Robust Feature Extraction based on Spectral Peaks of Group Delay and Autocorrelation Function and Phase Domain Analysis

G. Farahani[1], S.M. Ahadi[1] and M.M. Homayounpoor[2]

[1]Electrical Engineering Department, [2]Computer Engineering Department
Amirkabir University of Technology
Hafez Ave., Tehran 15914, Iran

f8023953@aut.ac.ir, sma@aut.ac.ir, homayoun@ce.aut.ac.ir

10.21437/Interspeech.2006-130

Abstract

This paper presents a new robust feature set for noisy speech recognition in phase domain along with spectral peaks obtained from group delay and autocorrelation functions.

The group delay domain is appropriate for formant tracking and autocorrelation domain is well-known for its pole preserving and noise separation properties. In this paper, we report on appending spectral peaks obtained in either group delay or autocorrelation domains to the feature vectors extracted originally in phase domain to create a new feature set.

We tested our features on the Aurora 2 noisy isolated-word task and found that it led to improvements over other group delay-based and autocorrelation-based methods that use magnitude instead of phase for feature extraction**.**

Index Terms: robust speech recognition, spectral peak, group delay, autocorrelation

1. Introduction

In many traditional methods the feature vector is obtained from methods exploiting short-time magnitude spectrum such as MFCC. However, features extracted using magnitude are known to be more sensitive to the changes in the environmental conditions such as noise and channel distortions. Therefore, the performance of such Automatic Speech Recognition (ASR) systems will severely degrade in noisy conditions.

Many methods have been proposed to reduce such performance degradations. These methods, from one point of view, could be classified into two major groups, i.e. magnitude and phase domains.

Some examples of the methods that work in the magnitude domain are RAS [1], AMFCC [2], DAS [3], Spectral Subtraction (SS), RelAtive SpcTrAl (RASTA) filtering etc.
On the other hand methods in the phase domain include Phase AutoCorrelation (PAC) [4] and methods that use group delay (differentiated phase) as a base for feature extraction [5-7]. Also the group delay domain is known as an appropriate domain for formant tracking and peak isolation [8].

The above-mentioned finding in the phase domain has persuaded us to use the signal phase information in the feature vector.

Autocorrelation domain is another domain that has attracted attention in robust speech recognition. A number of feature extraction algorithms have been devised using this domain as the initial domain of choice and have led to some improvements in the efficiency of ASR systems [1-3].

Also, it is well-known that spectral peaks convey important information of the speech signal and using these peaks in feature vector can help in improving the recognition rate of *ASR* systems [9, 10].

The focus of this paper is on the use of phase information of speech signal to improve the recognition rate of noisy signal. Also due to the advantages associated with the use of group delay and autocorrelation domains, we decided to use these domains for peak isolation and extension of the feature vector which is itself extracted in phase domain.

This paper is constructed as follows. The following section reviews the autocorrelation and phase domains and will describe the mathematical basics of our proposed method. Section 3 describes the proposed algorithm for feature extraction. In section 4, our experimental results will be discussed and section 5 concludes the paper.

2. Autocorrelation and phase domains

In this section we will describe the autocorrelation and phase domains and their associated mathematical formulas.

2.1. Autocorrelation domain

If we assume *w(t)* to be the additive noise, *x(t)* the clean speech signal and *y(t)* the noisy speech signal, then we can write:

$$y(t) = x(t) + w(t)\,. \tag{1}$$

In discrete domain, we will have

$$y(m,n) = x(m,n) + w(m,n) \qquad \begin{matrix} 0 \le n \le N-1 \\ 0 \le m \le M-1 \end{matrix} \tag{2}$$

where N is the frame length and n is the discrete time index in a frame, m is the frame index and M is the number of frames. If noise is considered to be uncorrelated with speech, we will have the following relationship between the autocorrelations of noisy speech, clean speech, and noise*, i.e.*

$$R_y(m,k) = R_x(m,k) + R_w(m,k) \qquad \begin{matrix} 0 \le m \le M-1 \\ 0 \le k \le N-1 \end{matrix} \tag{3}$$

where $R_y(m,k)$, $R_x(m,k)$ and $R_w(m,k)$ are the short-time autocorrelation sequences of the noisy speech, clean speech and noise respectively.

As mentioned earlier, feature extraction from magnitude spectrum will be obtained by applying DFT on the frame samples. DFT assumes each frame, *y(m,n),* is a part of periodic signal, $\tilde{y}(m,n)$, which is defined as :

$$\tilde{y}(m,n)=\sum_{k=-\infty}^{+\infty} y(m,n+kN) \quad 0\le m\le M-1,\ 0\le n\le N-1. \quad (4)$$

The estimator for the calculation of autocorrelation sequence is then given as:

$$R_y(m,k)=\sum_{i=0}^{N-1}\tilde{y}(m,i)\tilde{y}(m,i+k) \qquad \begin{matrix}0\le m\le M-1\\ 0\le k\le N-1\end{matrix} \quad (5)$$

Another view to equation (5) is that $R_y(m,k)$ gives the correlation between the samples spaced at interval k, which is computed as dot product of two vectors in *N*-dimensional domain, i.e.

$$\begin{aligned}&Y_0=\{\tilde{y}(m,0),\tilde{y}(m,1),...,\tilde{y}(m,N-1)\}\\ &Y_k=\{\tilde{y}(m,k),...,\tilde{y}(m,N-1),\tilde{y}(m,0),...,\tilde{y}(m,k-1)\}\\ &R_y(m,k)=Y_0{}^T Y_k.\end{aligned} \quad (6)$$

If we carry out these steps for clean speech, *x(n,m),* we would have

$$\begin{aligned}&R_x(m,k)=X_0{}^T X_k\\ &X_0=\{\tilde{x}(m,0),\tilde{x}(m,1),...,\tilde{x}(m,N-1)\}\\ &X_k=\{\tilde{x}(m,k),...,\tilde{x}(m,N-1),\tilde{x}(m,0),...,\tilde{x}(m,k-1)\},\end{aligned} \quad (7)$$

where $\tilde{x}(m,n)$ is the periodic signal obtained from *x(m,n)*. Clearly, the autocorrelation sequences for clean and noisy signals are different. Therefore, features extracted from autocorrelation sequences would be sensitive to noise.

2.2. Phase domain

As mentioned above, if the speech features are extracted from squared magnitude spectrum of signal (DFT of autocorrelation sequence), they will be sensitive to noise.

From (6), we can see that the magnitude of two vectors Y_0 and Y_k is the same. If we assume $|Y(m)|$ to be the magnitude of vectors and $\theta_y(m,k)$ the angle between them, then we could write the relationship between the autocorrelation, $R_y(m,k)$, magnitude of the vectors and the angle between them as follows:

$$R_y(m,k)=|Y(m)|^2\cos\theta_y(m,k) \qquad \begin{matrix}0\le m\le M-1\\ 0\le k\le N-1.\end{matrix} \quad (8)$$

Now the angle $\theta_y(m,k)$ between the two vectors will be calculated as:

$$\theta_y(m,k)=\cos^{-1}\left(\frac{R_y(m,k)}{|Y(m)|^2}\right) \qquad \begin{matrix}0\le m\le M-1\\ 0\le k\le N-1.\end{matrix} \quad (9)$$

2.2.1. Group delay function

For calculating group delay function, if we assume that *x(n), n*=0,1,...,*N*-1, is a segment of speech signal, first we calculate *y(n)* as

$$y(n)=n\,x(n) \qquad n=0,\,1\ldots N\text{-}1. \quad (10)$$

Now, if we define $X(k)$ and $Y(k)$ as Fourier Transforms of *x(n)* and *y(n)* respectively, then the group delay function is defined as follows [5]:

$$\tau_0(k)=\frac{X_R(k)Y_R(k)+X_I(k)Y_I(k)}{X_R(k)^2+X_I(k)^2} \qquad k=0,1\ldots N\text{-}1, \quad (11)$$

where $X_R(k)$, $Y_R(k)$, $X_I(k)$ and $Y_I(k)$ are real and imaginary parts of $X(k)$ and $Y(k)$ respectively.

In order to prevent the spikes on the group delay of signal, we will use a modified group delay as [6]

$$\tau(k)=\frac{X_R(k)Y_R(k)+X_I(k)Y_I(k)}{S(k)^{2\alpha}} \qquad k=0,\,1\ldots N\text{-}1 \quad (12)$$

$$\tau_P(k)=\tau(k)|\tau(k)|^{\beta-1} \qquad k=0,\,1\ldots N\text{-}1, \quad (13)$$

where *S(k)* is the cepstrally-smoothed spectrum of $|X(k)|$ and α and β are two constants in the range of 0 to 1. These two parameters should be fine tuned according to environmental condition. We have set the parameters as in [6], i.e. $\alpha=0.9$ and $\beta=0.4$.

3. Proposed method

In this section, our method of feature extraction in phase domain plus the extraction of extra feature parameters in group delay and autocorrelation domains will be proposed. As mentioned in [8], group delay domain is an appropriate domain for formant tracking. Therefore, the use of group delay function for tracking spectral peaks will be considered as a way to obtain robust features under noisy conditions. Also, according to the effectiveness of autocorrelation function for preserving peaks, we will also use the autocorrelation domain for extracting first 3 formants of the speech signal as well as the group delay domain [9, 10].

3.1. Feature extraction in phase domain

In Figure 1 we have shown the calculation of feature parameters in both group delay and autocorrelation domains to extend the features extracted in phase domain.

Similar to other front-end diagrams, first the speech signal was divided into frames and then a Pre-emphasis filter was used in each frame to give more weight to higher frequency components. Later, a Hamming window was applied to suppress the boundary effects of Frame Blocking. The next step was the calculation of the autocorrelation function according to (5) and the phase angle, $\theta_y(m,k)$, as mentioned in (9). The rest of the front-end calculations were similar to ordinary MFCC front-end calculations. As it is clear from (9), these features are related only to the phase variations, in contrast to the features based on

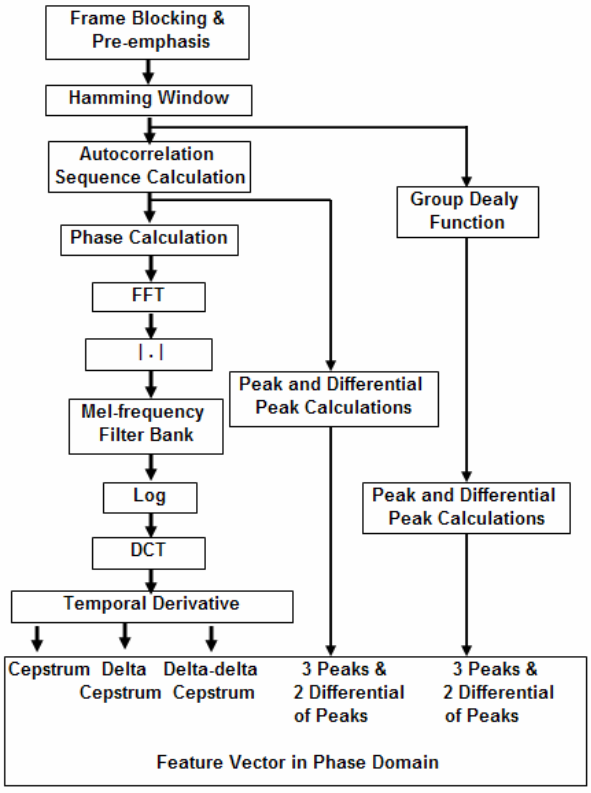

Figure 1 *Front-end diagram to extract features in phase domain along with autocorrelation and group delay functions.*

the magnitude, such as MFCC, that are related to both $|Y(m)|$ and $\theta_y(m,k)$[4].

3.2. Adding peaks to phase feature vector in group delay domain

As explained, group delay domain is a good candidate for formant tracking and also spectral peak tracking. For this reason we used a group delay function, as mentioned in (13), for peak isolation.

The peaks were calculated using the algorithm that will be discussed in section 3.4, extracted in group delay domain. Three peak frequencies and two differentials of them were then added to the feature vector.

As mentioned in [7], for group delay calculation in (13), the smoothed spectrum was calculated using the first 12 cepstral coefficients. The path for peak isolation in group delay domain is depicted in Figure 1. We called the new features, found after appending these parameters to the original feature vector, *Group Delay Peaks and Phase features* (GDPP).

3.3. Adding peaks to phase feature vector in autocorrelation domain

As depicted in Figure 1, our proposed method in autocorrelation domain is similar to that in group delay domain. The main difference is that in this domain, we have initially calculated the autocorrelation of the signal. Then, the first three peaks locations and their derivatives were calculated using the signal autocorrelation spectrum as will be explained in section 3.4. Finally, these values were added to the extracted feature vector in phase domain. The new coefficients were named *Autocorrelation Peaks and Phase features* (APP).

3.4. Peak threading algorithm

As mentioned in [9, 10], the peaks of the speech spectrum are important for speech recognition. Hence, we decided to add three peak frequencies and two peak derivatives to the feature vector.

For peak calculation, we used the peak threading method that is rather accurate in finding the location of peak frequencies in spectral domain [9]. For this, first we applied a set of triangular filters to the signal. These filters had bandwidths of 100 Hz for center frequencies below 1 kHz and bandwidths of one tenth the center frequency for the frequencies above 1 kHz. Then we applied an AGC (Automatic Gain Control) to the filter outputs.

In our implementation, we used a typical AGC that slowly adapts the output level, so that its value is maintained near that of the target level when the levels of input change. Therefore, the inputs below 30 dB are amplified linearly by 20dB and inputs above 30 dB are amplified increasingly less.

After finding the isolated peaks, the peaks were threaded together and smoothed. Then three peak frequencies and two peak derivatives were found and added to the feature vector.

4. Experimental work

The proposed approach was implemented on Aurora 2 task [11]. The feature vectors for both proposed methods were composed of 12 cepstral and one log-energy parameters, together with their first and second derivatives and five extra components of which three were for the first three formants and the other two for the frequency peak derivatives. Therefore, our feature vectors were of size 44. All model creation, training and tests in all our experiments have been carried out using the HMM toolkit [12].

Figure 2 displays the results obtained using MFCC, PAC (Phase AutoCorrelation) and our proposed methods (APP and GDPP). Also, for comparison purposes, we have included the results of adding spectral peaks to feature vectors calculated using magnitude spectrum named TSP (Threaded Spectral Peaks), GDFP (Group Delay Function Peak) and ACP (AutoCorrelation Peaks) [10]. As discussed in [10], the algorithm for the extraction of these features from magnitude spectrum was the same and a feature vector size of 44, the same as that used here, was used.

According to this figure, APP and GDPP methods have led to better recognition rates in comparison to most of the other methods while GDPP outperformed other methods for all test sets. This result is similar to the results mentioned in [10] where the group delay domain was found more appropriate, for peak isolation, than the autocorrelation domain. Here, we see that both domains lead to better results when combined with phase domain features.

In Table 1, we have summarized the average recognition rates obtained for each test set of Aurora 2. As can be seen, average recognition rates for features extracted using the group delay domain are better than those of autocorrelation-based features. This indicates that while spectral peaks extracted from

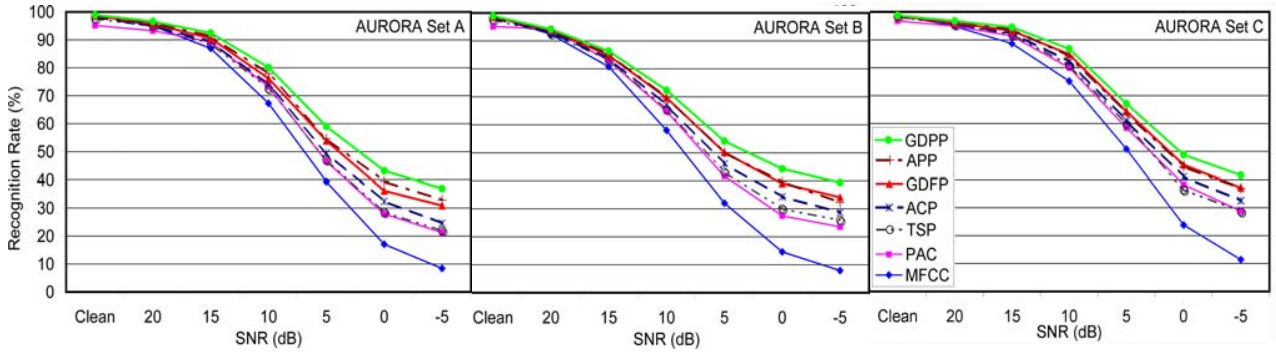

Figure 2 *Average recognition rate on Aurora 2 database. (a) Test set a, (b) Test set b, (c) Test set c. The results correspond to MFCC, PAC, TSP, ACP, GDFP, APP and GDPP methods.*

Table 1. *Comparison of Average recognition rates for various feature types on three test sets of Aurora 2 task.*

Feature type	Average Recognition Rate (%)		
	Set A	Set B	Set C
MFCC	61.13	55.57	66.68
PAC	66.02	62.25	72.60
TSP	66.31	62.86	72.89
ACP	68.03	64.86	74.46
GDFP	70.49	67.50	76.74
APP	71.83	67.69	76.53
GDPP	74.28	68.89	78.81

both these domains are very useful in improving the robustness of a recognition system, group delay domain achieves more robustness in comparison to autocorrelation domain, so that the method using group delay peaks and phase features (GDPP) tops all the results obtained.

5. Conclusion

In this paper two new robust feature extraction methods have been proposed. As the features extracted in magnitude domain are more sensitive to the background noise in comparison to phase domain, we used phase domain as a base for feature extraction and for further boosting the robustness, spectral peaks and their derivatives were added to the feature vector.

A similar procedure was carried out before using base features extracted in magnitude domain. In this paper, we have shown that features extracted using phase domain and extended by these spectral peak parameters can lead to even better results in comparison to the magnitude spectrum. Two domains that are found appropriate for robustness in speech recognition systems and also in formant extraction, namely autocorrelation and group delay domains, were used for spectral peak extraction.

Among the two, it was observed that the peaks found using group delay domain were more robust in comparison to the autocorrelation domain peaks.

6. Acknowledgment

This work was in part supported by a grant from the Iran Telecommunication Research Center (ITRC).

7. References

[1] You, K.-H. and Wang, H.-C., “Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences”, *Speech Communication*, 28:13-24, 1999.

[2] Shannon, B.J. and Paliwal, K.K., “MFCC Computation from Magnitude Spectrum of higher lag autocorrelation coefficients for robust speech recognition”, in *Proc. ICSLP*, Jeju, Korea, pp. 129-132, Oct. 2004.

[3] Farahani, G. and Ahadi, S.M., “Robust features for noisy speech recognition based on filtering and spectral peaks in autocorrelation domain”, in *Proc. EUSIPCO*, Antalya, Turkey, 2005.

[4] Ikbal, S., Misra, H. and Bourlard, H., “Phase autocorrelation (PAC) derived robust speech features”, in *Proc. ICASSP*, Hong Kong, pp. II-133-136, April 2003.

[5] Yegnanarayana, B., Murthy, Hema A. and Ramachandran, V. R., "Processing of noisy speech using modified group delay functions", Proc. ICASSP'91, pp. 945-948, 1991.

[6] Hegde, R. M., Murthy, H. A. and Gadde, V. R. R., “Continuous speech recognition using joint features derived from the modified group delay function and MFCC”, in Proc. ICSLP, Jeju, Korea, Oct. 2004.

[7] Farahani, G., Ahadi, S.M. and Homayounpoor, M.M., “Robust feature extraction using group delay function for speech recognition”, in *Proc. SPECOM*, Patras, Greece, 2005.

[8] Bozkurt, B., Doval, B., D'Alessandro, C. and Dutoit, T., “Improved differential phase spectrum processing for formant tracking”, in *Proc. ICSLP*, Jeju, Korea, Oct 2004.

[9] Strope, B. and Alwan, A., “Robust word recognition using threaded spectral peaks”, in *Proc. ICASSP*, pp. 625-628, Washington, USA, 1998.

[10] Farahani, G., Ahadi, S.M. and Homayounpoor, M.M., “ Use of spectral peaks in autocorrelation and group delay domains for robust speech recognition”, in *Proc. ICASSP*, Toulouse, France, 2006.

[11] Hirsch, H.G. and Pearce, D., "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proc. ISCA ITRW ASR*, 2000.

[12] The hidden Markov model toolkit available from http://htk.eng.cam.ac.uk.