USE OF SPECTRAL PEAKS IN AUTOCORRELATION AND GROUP DELAY DOMAINS FOR ROBUST SPEECH RECOGNITION

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

ABSTRACT

This paper presents a new front-end for robust speech recognition. Two scenarios are used for the features extracted in autocorrelation and group delay domains. These new front-end scenarios will focus on the spectral peaks of speech in two mentioned domains. Therefore we will address the issue of using spectral peak location information in a feature vector for robust speech recognition.

A task of speaker-independent isolated-word recognition was used to demonstrate the efficiency of these robust front-end diagrams. The cases of white noise and different colored noises such as babble, factory and car noises were tested. Experimental results show significant improvements in comparison to the results obtained using traditional front-end diagrams.

1. INTRODUCTION

A considerable amount of efforts have been devoted within the past few decades to improve the speech recognition robustness in adverse environments. Despite important achievements in improving recognition systems, these systems typically do not work well in cases of even slight changes in the acoustic environment. Many of these efforts have been made to reduce the mismatches between training and test conditions.

Successful research works in robust speech recognition using feature parameters, especially in autocorrelation and group delay domains, have been carried out. These efforts in autocorrelation domain were initiated with the introduction of SMC (short-time modified coherence) [1], OSALPC (one-sided autocorrelation LPC) [2] and RAS (relative autocorrelation sequence) [3]. Recently, further improvements in this field have been reported [4, 5]. An important property of the autocorrelation domain is its pole preserving property. Therefore, the spectral properties of the speech signal are well preserved after transformation to the autocorrelation domain. In most ASR systems, either amplitude or power spectrum of the speech signal has been used for feature extraction. However, recent studies on speech perception have revealed the importance of the phase of speech signal [6, 7]. The main problem in phase spectrum estimation are zeros of signal which are close to the unit circle, causing spikes on the group delay of signal (derivative of the phase spectrum) [8]. In order to overcome the problem of spikes in group delay, some researchers have suggested solutions such as modified group delay [8] and product spectrum [9]. Group delay is also found to be a good domain for formant tracking [10, 11].

Group delay is an important feature of the signal that can help in enhancing the signal quality in noisy conditions [12]. It has shown good performance in other speech applications.

Previous research works have revealed the usefulness of group delay in processing the speech signals. In [13], a method for determining the instants of significant excitations in speech signals using group delay function, which works well for all types of voiced speech, has been proposed. Due to the promising results obtained from applying group delay to similar signal processing applications, we decided to use it in a pre-processing stage of our speech recognizer. Also, we decided to combine the use of spectral peaks with group delay in order to further improve the robustness of our approach. Spectral peaks are found to be very important in obtaining a robust set of features in speech recognition [14].

Although the issue of incorporating spectral peak information in a speech recognition system has been dealt with previously [14], direct use of spectral peaks as speech features might lead to inconsistencies in feature vector lengths due to unavailability of enough number of peaks in all frames. In this paper we propose an algorithm that incorporates maximum information into feature vector, extracted in autocorrelation and group delay domains. Furthermore, a new method for creating feature vector in autocorrelation and group delay domains is introduced and the results compared to methods such as RAS (relative autocorrelation sequence), GDF (group delay features) [15] and MFCC (mel-frequency cepstral coefficients).

2. AUTOCORRELATION AND GROUP DELAY FUNCTION OF SPEECH SIGNAL

In this section a procedure for calculating autocorrelation and group delay function of speech signal will be described.

2.1. Autocorrelation Function

If *v(t)* is the ambient noise and *x(t)* the clean speech signal, then noisy speech signal *y(t)* can be modeled as

$$y(t) = x(t) + v(t) \tag{1}$$

Since the speech signal is time-variant and non-stationary, it is usually analyzed in the discrete domain. Thus in each frame we have:

$$y(n) = x(n) + v(n) \qquad 0 \le n \le N-1 \tag{2}$$

where N is the frame length and n is the discrete time index in a frame. If the noise and speech are assumed uncorrelated, the autocorrelation of the noisy speech is the sum of the autocorrelations of the clean speech, *x(n),* and the noise v*(n), i.e.*

$$r_y(k) = r_x(k) + r_v(k) \qquad 0 \le k \le N-1 \tag{3}$$

where $r_y(k)$, $r_x(k)$ and $r_v(k)$ are the short-time autocorrelation sequences of the noisy speech, clean speech and noise respectively.

The unbiased estimator for the calculation of one-sided autocorrelation sequence is given as follows:

$$r_y(k) = \frac{1}{N-k} \sum_{i=0}^{N-1-k} y(i)y(i+k) \qquad 0 \le k \le N-1 \tag{4}$$

As the autocorrelation of noise is considered relatively constant with time, a high-pass filter should be able to reduce its effect. Therefore a filter with the following transfer function is used [15]:

$$H(z) = \sum_{t=-L}^{L} t.z^t \Big/ \sum_{t=-L}^{L} t^2 \tag{5}$$

In our experiments we have used L=2.

2.2. Group delay Function

Given a segment of speech signal, *x(n), n*=0,1,...,*N*-1, the group delay function can be computed as follows:

First we calculate *y(n)* as

$$y(n) = n\,x(n) \qquad n=0, 1\ldots N\text{-}1 \tag{6}$$

We denote $X(k)$ and $Y(k)$ as Fourier Transforms of *x(n)* and *y(n)* respectively. Then the samples of group delay function can be written as follows [15]:

$$\tau_0(k) = \frac{X_R(k)Y_R(k) + X_I(k)Y_I(k)}{X_R(k)^2 + X_I(k)^2} \quad k=0, 1\ldots N\text{-}1 \tag{7}$$

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 follows [8]

$$\tau(k) = \frac{X_R(k)Y_R(k) + X_I(k)Y_I(k)}{S(k)^{2\alpha}} \qquad k=0, 1\ldots N\text{-}1 \tag{8}$$

$$\tau_P(k) = \tau(k)\left|\tau(k)\right|^{\beta-1} \qquad k=0, 1\ldots N\text{-}1 \tag{9}$$

where *S(k)* is cepstrally smoothed spectrum of $|X(k)|$ in order to reduce spikes in group delay function [15]. α and β should be fine tuned according to environment. We have set the parameters as in [8], i.e. $\alpha = 0.9$ and $\beta = 0.4$.

3. PROPOSED METHOD

In this section the method of feature extraction in autocorrelation and group delay domains is proposed. As mentioned in [15], the speech spectrum reconstruction using group delay will reduce fluctuations caused by the variance of noise. Also, group delay domain is an appropriate domain for formant tracking. Therefore, the use of group delay function for tracking spectral peaks can be considered as a way to obtain robust features under noisy conditions. Furthermore, due to the importance of spectral peaks and also the effectiveness of autocorrelation function, we will also use the autocorrelation domain for extracting first 3 formants of the speech signal.

3.1. Feature extraction in group delay domain

Figure 1 depicts the algorithm followed to compute the new features. Similar to other front-ends, the speech signal, *s(n)*, was initially divided into frames by the frame blocking process. Pre-emphasis was then carried out to give more weight to higher frequency components. Later, a Hamming window was used to suppress the effects of Frame Blocking. The next step was the calculation of the group delay function as mentioned in equation (9). Then, the algorithm mentioned in section 3.3 was utilized to perform peak calculation and differentiate them and also magnitude estimation carried out as in [15]. Finally, three frequencies and two differentials were added to the feature vector.

As mentioned in [15], the smoothed spectrum was calculated using the first 12 cepstral coefficients. As seen in Figure 1, the dashed lines show the path for feature extraction in group delay domain. The rest of the front-end calculations were followed similar to ordinary MFCC calculations. The new coefficients were named *group delay function peaks* (GDFP).

As shown, in the GDFP method, three more stages were introduced in front-end processing, such that first we calculate the group delay function, then estimate the magnitude of the signal after modifying the group delay and finally calculate the peaks from group delay function. Therefore, a smoother signal magnitude is obtained while the effect of noise is reduced and also better formant frequencies are resulted, which will be added to the feature vector.

In Figure 1, the box named “peaks calculation and differential of peaks” displays how the peak threading method can be integrated in our front-end. Since after the application of group delay function, the effect of noise is reduced and also in this domain the peaks of signal will be more clear, we expect the resultant feature vector to be more robust to noise.

3.2. Feature extraction in autocorrelation domain

Figure 1 also includes our proposed method for autocorrelation domain feature extraction (solid lines). The algorithm, in autocorrelation domain, has much similarity, in many steps, to the group delay algorithm. In this domain, we have initially calculated the unbiased autocorrelation of the signal. Then, the first three peaks locations and their derivatives were calculated using the signal autocorrelation. These values were later added to the feature vector. Furthermore, we continued the front-end diagram with/without filtering as mentioned in (5). FFT was applied after the application of the Hamming window and the remainder was the same as an ordinary front-end typically used for calculating MFCC parameters. If we use the filter after the autocorrelation of the signal, we obtain a cleaner signal, compared to the original noisy signal. Therefore, we expect the features calculated after the application of this filter to perform better, in comparison to those obtained using the unfiltered autocorrelation.

3.3. Adding peaks to features

It is well-known that the peaks of the speech spectrum are important for speech recognition. Therefore, we have added three peak frequencies and two peak derivatives to the feature vector. For peak calculations, we have used the peak threading method

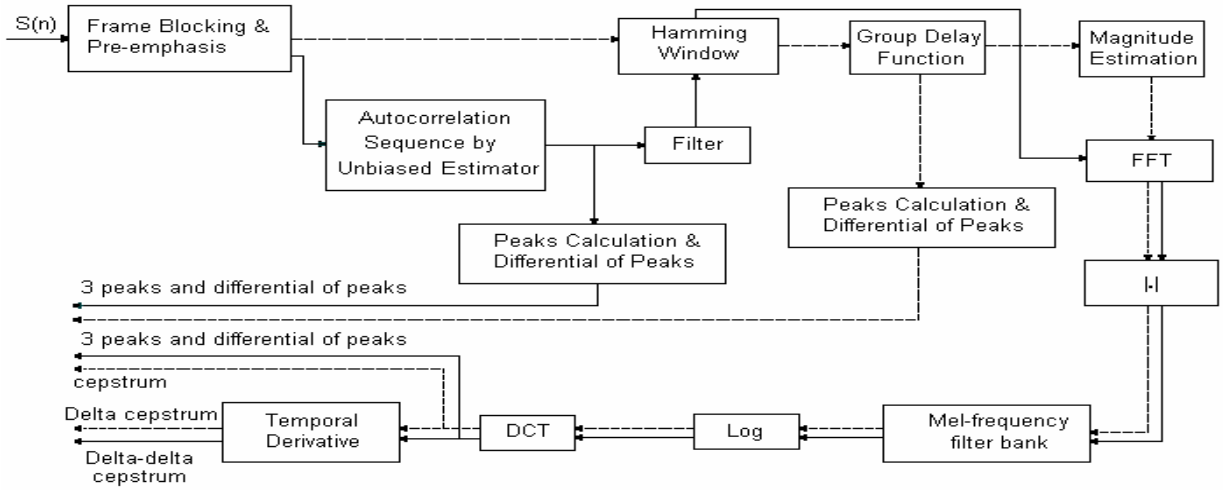

Figure 1: Front-end diagram to extract features in group delay and autocorrelation domains. Solid lines show the procedure for feature extraction in autocorrelation domain and dash lines show the procedure for feature extraction in group delay domain.

that is more accurate in finding the location of peak frequencies in spectral domain [14].

First of all, we have 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. The next step was to apply AGC (Automatic Gain Control) to the filter outputs. In our algorithm, we used a typical AGC for our dynamic model. This AGC slowly adapts to maintain the output level near 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 at this stage, the peaks were threaded together and smoothed. Then, as mentioned, three peak frequencies and two peak derivatives were chosen and added to the feature vector. Details of peak isolation and threading can be found in [14]. In this paper, spectral peaks obtained using unfiltered signal autocorrelation, as depicted in the front-end diagram, are called ACP (autocorrelation peaks) and those obtained using filtered signal autocorrelation, ACPAF (autocorrelation peaks after filtering). We have also implemented a feature extraction procedure similar to [14], except that a different AGC was used, as explained above. This will be called *threaded spectral peaks* (TSP).

4. EXPERIMENTAL WORK

The speech corpus used in these experiments is a speaker-independent isolated-word Farsi (Persian) corpus. The corpus was collected from 65 male and female adult speakers uttering the names of 10 Iranian cities. The data was collected in normal office conditions with SNRs of 25dB or higher and a sampling rate of 16 kHz. Each speaker uttered 5 repetitions of words, some of which were removed from the corpus due to problems that occurred during the recordings. The 2665 utterances from 55 speakers were used for HMM model training. The test set contained similar data from 10 speakers (5 male & 5 female) that were not included in the training set.

The noise was then added to the speech in different SNRs. The noise data was extracted from the NATO RSG-10 corpus [16].We have considered babble, car, factory1 and white noises and added them to the clean signal at 20, 15, 10, 5, 0 and -5 dB SNRs. Our experiments were carried out using MFCC (for comparison purposes), RAS, GDF, TSP and our three new methods, GDFP, ACP and ACPAF. The features in all cases were computed using 25 msec. frames with 10 msec. of frame shifts. Pre-emphasis coefficient was set to 0.97. For each speech frame, a 24-channel Mel-scale filter-bank was used. Each word was modeled by an 8-state left-right HMM and each state was represented by one Gaussian PDF. The feature vectors for three proposed methods were composed of 12 cepstral and a log-energy parameter, together with their first and second derivatives and five extra components of which three are for the first three formants and the other two for the frequency peak derivatives. Therefore, our feature vectors were of size 44. All feature extractions (for MFCC), model creation, training and tests have been carried out using the HMM toolkit (HTK) [17]. Figure 2 depicts the results of our implementations. Also the averages of our results are displayed in Table 1. The average values mentioned in this table are calculated over the results obtained from 0 dB to 20 dB SNRs, omitting the clean and -5 dB results.

As seen in Table 1, the recognition rate using MFCC features is seriously degraded by different noises, while RAS and GDF methods exhibit more robustness. Adding the peaks, especially in group delay domain, approved the effectiveness of group delay domain for peak tracking. As the results indicate, while ACP achieves a noticeable improvement in the baseline performance in noise, the combination of ACP with FIR filter works better than ACP alone. As seen in Table 1, features extracted using the group delay domain work better than the autocorrelation-based features, outperforming other methods in noisy conditions. This indicates that while both these domains are useful in improving the robustness of the recognition system, group delay domain achieves more robustness in comparison to autocorrelation domain, so that the group delay performance tops all the results obtained. The results obtained can be listed in brief as follows:

(1) The GDFP outperforms other methods.
(2) The improvements for car noise are very slight, as most of the feature extraction techniques perform almost similar in that case.
(3) Appending frequency peaks to the feature vector can further improve the results obtained using autocorrelation and group delay based features.

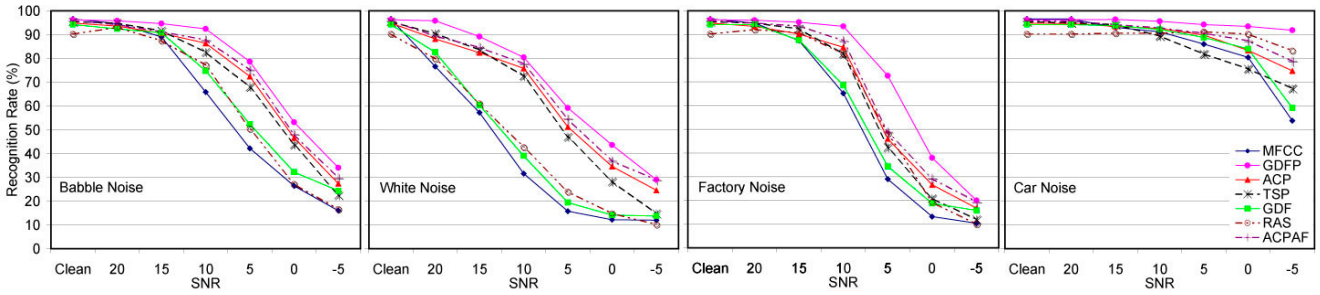

Figure 2: Recognition Rates for different noises and various methods. Acronyms stand for: MFCC: mel frequency cepstral coefficients; GDFP: group delay function peaks; ACP: autocorrelation peaks; TSP: threaded spectral peaks; GDF: group delay features; RAS: relative autocorrelation sequence; ACPAF: autocorrelation peaks after filtering.

ACKNOWLEDGMENT

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

Table 1: Comparison of average recognition rates for various feature types with babble, car, factory1 and white noises.

Feature type	Average Recognition Rate			
	Babble	Car	Factory1	White
MFCC	63.60	89.44	57.92	38.68
GDFP	82.92	95.16	79.04	73.64
ACP	77.92	90.76	68.16	66.44
ACPAF	79.28	92.12	70.76	68.64
TSP	76.16	87.20	66.44	64.36
GDF	68.52	90.60	60.72	43.12
RAS	67.04	90.56	66.40	44.44

5. CONCLUSION AND FUTURE WORK

Improved robustness in speech recognition using new feature extraction methods was discussed in this paper. A key issue for practical applications of ASR systems is robustness. Two domains that are found appropriate for robustness in speech recognition systems are autocorrelation and group delay domains. Techniques based on the above methods and the use of spectral peaks was discussed in this paper. Our proposed front-end diagrams in autocorrelation and group delay domains were evaluated together with several different robust feature extraction methods. The usefulness of these techniques was shown and the results indicate that the spectral peaks inherently convey robust information for speech recognition, especially in autocorrelation and group delay domains. However, it is observed that the peaks in group delay domain are more robust in comparison to autocorrelation domain peaks. The method used for the estimation of the signal magnitude is believed to partially suppress the effect of noise in group delay domain before spectral analysis is performed. The results indicate the potential of these two domains for use in speech recognition.

Better parameter optimization can be a basis for the future work as it is believed to have important influence on the system performance.

6. REFERENCES

[1] D. Mansour, B.-H. Juang, " The short-time modified coherence representation and noisy speech recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 37, no. 6, pp. 795-804, 1989.

[2] J. Hernando, C. Nadeu, "Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 5, no.1, pp. 80-84, 1997.

[3] Kuo-Hwei Yuo and Hsiao-Chuan Wang, "Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences," *Speech Communication*, vol. 28, pp.13-24, 1999.

[4] B.J. Shannon, K.K. Paliwal " MFCC Computation from Magnitude Spectrum of higher lag autocorrelation coefficients for robust speech recognition," in *Proc. ICSLP*, pp. 129-132, 2004.

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

[6] K. K. Paliwal and L. D. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003.

[7] B.Bozkurt and L.Couvreur, "On the use of phase information for speech recognition," in *Proc. EUSIPCO*, Antalya, Turkey, 2005.

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

[9] D. Zhu and K. K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *Proc. ICASSP*, Montreal, Canada, May 2004.

[10] G.Duncan, B. Yegnanarayana and Hema A. Murthy, "A nonparametric method of formant estimation using group delay spectra," in *Proc. ICASSP*, pp. 572-575, May 1989.

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

[12] Aruna Bayya and B. Yegnanarayana , "Robust features for speech recognition Systems," in *Proc. ICSLP '98*, December 1998.

[13] P. Satyanarayana Murthy and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals," *IEEE Trans. on Speech and Audio Processing,* vol. 7, no. 6, November 1999.

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

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

[16] Available from http://spib.rice.edu/ spib/select_noise.html

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