THE ANNALS OF "DUNAREA DE JOS" UNIVERSITY OF GALATI
FASCICLE III, 2009, Vol.32, No.1, ISSN 1221-454X
ELECTROTECHNICS, ELECTRONICS, AUTOMATIC CONTROL, INFORMATICS

THE PHASE SPECTRA BASED FEATURE FOR ROBUST SPEECH RECOGNITION

Abbasian Ali, Marvi Hossien

Faculty of Electrical and Robotic Engineering
Shahrood University of Technology,
Shahrood, Iran
Ali_abbasin2002@yahoo.com , marvi_hossien@yahoo.co.uk

Abstract: Speech recognition in adverse environment is one of the major issue in automatic speech recognition nowadays. While most current speech recognition system show to be highly efficient for ideal environment but their performance go down extremely when they are applied in real environment because of noise effected speech. In this paper a new feature representation based on phase spectra and Perceptual Linear Prediction (PLP) has been suggested which can be used for robust speech recognition. It is shown that this new features can improve the performance of speech recognition not only in clean condition but also in various levels of noise condition when it is compared to PLP features.

Keywords: Group delay function, Phase Spectrum, Robust phoneme recognition .

1. INTRODUCTION

It is widely acknowledged that the performance of the current state-of-the-art speech recognizers starts to drop drastically in noisy conditions. It is hence clear that new technological breakthroughs are required for a major performance improvement. In order to make significant improvements, we need to acquire more basic knowledge in the area of feature extraction. As we know, modern speech recognizers still perform much worse than humans both in clean and noisy environments (P.Woodland, 1996). Modeling the complete human auditory system is, however, not possible since the system is only partially understood. Nevertheless, some parts of the system are known and can hence be utilized to improve the feature extraction unit. Spectral representation of speech is complete when both the Fourier transform magnitude and phase spectra are specified. In conventional speech recognition system, features are generally derived from the short-time magnitude spectrum while, the phase spectrum of the signal has been ignored. Recently, some features derived from phase spectrum have been suggested (Ray Schliitel, 2001; Guangji Shi, 2006). Often, the group delay function, which has properties similar to the phase, is studied (Rajesh M. Hegde, 2004). The group delay function has been used in earlier efforts to extract pitch and formant from speech signal reconstruction (Andrew C. Lindgren, 2003), and

This paper was recommended for publication by Rustem Popa

spectrum estimation. In all these efforts, no attempt was made to extract features from the speech signal and use them for speech recognition applications. Moreover, the cepstral features derived from the modified group delay function (MGDF) have been studied for speech recognition (H.A. Murthy, 2003). In this paper a new feature representation based on phase spectra and Perceptual Linear Prediction (PLP) has been suggested which can be used for robust speech recognition. The experiment show promising result. The rest of this paper is organized as followed. In section II, we review the conventional robust speech recognition methods. In section III, we described how can extract the features from the group delay function. The proposed method is introduced in section IV. Experimental results are given in section V.

2. ROBUST SPEECH RECOGNITION

As one of issues for the design of a robust speech Recognition system, the extraction of robust speech features should be considered. It is known that cepstrum data are usually corrupted by noise. Various noise robust methods have been developed such as noise-robust LPC analysis (Tierney J,1980), Hidden Markov Model (HMM) decomposition and composition (Gales M.J.F. and Young S.J,1993), (Martin F,1992),and the extraction of dynamic cepstrum, (Aikawa K. and Saito T,1994), (Aikawa K. and Hattori H,1996) etc. In spite of such research activities, the useful noise-robust techniques are still limited as a spectral subtraction (SS), Cepstral mea subtraction (CMS), RASTA and Perceptual Linear Prediction methods (PLP) (Boll S, 1979). Now we review some of these methods:

2.1. Spectral Subtraction (SS)

A lot of problems arise when q priori unpredictable ambient or electrical noise is present in the recorded signal. In order to be able to cope with that, it is usually assumed that the speech and the noise are additive and uncorrelated, and that the noise signal exhibits only slow variations relative to the speech signal. If this is true, one can estimate the noise spectrum during silent intervals and subtract this estimated noise spectrum from the signal spectrum during speech intervals. This technique which stems from the speech enhancement domain is called spectral subtraction. One problem with SS is that it is more thoroughly investigated in the context of speech enhancement than in the context of speech recognition. This means that one has to be very careful in blindly adopting results obtained from speech enhancement experiments. In fact, some deformations introduced by SS may be intolerable to the human ear but not very harmful for Recognizer, while other deformations maybe tolerated by the ear but not by the Recognizer. Another problem which needs further consideration is that an inaccurate spectral estimation of the noise power spectrum can result in negative power values which need to be set equal to a non-negative threshold. This non-linear operation produces residual noise commonly known as musical noise. The Signal-to-Noise Ratio (SNR) improvement is thus achieved at the expense of introducing other distortions into the speech signal, and it is to be seen how these may degrade the recognition results.

2.2. Cepstral Mean Subtraction (CMS)

If the frequency characteristic of channel is not flat, one will observe a signal which is obtained by convolving the original speech signal with the impulse response .One often says that the observed signal is corrupted by convolutional noise. However, if the channel characteristics vary only slowly in time compared to the characteristics of the speech signal, this noise can be considered multiplicative in the spectral domain. One may thus hope to suppress it in any feature space directly representing the log-spectrum of the signal. The simplest and most popular technique for doing this is Cepstral Mean Subtraction (CMS). Since the cepstrum does represent the log-spectrum of the signal, the convolutional noise is additive in the MFCC space. Therefore, CMS is based on the very simple assumption that the long-term average of the cepstrum can be estimated accurately on the basis of a few seconds of speech , and that the effects of the convolutional distortion will be removed by subtracting this long-term average from the original cepstra .

2.3. Realative Spectra (RASTA)

A generalization of CMS is RASTA (Realative Spectra) filtering. If each cepstral coefficient be as the sample of a time signal, this filter removes the low and high frequency modulations from this signal. The relative spectral (RASTA) technique proposed in (Hermansky, 1993) to enhance the temporal features was shown to increase the recognition performance with convolutional channel noise.

2.4. Perceptual Linear Prediction (PLP)

Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) are the most popular acoustic features used in speech recognition. Often it depends on the task, which of the two methods leads to a better performance. PLP features are reported (H. Hermansky, 1990) to be more robust when there is an acoustic mismatch between training and test data.

3. MAIN POINT OF GROUP DELAY

A brief summary of the methods used to extract group delay in speech is provided in this section.

Group delay is defined as the negative derivative of the phase of the Fourier transform of a signal. It is the time-domain delay of each frequency component of the signal, as a function of frequency. Let x(n) be a frame of digited speech and its Fourier transform is given by:

$$(1) \quad X(\omega) = \sum_n x(n) e^{-j\omega n}$$

The $X(\omega)$ can also be expressed as:

$$(2) \quad X(\omega) = |X(\omega)| e^{-j\theta(\omega)}$$

Then the Group delay is defined as in (3):

$$(3) \quad \tau(\omega) = -\frac{d\theta(\omega)}{d\omega}$$

In fact, the phase should be unwrapped before taking the derivative, which would be problem (K.K. Paliwal, 2007). To avoid unwrapping, another method calculates the group delay directly is computed from the speech signal that we can use logarithm in (2) as:

$$(4) \quad \log X(\omega) = \log(|X(\omega)|) + j\theta(\omega)$$

Then again we can write (3) as:

$$(5) \quad \tau(\omega) = -\mathrm{Im}\left(\frac{d(\log X(\omega))}{d\omega}\right)$$

Now it can be versus derivative as in (6):

$$(6) \quad \tau(\omega) = \left\{\frac{-X(\omega)_R \frac{d(X(\omega)_I)}{dw} + X(\omega)_I \frac{d(X(\omega)_R)}{d\omega}}{|X(\omega)|^2}\right\}$$

Where the subscripts R and I denote the real and imaginary parts. As differentiation can only be approximated in the discrete-time domain, another method is proposed in (Banno, *et al.,* 1998; Murthy, *et al.,* 1991) with the use of the following Fourier transform property (7) to avoid differentiation.

$$(7) \quad -jF\{nx(n)\} = \frac{dX(\omega)}{d\omega}$$

Where F denotes the Fourier transform. Separating the real and imaginary parts, we get

$$(8) \quad F\{nx(n)\}_I - jF\{nx(n)\}_R = \frac{dX(\omega)_R}{d\omega} + j\frac{dX(\omega)_I}{d\omega}$$

Using the above expression group delay as in (6) can be rewritten as in (9) :

$$(9) \quad \tau(\omega) = \left\{\frac{X(\omega)_R F\{nx(n)\}_R + X(\omega)_I F\{nx(n)\}_I}{|X(\omega)|^2}\right\}$$

If $Y(\omega)$ be the Fourier transform of nx(n), $F\{nx(n)\}$,and the subscripts R and I denote the real and imaginary parts.

We can rewrite Eq. (9) as:

$$(10) \quad \tau(\omega) = \left\{\frac{X(\omega)_R.Y(\omega)_R + X(\omega)_I.Y(\omega)_I}{|X(\omega)|^2}\right\}$$

As the group delay of speech suffers from spiky characteristics a major modification is proposed in is to use the cepstrally smoothed power spectrum. Then If we assume that speech is produced by a source-system model, the speech power spectrum, $|X(\omega)|^2$, can be expressed as the multiplication of the system component of the power spectrum, $S(\omega)^2$, with the source (or excitation) component of the power spectrum, $E(\omega)^2$:

$$(11) \quad |X(\omega)|^2 = S(\omega)^2.E(\omega)^2$$

The excitation contributes zeros near the unit circle which cause meaningless peaks in the GDF. The modified group delay function (MGDF), is formed by multiplying the GDF by the source component of the power spectrum:

$$(12) \quad \bar{\tau}(\omega) = \tau(\omega) E(\omega)^2$$

This operation gives less weight to peaks in the GDF which are the result of excitation-induced zeros near the unit circle. This is equivalent to replacing the denominator in Eq.10 with the system component of the power spectrum, $S(\omega)^2$:

$$(13) \quad \bar{\tau}(\omega) = \left\{\frac{X(\omega)_R.Y(\omega)_R + X(\omega)_I.Y(\omega)_I}{S(\omega)^2}\right\}$$

To further suppress the peaks, two new parameters (α and γ) were introduced in (Murthy, *et al.,* 1991) and the resulting group delay was named the modified group delay (MODGD) function by the authors, and is given in (8). The exact values of and can be determined experimentally, but the ranges suggested as $0 < \alpha < 1$, $0 < \gamma < 1$.

$$(14) \quad \bar{\tau}_{r,\alpha}(\omega) = sign\left|\frac{X(\omega)_R.Y(\omega)_R + X(\omega)_I.Y(\omega)_I}{S(\omega)^{2r}}\right|^{\alpha}$$

Fig.1 (a), (b) and (c) shows a frame of the fricatives sound /sh/, group delay (GDF) and modified group delay (MGDF), respectively. Before the Fourier transform, the speech signal in Fig.1 (a) has been multiplied with Hamming window. In Fig.1 (b), there are meaningless peaks and valleys in the GDF. It occurs due power spectrum in denominator in Eq. (10). It was shown that the spikes are caused by the Zeros of the speech signal which are close to unit circle. In Fig.1 (c), we can show the GDF meaningless peaks are lost and also MGDF has a rather flat envelope, which is caused by the presence of the smoothed power spectrum term in the denominator in Eq. (13).

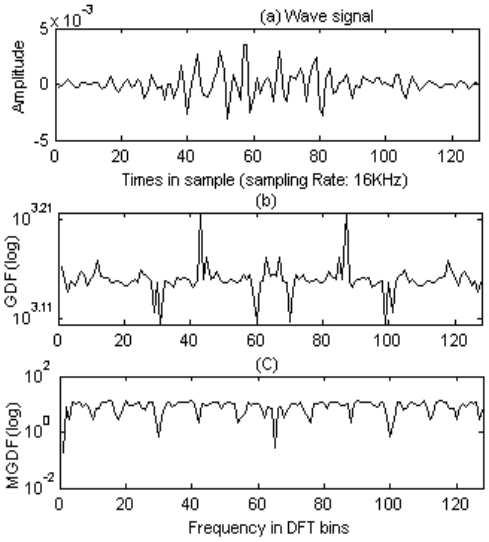

Fig1: (a) A frame of sound /sh/ ,
(b) Group delay function (GDF), modified group delay function (MGDF)

4. NEW FEATURE EXTRACTION ALGORITHM

The proposed feature are based on GDPPS and Bark scale, it can be obtained by the following stages:

4.1. At The first stage, we apply the Fourier transform on a pre-emphasized and hamming windowed speech signal.

4.2. The Group Delay Product Power spectrum (GDPPS) , $B(\omega)$ is obtained as follow:

$$(15)\quad B(\omega) = \left|X(\omega)\right|^2 .\tau(\omega)$$

where $\tau(\omega)$ is obtained in Eq.10. $B(\omega)$ is influenced by both the magnitude spectrum and the phase spectrum. Fig.2 shows GDPPS and MGDF. The MGDF has a small variation and also has a more flat envelope than GDPPS but the GDPPS has an envelope comparable to that of the power spectrum.

4.3. GDPPS cannot be used directly to train the phoneme recognition system, since the length of the vector is as long as that of the length of the DFT window size. Then now we apply Bark Filter bank over the GDPPS.

The three steps frequency warping, smoothing and sampling are integrated into a single filter-bank called Bark filter-bank. The Bark scale provides an alternative perceptually motivated scale to the Mel scale. Speech intelligibility perception in humans begins with spectral analysis performed by the basilar membrane (BM). Each point on the BM can be considered as a bandpass filter having a bandwidth equal to one critical bandwidth or one Bark (Ben J. Shannon, 2003). The bandwidth of several auditory filters were empirically observed and used to formulate the Bark scale as in (16):

$$(16)\quad Bark(f) = 6\log_e\left((f/600) + \sqrt{(f/600)^2 + 1}\right)$$

4.4. In this stage we used the intensity-to-laudness conversion, which raises the filter-bank outputs to the power of 0.33.This conversion, decreases the dynamic variability and also it is as a tuning of the spectral envelope approximation.

4.5. Finally, Cepstral features are derived, which can be decorrelated and relatively robust to channel mismatch and noise.

5. PERFORMANCE EVALUATION

5.1. Experimental Setup and Data Base

To evaluate the performance of proposed method, experiments have been performed on TIMIT database. This database is divided into training and testing sections. The experiments have been done an phonemes based speech recognition. In all 280 phonemes have been extracted from utterances and used on training set, while 140 phonemes are used for testing performance. In experiment different categories of phonemes such as vowels, semivowels, nasal, fricatives and steps are used. The noisy utterance is simulated by adding artificially generated white Gaussian noise to clean speech signal with various SNR levels by the following Equation:

$$(17)\quad \tilde{s}(m) = s(m) + kN(m)$$

$$k = \frac{\sum_{m=0}^{M-1} s^2(m)}{10^{\frac{SNR}{10}} . \sum_{m=0}^{M-1} N^2(m)} \qquad SNR = 10\log\left(\frac{\sum_{m=0}^{M-1} s^2(m)}{\sum_{m=0}^{M-1} N^2(m)}\right)$$

where $\tilde{s}, s$ and N represent noisy speech signal, clean speech signal and noise signal respectively, M

Is the length of s and SNR denotes the signal to noise ratio.

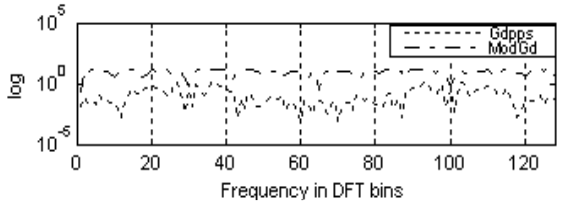

Fig.2: GDPPS and MGDF for a frame of sound /sh/

5.2. Evaluation of the Performance

In the first set of experiments, we used previous commonly speech recognition feature, Perceptual Linear Prediction (PLP). We computed PLP features with the bark filter-bank that consists of 22 asymmetrically-shaped filters.

In all cases, speech is pre-emphasised before analysis (Coeff. 0.97) and a Hamming analysis window of duration 16 ms is used, with 10 ms frame-shift.

The second set of experiments was conducted to evaluate the performance of the new feature, Bark Group Delay Product Power Spectrum (BGDPPS) which was described in Section 4.

The frame-rate is 9.3ms and frame-shift is 8ms for hamming windowing. We used 10 filters in bark filter-bank that obtained experimentally. In the third set of experiments, combination of the new feature with PLP are used which we denoted them as (PLP-BGDPPS).

Table1 shows the recognition of some phonemes for 5 test speakers. From the table 1, it can be seen that for the speaker one only the phoneme "w" has been misunderstand by the system. The worse case is for speaker 4 which 4 phoneme are misunderstood.

Table1: The recognition results of some phoneme for 5 speaker of the testing procedure

	s	**ao**	**r**	**w**	**d**	**k**	**m**	**ow**	**ay**
1	*s*	***aa***	*r*	***ay***	*d*	*k*	*m*	*ow*	*ay*
2	***sh***	***L***	***m***	*w*	***m***	***sh***	*m*	*ow*	*ay*
3	*s*	*ao*	*r*	*w*	*d*	***a o***	*m*	*ow*	***ae***
4	*s*	***ow***	*r*	***ow***	***g***	*k*	***n***	*ow*	***ao***
5	*s*	***aa***	*r*	***ao***	***m***	*k*	*m*	***ao***	*ay*

The overall performance of all experiment is shown in the Fig. 2. As can be seen from the figure, the proposed feature offers better accuracy compared to PLP.

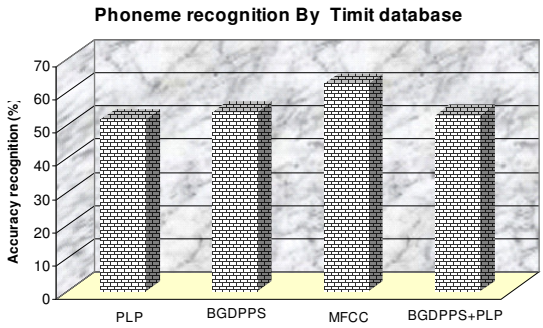

Fig3: The overall performance of method

Moreover as evident from this figure, the combination of new feature and PLP is also increased the accuracy of the original PLP about 2.2%. Also, according to the figure the MFCC (Mel scale frequency cepstrum) offers the best performance among all of the methods.

To verify the robustness of the features to noise the clean test utterance are corrupted with various levels of noise. The result for clean condition and various levels of noise are given in Table 2.

Table 2: Performance of different features on TIMIT database

SNR(dB)	**PLP**	**BGDPPS (proposed -feature)**	**PLP+BGDPPS**
-5dB	15	11.11	21
0 dB	15.78	22.2	21.1
5 dB	26	27.7	26.31
20 dB	31.5	33.3	42.1
clean	**51.64**	**53.80**	**52.8**

As can be seen from the table the proposed feature offers better accuracy compared to PLP not only in clean condition but also for various level of noise condition. Moreover as evident from the table the combination of new feature and PLP is also increased the accuracy of the original PLP for both clean and noisy.

6. CONCLUSION

This paper suggested a new feature representation based on phase spectra and Perceptual Linear Prediction (PLP) which is used for robust speech recognition. These new feature are derived from Group Delay Product Spectrum and Bark Scale which are combined with PLP. It is demonstrated that phase spectrum can improved the performance of speech recognition system not only in the clean condition but also in various level of noise condition.

7. REFERENCES

Alsteris L. D. and Paliwal K.K. (2007). *Short-time phase spectrum in speech processing: A review and so me experimental results*, Digital Signal Processing: A Review Journal, **vol. 17**, pp. 578-616.

Lindgren Andrew C. and Johnson Michael T., Povinelli Richard J. (2003). *Speech Recognition using Reconstructed Phase Space features*, pp.60-63, IEEE, ICASSP.

Aikawa K. and Saito T. (1994). *Noise robustness evaluation on speech recognition using a dynamic cepstrum*, IEICE Technical Report, **Vol. SP94-14**, pp. 1-8.

Aikawa K., Hattori H., Kawahara H. and Tohkura Y (1996). *Cepstral representation of speech motivated by time-frequency masking: an application to speech recognition*, J. Acoust. Soc. Am., **Vol. 100, No. 1**, pp. 603-614.

Boll S. (1979). *Suppression of acoustic noise in speech using spectral subtraction*, IEEE Trans. ASSP,ol. ASSP-27, No. 2, pp. 113-120.

Gales M.J.F. and Young S.J. (1993). *Cepstral parameter compensation for HMM recognition in noise*, Speech Communication, **Vol. 12, No. 3**, pp. 231- 239.

Guangji Shi, Maryam Modir Shanechi and Parham Aarabi (2006). *On the Importance of Phase in Human Speech Recognition*, IEEE Transactions on Acoustic, Speech and Language Processing, **Vol 14, No. 5.**

Gadde V. and Murthy H.A. (2003). *The modified group delay function and its application to phoneme recognition*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 68-71.

Hermansky, H., et al. (1992). *RASTA-PLP speech analysis technique*, Proc. of ICASSP'92, pp I-121-124, 1992.

N. Kawasaki, (1993). *Parametric study of thermal and chemical nonequilibrium nozzle flow*, M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan.

Hermansky H.(1990). *Perceptual linear predictive (PLP) analysis of speech*, Journal Acoust. Soc. Amer., **vol. 87, no. 4**, pp. 1738–1752.

Hegde R. M., Murthy H. A, Gadde V. Ramana Rao (2004). *Application of the Modified Group Delay function to Speaker Identification and Discrimination* , IEEE, ICASSP.

Martin F., Shikano K., Minami Y. and Okabe Y. (1992). *Recognition of noisy speech by composition of hidden Markov models*, IEICE Technical Report, **Vol. SP92-96**, pp. 9-16.

Lindgren Andrew C., Johnson Michael T., Povinelli Richard J. (2003). *Speech Recognition using Reconstructed Phase Space features*, pp. 60-63, IEEE, ICASSP.

Ney Hemann, Ray Schliitel (2001). *Using phase spectrum information for improved speech recognition performance*, pp.133-136, IEEE.

Paliwal Kuldip K., Shannon Ben J. (2003). *A Comparative Study of Filter Bank Spacing for Speech Recognition*, Microelectronic Engineering Research Conference.

Tierney J. (1980). *A study of LPC analysis of speech in additive noise*, IEEE Trans. on Acoust., Speech, and Signal Process., Vol. ASSP-28, No. 4 , pp.389-397.

Woodland P., Gales M. and Pye D., (1996). *Improving environmental robustness in large vocabulary speech recognition*, in Proc. ICASSP, **vol. 1**, pp. 65–68.

Yeganarayana B. and Murthy H. A., (1991). Speech processing using group delay functions, Signal Processing, **vol. 22**, pp. 259-67.