# PHASE AUTOCORRELATION (PAC) DERIVED ROBUST SPEECH FEATURES

*Shajith Ikbal*[*]*, Hemant Misra, Hervé Bourlard*[*]

IDIAP, Martigny, Switzerland
{ikbal, misra, bourlard}@idiap.ch

**ABSTRACT**

In this paper, we introduce a new class of noise robust acoustic features derived from a new measure of autocorrelation, and explicitly exploiting the phase variation of the speech signal frame over time. This family of features, referred to as "Phase AutoCorrelation" (PAC) features, include PAC spectrum and PAC MFCC, among others. In regular autocorrelation based features, the correlation between two signal segments (signal vectors), separated by a particular time interval $k$, is calculated as a dot product of these two vectors. In our proposed PAC approach, the angle between the two vectors is used as a measure of correlation. Since dot product is usually more affected by noise than the angle, it is expected that PAC-features will be more robust to noise. This is indeed significantly confirmed by the experimental results presented in this paper. The experiments were conducted on the Numbers 95 database, on which "stationary" (car) and "non-stationary" (factory) Noisex 92 noises were added with varying SNR. In most of the cases, without any specific tuning, PAC-MFCC features perform better.

## 1. INTRODUCTION

Traditional features used for speech recognition are typically extracted from the magnitude spectrum [1, 2] of the speech signal, estimated by Discrete Fourier Transform (DFT) of the autocorrelation coefficients [3, 4]. Unfortunately, these features are extremely sensitive to external noise added to the signal as the basic autocorrelation coefficients, from which they are extracted, are highly sensitive to external noise. This generally results in poor performance of the speech recognition systems in presence of noise.

Several techniques have been developed so far to cope with the sensitivity of the feature vectors to external noise. These techniques typically work at the spectral level of the feature extraction trying to get rid of the effect of the external noise on the spectrum. One early method called Spectral subtraction [5] gets an estimate of noise power spectrum from the non-speech intervals of the signal and subtracts it from the power spectra of the overall speech signal. This technique can be employed for the cases where the noise characteristic is stationary. However, in case of non-stationary noise, this technique may result in the removal of significant speech information and hence may result in poor recognition performance. A relatively new technique called RASTA processing [6], which has been shown to be quite successful for noise robust speech recognition, tries to remove those noise components in the power spectrum whose temporal properties are quite different from that of the speech component. Band-pass filters, with bandwidths equal to the bandwidths of the temporal characteristic of the speech component is applied to each frequency band of the spectrum, to get rid of the noise components.

In this paper, we introduce a class of noise robust speech features called Phase AutoCorrelation (PAC) derived features. These features are derived from a new measure of autocorrelation, we propose in this paper, called Phase AutoCorrelation. Regular autocorrelation coefficients, which are computed by performing dot product between signal vectors separated by a particular time interval, are extremely sensitive to the external noise. In phase autocorrelation, angle between the signal vectors is used as the measure of correlation, instead of the dot product. The angle is less sensitive to external noise, as compared to the dot product. As a result of this, we expect the PAC derived features to be more robust to noise as compared to the traditional features, which are derived from the regular autocorrelation.

In the next section, we first explain the draw-backs of the traditional autocorrelation in the presence of external noise and then propose a new measure of autocorrelation called Phase AutoCorrelation. We end that section by introducing the PAC derived features. In Section 3, we explain the experimental setup used to evaluate the PAC derived features under noisy conditions. In Section 4, we present and discuss the results of the experiments.

## 2. PHASE AUTOCORRELATION (PAC)

### 2.1. Autocorrelation

Feature extraction block in a typical speech recognition system divides the speech signal $s[n]$ into a sequence of frames

[*] Also with EPFL, Lausanne, Switzerland.

given by,

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

where $T$ is the total number of frames and $s_t[n]$ is given by,

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

$N$ is the frame length and $K$ the frame shift. Feature vectors are extracted from each of these frames assuming that the characteristic of the signal within a single frame is stationary. Features extracted from the frames are typically some or other form of the magnitude spectrum. The magnitude spectrum is obtained by first performing the Discrete Fourier Transform (DFT) of the frame samples and then taking the magnitude of the resulting coefficients for various frequencies. DFT assumes each frame $s_t[n]$ to be part of a periodic signal $\tilde{s}_t[n]$ [3] defined as:

$$\tilde{s}_t[n] = \sum_{k=-\infty}^{+\infty} s_t[n+kN]$$

As well known, the squared magnitude spectrum is the DFT of the autocorrelation $R[n]$ of the periodic sequence $\tilde{s}_t[n]$ over the length equal to the length of the frame. The equation for autocorrelation is given as follows:

$$R[k] = \sum_{n=0}^{N-1} \tilde{s}_t[n]\tilde{s}_t[n+k], \quad k = 0, 1, ..., N-1. \quad (1)$$

The above operation of autocorrelation basically removes the phase differences between various sinusoidal components in the speech signal to yield $R[k]$. Another view to the above equation is that $R[k]$ gives a measure of the correlation between the samples spaced at an interval of $k$, which is computed as a dot product between the two vectors in $N$ dimensional space as given below. If,

$$\begin{aligned} \mathbf{x}_0 &= \{\tilde{s}_t[0], \tilde{s}_t[1], ..., \tilde{s}_t[N-1]\} \\ \mathbf{x}_k &= \{\tilde{s}_t[k], ..., \tilde{s}_t[N-1], \tilde{s}_t[0], ..., \tilde{s}_t[k-1]\} \end{aligned} \quad (2)$$

$$R[k] = \mathbf{x}_0^T \mathbf{x}_k \quad (3)$$

If the samples spaced at an interval of $k$ are highly correlated, $\mathbf{x}_0$ will be closer to $\mathbf{x}_k$ in the $N$ dimensional space and hence will result in higher value of the dot product.

In the presence of an additive noise, say $r[n]$, the resultant signal, $s^n[n] = s[n] + r[n]$, will result in a frame $s_t^n[n]$ at time $t$. The autocorrelation $R^n[k]$ for that frame now is the dot product between two vectors given by,

$$\begin{aligned} \mathbf{x}_0^n &= \{\tilde{s}_t^n[0], \tilde{s}_t^n[1], ..., \tilde{s}_t^n[N-1]\} \\ \mathbf{x}_k^n &= \{\tilde{s}_t^n[k], ..., \tilde{s}_t^n[N-1], \tilde{s}_t^n[0], ..., \tilde{s}_t^n[k-1]\} \end{aligned}$$

where $\tilde{s}_t^n[n]$ is the periodic signal obtained from the frame $s_t^n[n]$. This $R^n[k]$ is clearly different from the $R[k]$ and is a function of the noise component present in the speech signal. As a result, whatever features we extract from these autocorrelation coefficients, these will be sensitive to the noise present in the signal.

### 2.2. Phase Autocorrelation

In an attempt to reduce the sensitivity of the correlation coefficients to the external noise present in the signal, we propose here a new measure of autocorrelation called Phase AutoCorrelation.

The magnitude of the two vectors $\mathbf{x_0}$ and $\mathbf{x_k}$ given in (2) are the same, since the set of individual vector components in these two vectors are the same. If $\|\mathbf{x}\|$ represents the magnitude of the vectors and $\theta_k$ the angle between them in the $N$ dimensional space, then (3) can be rewritten as:

$$R[k] = \|\mathbf{x}\|^2 \cos(\theta_k) \quad (4)$$

In the proposed method for correlation computation we just use the angle $\theta_k$ between the two vectors, instead of the dot product, as the measure of correlation, resulting in a new set of correlation coefficients $P[k]$ defined as:

$$P[k] = \theta_k = \cos^{-1}\left(\frac{R[k]}{\|\mathbf{x}\|^2}\right) \quad (5)$$

This new measure of correlation is referred to as the 'Phase AutoCorrelation' (PAC), as the angle between the vectors is used as the measure of correlation.

The presence of noise in the signal will affect both $\|\mathbf{x}\|$ and $\theta_k$. From the above equations, the regular autocorrelation coefficients $R[k]$ depends both on $\|\mathbf{x}\|$ and $\theta_k$, whereas the PAC coefficients $P[k]$ depend only on $\theta_k$. Consequently, $P[k]$ can be expected to be less susceptible to the external noise, as compared to $R[k]$.

### 2.3. PAC derived features

An entire class of features, which are usually derived from the regular autocorrelation coefficients, can now be derived from the PAC coefficients. DFT performed on the PAC coefficients will yield an equivalent of the regular spectrum, called PAC spectrum. Plots of the regular spectrum and the PAC spectrum for a frame of phoneme 'ih' are given in Figures 1 and 2, respectively. From the PAC spectrum, we can compute filter-banked PAC spectrum, PAC MFCC, and other features.

## 3. EXPERIMENTAL SETUP

We have conducted several experiments to illustrate the robustness of the PAC derived features. In these experiments, the speech recognition performance of the PAC derived features are compared with that of traditional features for various noise conditions. Specifically, PAC MFCCs are used in all the experiments and are compared with the regular MFCCs as well as J-RASTA-PLP features. PAC MFCCs and MFCCs were of dimension 39, including 13 static coefficients, 13 delta coefficients, and 13 delta-delta coefficients. Hidden Markov Model (HMM) emission probabilities were estimated by a Multi-Layer Perceptron (MLP) [1]

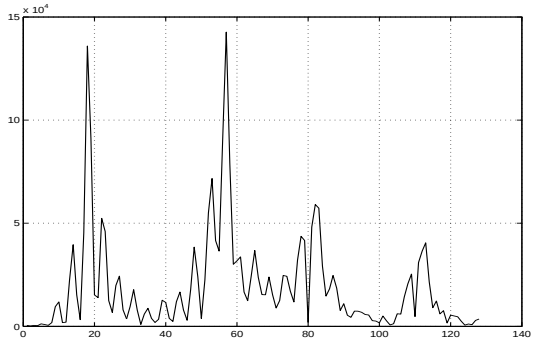

**Fig. 1**. Regular spectrum for a frame of phoneme 'ih'.

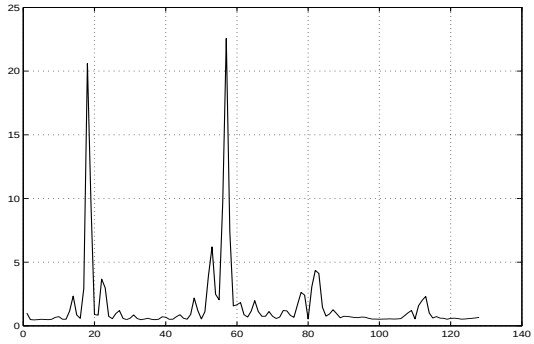

**Fig. 2**. PAC spectrum for a frame of phoneme 'ih'.

with 9 frames of contextual input and 500 hidden units. The number of output units is 27, corresponding to the number of phonemes.

All the experiments reported in this paper were conducted on OGI Numbers95 connected digits telephone speech database [7], described by a lexicon of 30 words, and 27 different phonemes. For the additive noise experiments, Factory and Lynx noises from Noisex92 database [8] and car noise from a database supplied by Daimler Chrysler Inc. (reported in this paper as 'Car') have been used. The experiments were performed at various noise levels, namely 0 dB SNR, 6 dB SNR, 12 dB SNR, and 18 dB SNR.

## 4. RESULTS AND DISCUSSION

Table 1 compares the performance of the PAC MFCC with regular MFCC for clean speech. From the table, it is clear, the performance of PAC MFCC for clean speech is inferior to the performance of MFCC. This may be because of the fact that the magnitude term in the (4) may also have significant phonetic discriminatory information, and dropping it out in the computation of $P[k]$, as given in the (5), leads to the degradation of the performance. But as explained in the previous sections, the magnitude term would certainly serve more as a confusing factor rather than as an useful factor, in the presence of external noise. Hence drop out of the magnitude term for the PAC coefficients should result in improved performance in case of noisy conditions. Experimental results obtained using noisy data show that this is indeed the case.

| Feature | Word Recognition Rate, % acc. |
|---|---|
| MFCC | 90.1 |
| PAC MFCC | 86.0 |

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

Figures 3, 4, and 5 show the performance comparison of the PAC MFCCs with the regular MFCCs for various noise conditions and various noise levels. From these figures it is clear that the performance of the PAC MFCCs is far superior than what can be achieved with regular MFCCs in the presence of the external noise. For all the noise conditions shown, the degradation of the performances for the PAC MFCCs are much slower than that of the MFCCs.

Moreover, we have also tried to compare the performances of PAC MFCC features with the J-RASTA-PLP features [6], which is a well known approach for noise robust speech feature extraction. Figures 6, 7 show the results of the experiments for Factory and Lynx noises, respectively. The PAC MFCC features are performing even better than J-RASTA-PLP features in extreme noise conditions, like Factory noise. For Lynx noise, which is a well behaved noise, RASTA processing works better. The above comparison between the PAC MFCC and J-RASTA-PLP is just to illustrate the usefulness of the PAC derived features. Otherwise, the comparison is not really valid since RASTA processing could also be applied to the PAC spectrum for further improving their robustness.

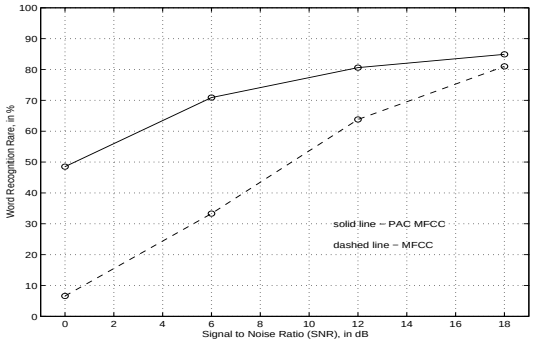


**Fig. 3**. Performance curves for Factory noise.

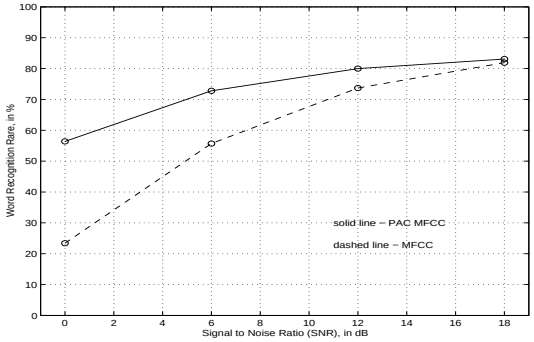


**Fig. 4**. Performance curves for Lynx noise.

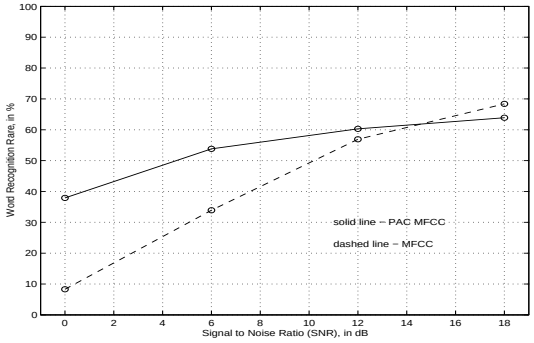


**Fig. 5**. Performance curves for Car noise.

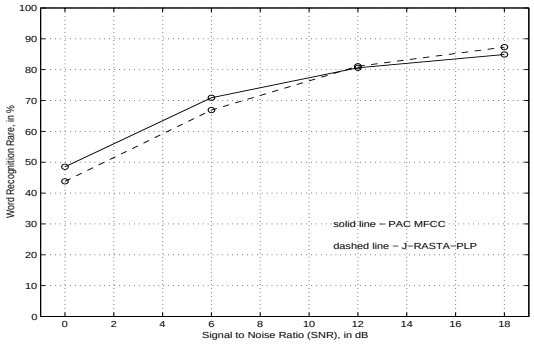


**Fig. 6**. Performance comparison of PAC MFCC and J-RASTA-PLP features for Factory noise.

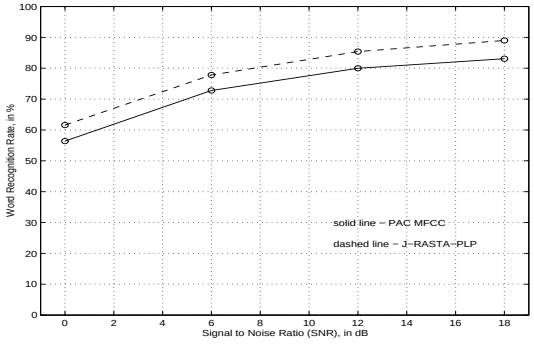


**Fig. 7**. Performance comparison of PAC MFCC and J-RASTA-PLP features for Lynx noise.

## 5. CONCLUSION

We have introduced a new category of features called Phase AutoCorrelation derived features. These features are extracted from the phase autocorrelation coefficients which are computed as the angle between two signal vectors separated in time by a particular interval. This use of angle as a measure of correlation makes the phase autocorrelation coefficients less sensitive to noise as compared to regular autocorrelation coefficients, which are computed as the dot product of the two vectors. This fact makes the PAC derived features significantly more robust to noise than the traditional features. The noise robustness of PAC derived features has been illustrated through the experimental results we have provided on Numbers 95 and Noisex92 databases.

As future work, the robustness of these features can be further improved by applying robust techniques such as RASTA processing over the PAC spectrum. Furthermore, these PAC features can be used as stand alone features or as complementary features in addition to regular features, e.g., in the multistream speech recognition framework.

**Acknowledgments:** 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)'. The authors also thank Nelson Morgan and Hynek Hermansky for their comments on the initial version of this paper.

## 6. REFERENCES

[1] 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.

[2] L. Rabiner, and B. H. Juang, "Fundamentals of Speech Recognition," *PTR Prentice-Hall, Inc.*, Englewood Cliffs, NJ, USA, 1993.

[3] A. V. Oppenheim, and R. W. Schafer, "Digital Signal Processing," *PTR Prentice-Hall, Inc.*, Englewood Cliffs, NJ, USA, 1975.

[4] L. R. Rabiner, and R. W. Schafer, "Digital Processing of Speech Signals," *PTR Prentice-Hall, Inc.*, Englewood Cliffs, NJ, USA, 1978.

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

[6] 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.

[7] 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.

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