our attention to the set

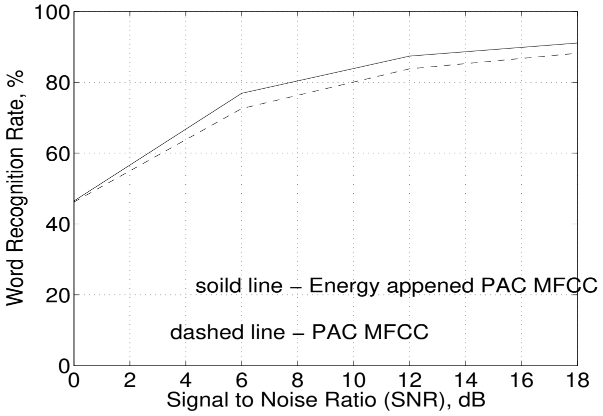

Figure 10: Performance comparison of energy appended PAC MFCC with PAC MFCC for additive Factory noise.

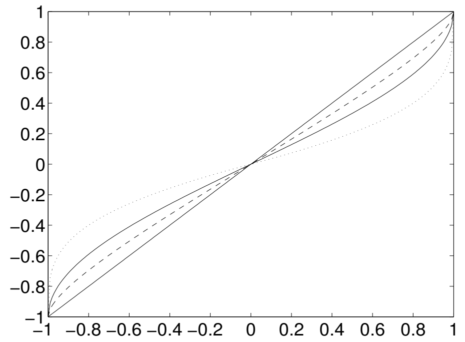

Figure 11: Alternative nonlinear functions to the inverse cosine.

of functions shown by the dashed line, because they cause milder modifications during transformation than the inverse cosine. Figure 12 shows plots of recognition performance for the clean speech and the 6dB noise corrupted speech, for various values of f. For clean speech, with highest recognition performance for $f = -1.0$, which corresponds to energy normalized MFCC, the performance drops down gradually with increasing f and reaches a low value when $f = 1.0$, which corresponds to PAC MFCC. This leads to a conclusion that all the nonlinear transformations hurt the recognition performance of clean speech. The milder the nonlinearity, lesser the degradation. But the nonlinear transformation certainly helps in the noisy speech. Even for the lower values of f, the recognition performance is reasonably better than the linear transformation. The performance curves also show that inverse cosine is not the optimal nonlinear function.

6 Conclusion

In this paper, we have analyzed the two operations performed during the computation of the PAC coefficients, namely energy normalization and inverse cosine transformation. In spite of the improved robustness in noise, these operations cause degradation of recognition performance in clean speech. As a remedial solution, we have tried introducing the energy information in the PAC based features. Introducing energy as an additional coefficient in the PAC based features has resulted in noticeable improvement in the recognition rate for clean as well as noisy speech. Questioning the optimality

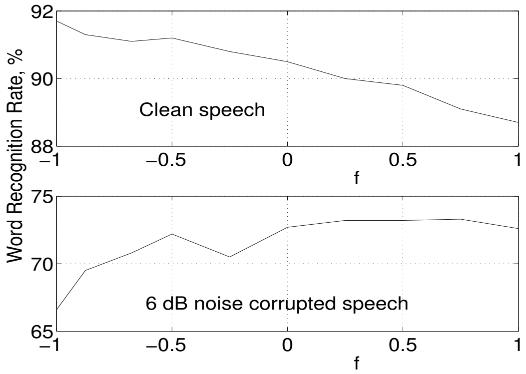

Figure 12: Recognition performance of the alternative nonlinear functions.

of inverse cosine transformation we have studied the suitability of a few other nonlinear functions, that are yet close to inverse cosine function. For clean speech best performance is still achieved with linear transformation, i.e., energy normalized MFCC, while the nonlinear function always degrades the performance for clean speech. However, for noisy speech, nonlinear functions help to improve the robustness.

These results point to future work where PAC-like feature derived using linear, inverse cosine, and other nonlinear functions, could be used as features in multi-stream frame work [8]. As inverse cosine do not turn out to be the optimal nonlinear function, suitability of other nonlinear functions that might enhance the speech specific information present in the speech signal would be worth exploring.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] T. A. Stephenson, M. Mathew, and H. Bourlard, "Speech Recognition with Auxiliary Information," accepted for publication in *IEEE Transactions on Speech and Audio Processing*,

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