A NEW PHASE-BASED FEATURE REPRESENTATION FOR ROBUST SPEECH RECOGNITION

Erfan Loweimi[1], Seyed Mohammad Ahadi[1], and Thomas Drugman[2]

[1] Speech Processing Research Laboratory (SPRL), Electrical Engineering Department, Amirkabir University of Technology,
424 Hafez Ave, Tehran, Iran
[2] TCTS Lab, University of Mons, 31, Boulevard Dolez, B7000 Mons, Belgium

{eloveimi, sma}@aut.ac.ir, thomas.drugman@umons.ac.be

ABSTRACT

The aim of this paper is to introduce a novel phase-based feature representation for robust speech recognition. This method consists of four main parts: autoregressive (AR) model extraction, group delay function (GDF) computation, compression, and scale information augmentation. Coupling GDF with an AR model results in a high-resolution estimate of the power spectrum with low frequency leakage. The compression step includes two stages similar to MFCC without taking a logarithm of the output energies. The fourth part augments the phase-based feature vector with scale information which is based on the Hilbert transform relations and complements the phase spectrum information. In the presence of additive and convolutional noises, the proposed method has led to 15% and 12% reductions in the averaged error rates, respectively (SNR ranging from 0 to 20 dB), compared to the standard MFCCs.

Index Terms— Speech phase spectrum, feature extraction, group delay, compression, scale information

1. INTRODUCTION

In the majority of current speech processing systems, information is captured from the amplitude spectrum and the use of phase-related information is generally discarded. Three main reasons may explain why phase processing has been avoided. The first issue is some historical considerations that biased researchers to some extent against the phase spectrum. In 1843, Ohm [1] showed that human ear performs a Fourier analysis and only the magnitude spectrum is utilized in perception. This theory (Ohm's acoustic law) was also verified by Helmholtz in 1875 [2] and implies that human ear is phase deaf. The second reason is phase wrapping, a key problem with the phase spectrum, which results in an intractable, noise-like, and chaotic shape lacking any informative trend and meaningful extremum points. Dealing with this problem, researchers proposed unwrapping methods [3]-[5] and also tried to evade it by working with phase-derived representations such as group delay function (GDF). The third problem with the speech phase spectrum is that it has been shown that it does not contain much intelligibility information in short frames (20 to 40 ms) and by frame length extension its information content increases [6]-[14]. However, due to the relative non-stationary nature of speech, applying longer frame sizes does not make sense. Incidentally, the reason behind this trend was not discussed and it remained unjustified since 1979 [6].

Despite these drawbacks, three GDF-based features were recently proposed for Automatic Speech Recognition (ASR): the modified group delay function (MODGDF) [15], product spectrum (PS) [16], and chirp group delay function (CGDF) [17]. Recognition rates of these methods in the presence of additive noise are comparable to MFCC, although slightly worse in most cases. In the presence of a convolutional noise, their performance has been reported to be notably lower than that of MFCC [18].

Notwithstanding these advances, two fundamental questions about the speech phase spectrum remain open. First, why does the quality of phase-only reconstructed speech improve by frame length extension? Secondly, if it is really the fact that the phase spectrum of short frames (20 to 40 ms) is not informative, why are recognition rates of phase-based features comparable to those of magnitude-based features? In [19], we have shown that the reason behind both of these two basic issues is the Scale Incompatibility Error (SIE). Besides, we have demonstrated that in contrary to the prevalent belief, speech phase spectrum could be more informative than its magnitude counterpart, even in short frame lengths.

Relying on this idea, we first proposed a feature extraction method called ARGDD which is based on AR modeling and GD processing [18]. Its performance was quite satisfactory in the presence of both additive and convolutional noises. However, it did not benefit from any specific psychoacoustic finding. In addition, it did not deploy the SIE, despite its great importance. In this paper we further develop the ARGDD method notably by incorporating these steps. The performance of the resulting method over both additive and convolutional noises is strikingly higher than standard MFCC and other phase-based features.

The organization of this paper is as follows. In Section 2 we review the SIE and give an answer to the two aforementioned fundamental questions. Section 3 introduces the new method and examines its main properties. In Section 4 the recognition results on the Aurora 2 database are presented and thoroughly analyzed. Finally Section 5 concludes the paper.

2. SCALE INCOMPATIBILITY ERROR (SIE)

In this section we will deal with the two aforementioned fundamental questions and will show that their answer lies in Hilbert transform relations. As well, the high potential of the phase spectrum, particularly in short frame lengths, will be proved.

Hilbert transform relations determine the link between the phase and magnitude spectra of a minimum or maximum phase signal. For a speech signal *x(n)* whose Fourier Transform is denoted by $X(\omega)$, they are defined as follows [20]:

$$Ln|X(\omega)| = \hat{c}(0) + \frac{1}{2\pi}\rho\int_{-\pi}^{\pi} arg[X(\theta)]\cot(\frac{\omega-\theta}{2})\,d\theta \quad (1)$$

$$\hat{c}(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} ln|X(\omega)|\,d\omega, \quad (2)$$

where *Ln*, $\hat{c}(0)$, ρ, and *arg* indicate the natural logarithm, complex cepstrum in zero, Cauchy principal value of the integral, and unwrapped (continuous) phase spectrum, respectively. Although Eq. 1 does not work for a mixed-phase signal like speech, which contains both causal and anti-causal components in its complex cepstrum [20], it may be interestingly inspiring.

It is evident that multiplying a speech signal by a constant number has no psychoacoustic effect and just changes the intensity. Equation (1) shows that through phase spectrum one may reconstruct the signal up to a scale error. It implies that in phase-only signal reconstruction, the scale error that is introduced after reconstructing each frame is equal to *exp*($\hat{c}(0)$). Looking at this issue on an intra-frame basis, there is no serious problem since we will obtain each frame with a sole scale error, which does not affect the intelligibility information content of the frame.

On the other hand, it is obvious that the value of the complex cepstrum in zero for different frames is not identical. Consequently, each phase-only reconstructed frame will have a particular scale error. When they are brought together for synthesizing the main signal, the scale incompatibility problem imposes its adverse effect. The frames which should be overlapped and added together have no scale compatibility because each one suffers from a different scale error. This will negatively affect the quality and/or intelligibility of the synthesized signal. We call this error the *scale incompatibility error* (SIE) [19].

In [19] we removed the scale incompatibility by multiplying each phase-only reconstructed frame by *exp*($\hat{c}(m,0)$) (where *m* denotes frame index) and observed that the phase-only reconstructed signal in this situation has an interesting high quality. It should be emphasized that $\hat{c}(m,0)$ in each frame is just a constant number, which is used solely for properly joining the overlapping frames that are added to each other. In other words, it has only an inter-frame role and has no intra-frame importance. Therefore, the remarkable quality of the phase-only reconstructed speech is only rooted in the high information content of the phase spectrum.

Besides, we quantitatively showed that by extending the frame length this error decreases and after removing it, phase-only reconstructed speech will have a higher quality and intelligibility compared to its magnitude-only counterpart [19]. This is true for all frame lengths, including the shorter ones. Taking into account these points, answers may have been found for the two aforementioned questions. In conclusion, SIE might be an argument to contradict the prevailing belief by emphasizing the huge potential in using the phase spectrum.

3. THE PROPOSED METHOD

In this section we introduce the proposed method and discuss its main properties. Its workflow is displayed in Fig. 1. As seen, it consists of four main parts: autoregressive model (AR) extraction, group delay (GD) computation, compression, and scale information augmentation. The role of each step is discussed further ahead. The feature warping which is applied here is based on [21] and was originally proposed for speaker verification. Its influence on the recognition rates will be discussed in Subsection 4.3.

3.1. Coupling the Group Delay with the AR Model

Coupling the AR model with the group delay function has a number of advantages. Before describing them, it is necessary to briefly review the main properties of the GDF. It has two important properties which are additivity and high resolution. The former refers to the point that if two signals are convolved in time domain their GDFs will be added to each other, and the latter indicates the potential of this function in providing a high-resolution estimate of the power spectrum. The main problem with GDF is that when the zeros and/or poles of a signal get close to the unit circle, this function becomes overwhelmingly spiky and chaotic and consequently useless.

In the case of a speech signal, the excitation or source component introduces some zeros in the vicinity of the unit circle. So, the applicability of GDF will be vastly decreased and the aforementioned properties become unusable. MODGDF, PS, and CGDF were basically proposed for alleviating this problem. Another possible approach to overcome this issue is to only keep the vocal tract (filter) component of the speech signal and discard the excitation (source) contribution. Here, this is approximated by extracting the spectral envelope of the signal via AR modeling. It not only removes the adverse effect of the zeros but also paves the way for deploying the two properties of the GDF in a more efficient way.

Another benefit of this coupling is compression (dynamic range reduction) and adjusting the bandwidth of the formants. In MODGDF [15], the authors introduced two extra parameters in this regard. Finding the appropriate values for these parameters is not straightforward due to the lack of theoretical insight into their role. So, the optimum values may be found through a line search. Due to the relatively broad span of search and the high computational load of each speech recognition test, searching for optimum values, which lead to maximal recognition rates is very laborious. Furthermore, there is no guarantee that these values remain optimum choices under different databases, SNRs, and noise types. However, in the proposed method we just need to find the AR model order, which depends on the sampling rate, and its possible options are very limited.

Figure 2 depicts the behavior of different group delay-based representations for vowel /ee/ in both clean and noisy conditions. It is seen that MODGDF, PS, and CGDF are noticeably sensitive to additive noise. On the other hand, the power spectrum of the AR model (extracted via the LPC method) and consequently its GDF are more robust. However, the GDF of the AR model (AR+GDF) results in an estimate of the power spectrum with a higher resolution and less frequency leakage, which is very noteworthy. This could be exploited in a feature extraction method with high robustness and discriminability abilities.

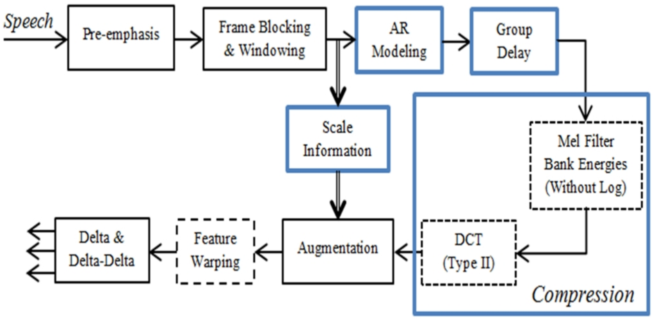
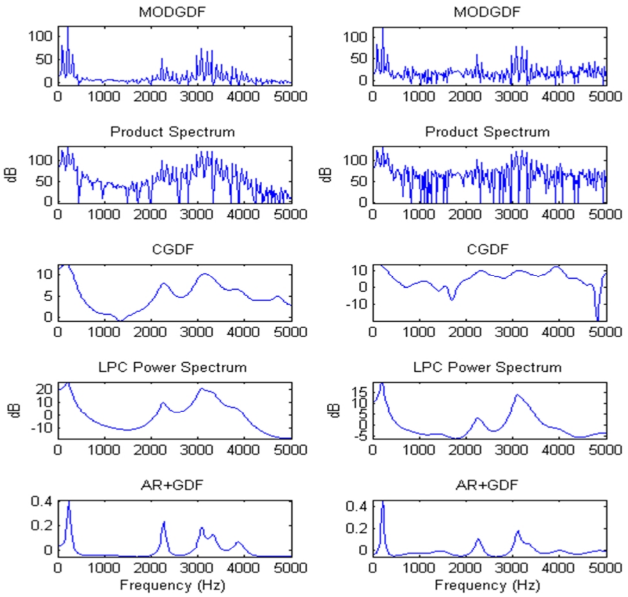


Fig. 1. Workflow of the proposed method.



(a) (b)

Fig. 2. Various group delay-based representations for vowel /ee/, (a) clean signal and (b) signal contaminated with white noise at 10 dB SNR.

3.2. Compression

The next important step is compression. In this stage, the power spectrum should be compressed in a rather low number of coefficients with minimum loss of vocal tract information. To do so, we have employed the Mel filterbank as well as DCT similar to what is done for MFCC. However, the logarithm applied to the output energies is not considered here. The reason is that the logarithm is used for converting the multiplication of the magnitude spectra of the source and filter into addition. This will smooth the way for separating these two components and preserve the filter component for performing speech recognition. However, multiplication of the Fourier transforms of two signals corresponds to the addition of their phase spectra and group delays (additive property). Therefore, the logarithm is not required. The same is true for CGDF and MODGDF-CC [22].

3.3. Scale Information Augmentation

In Section 2, we discussed the scale incompatibility error (SIE). Due to its importance, it should be taken into account in any phase-based speech processing task. It was shown that speech phase spectrum can provide a good estimate of each frame waveform, but scale information is missing [19]. It is reasonable to augment the phase-based features with such complementary information. Based on Hilbert transform relations we investigate augmenting the feature vector of each frame with the $exp(\hat{c}(0))$ of that frame. Since the target is not synthesizing the signal, we are not obligated to necessarily apply the $exp(\hat{c}(0))$. The possibility of using $\hat{c}$ (0) instead of its exponential form will be also addressed in Subsection 4.3.

It is worth highlighting two points at this stage. First, $\hat{c}(0)$ must not be confused with the c_0 coefficient, which is found after taking DCT from the output energies of Mel filterbank. Secondly, $exp(\hat{c}(0))$ does not change the GDF shape of the AR model and should therefore convey complementary information. Hereafter, we will call our proposed method ARGDMF.

4. EXPRIMENTAL EVALUATION

4.1. Database

The performance of ARGDMF was assessed on the Aurora 2 database [23]. The source speech for this corpus is TIDigits, consisting of connected digits, spoken by American English speakers, which is later downsampled to 8 kHz. It includes three test sets (A, B, and C) with SNRs varying from -5 dB to 20 dB by steps of 5 dB. Test sets A and B include additive noises while speech signals in test set C are contaminated with both additive and convolutional (MIRS [23]) distortions. We have used the clean-data training in all our experiments and HMMs were standardly trained with HTK [24].

4.2. Feature Extraction Settings

For all feature extraction techniques, we have used the default parameters reported in their respective publications. The frame width, frame shift, and number of filters of Mel filterbank are set to 32 ms, 12 ms, and 23, respectively. For the pre-emphasis coefficient, two options are investigated: a fixed value of 0.97, and an adaptive approach using $r(1)/r(0)$ as suggested in [25], where $r(n)$ is the autocorrelation sequence of the signal. For all techniques except for ARGDD and ARGDMF, a Hamming window has been used. For these two latter methods we used a Chebyshev window with dynamic range of 30 dB based on [19] and the AR model was extracted via the LPC method with an order of 12, considering the sampling rate of Aurora 2 (8 kHz). For all methods, 12 coefficients were used and cepstral mean normalization (CMN) was applied. In the following, we investigate in a step-by-step manner the influence of various blocks of the algorithm: pre-emphasis, window shape, compression method, and scale information. Table 1 summarizes different variants of ARGDMF.

4.3. Results and Discussion

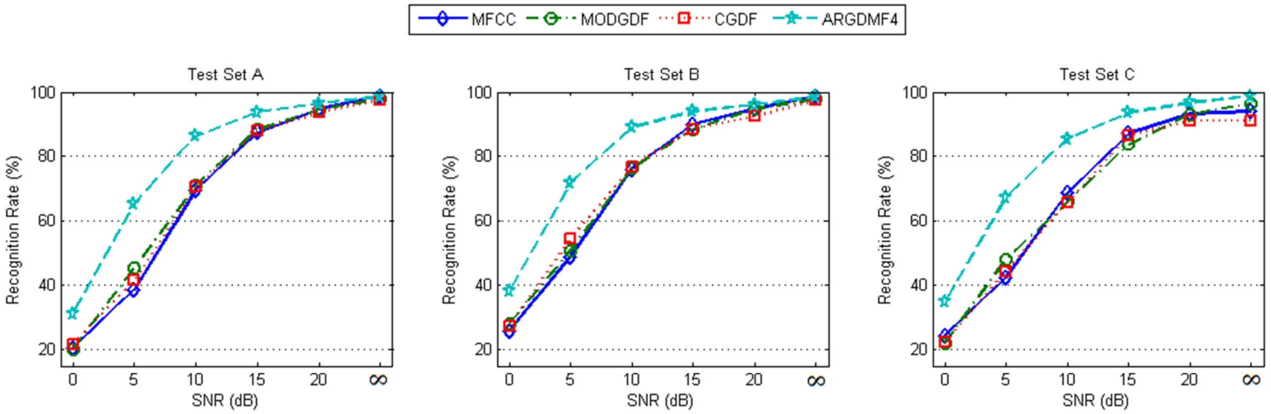
The performance of the compared techniques, averaged for SNRs varying from 0 to 20 dB, are presented in Tables 2 and 3. In the latter, the first and second derivatives are added, leading to a vector with 36 features. In general, the ARGDMF-based techniques are observed to clearly outperform the state-of-the-art.

First, the effect of an adaptive pre-emphasis is analyzed (results outside parentheses in Table 2 and 3). On test sets A and B, no notable differences are observed due to adaptive pre-emphasis. On the contrary, a clear improvement is noticed on test set C (except for MFCC and PS), with an increase of recognition rates up to 7%. Therefore, adaptive pre-emphasis may be considered as a complementary block for phase-based features in the presence of convolutional noises.

Table 1: Various forms of ARGDMF.

	Compression	Scale Information	Window Type	Warping
ARGDD	Double DCT	•	Ch. 30 dB	•
ARGDMF1	MFB + DCT	•	Ch. 30 dB	•
ARGDMF2	MFB + DCT	$exp(\hat{c}(0))$	Ch. 30 dB	•
ARGDMF3	MFB + DCT	$exp(\hat{c}(0))$	Hamming	•
ARGDMF4	MFB + DCT	$\hat{c}(0)$	Ch. 30 dB	•
ARGDMF5	MFB + DCT	$\hat{c}(0)$	Ch. 30 dB	ü

v MFB: Mel FilterBank,
v DCT: Discrete Cosine Transform
v Ch.: Chebyshev window



(a) (b) (c)

Fig. 3. Recognition rate as a function of SNR, after adaptive pre-emphasis (except for MFCC). In all cases, the feature vector consists of 36 coefficients including their static (12), Delta (12), and Acceleration (12) forms. (a) test set A, (b) test set B, (c) test set C.

The role of compression can be evaluated by comparing the performance of ARGDD and ARGDMF1. As seen, compressing the power spectrum through the application of a Mel filterbank followed by a DCT is relatively better than a double DCT (compression method used in ARGDD). This results in an absolute increase in the recognition rates by around 2%.

Inspecting the importance of considering the scale information (ARGDMF1 vs. ARGDMF2), it turns out from Tables 2 and 3 that employing scale information yields a notable overall improvement (up to 7.8% in absolute accuracies). This supports the justification of using scale information and highlights its complementary role.

A noteworthy point is that when using $exp(\hat{c}(0))$ as scale information (ARGDMF2), augmenting the feature vector with Δ and $\Delta\Delta$ coefficients almost does not alter recognition rates. Nonetheless, in the case of applying the scale information in the form of $\hat{c}(0)$ (ARGDMF4), including the temporal derivatives enhances the method. Although based on Hilbert transform relations, the proper representation for the scale information is $exp(\hat{c}(0))$, its sharp changes over adjacent frames, which are intensified after derivation, reduce the usefulness of augmenting the feature vector with dynamic coefficients. As a consequence, using the smoother $\hat{c}(0)$ contour seems to be more suited for a HMM-based modeling when Δ and $\Delta\Delta$ are intended to be applied.

Regarding the issue of choosing the appropriate window shape, we have shown in [19] that phase-only reconstructed speech with a Chebyshev window with dynamic range of 25 to 35 dB has led to the best quality. Results from Tables 2 and 3 (ARGDMF2 vs. ARGDMF3) corroborate that using an appropriate Chebyshev window has a considerable influence on the recognition abilities.

As a last factor, we inspected the effect of a feature warping method which was proposed originally for speaker verification [21]. The gap of performance after applying this technique (ARGDMF5) is striking: it achieves higher recognition rates with an absolute increase of 9% in Table 2 and of around 6% in Table 3.

Finally Fig. 3 shows the evolution of the performance as a function of SNR, for our ARGDMF4 technique and 3 representative state-of-the-art methods. For a fair comparison, results of ARGDMF5 are not displayed in Fig. 3 since feature warping can be applied to other phase-based methods and is not a block specific to our proposed technique. The improvement brought by the proposed approach becomes clear at SNRs below 20 dB. It is observed to significantly outperform other methods at low SNRs with an absolute raise of recognition rates up to 20%.

Table 2: Average (0-20 dB) word accuracy in percent*.

	TEST SET A	TEST SET B	TEST SET C
MFCC	56.5 (56.1)	59.8 (60.4)	54.4 (57.8)
MODGDF	59.4 (59.3)	62.2 (62.6)	58.1 (54.3)
MODGDF-CC	59.7 (56.0)	62.5 (61.4)	58.2 (52.0)
PS	56.3 (55.2)	60.0 (58.5)	54.7 (57.3)
CGDF	55.8 (55.0)	59.3 (59.0)	55.3 (47.1)
ARGDD	62.5 (62.2)	65.9 (65.7)	65.8 (61.5)
ARGDMF1	64.5 (64.6)	65.8 (66.1)	67.2 (60.5)
ARGDMF2	**72.3 (72.0)**	**73.3 (73.8)**	**72.6 (67.4)**
ARGDMF3	67.0 (66.9)	70.1 (70.6)	66.8 (64.5)
ARGDMF4	66.4 (64.1)	67.8 (67.1)	67.0 (57.3)
ARGDMF5	**75.2 (74.1)**	**76.3 (75.4)**	**76.1 (69.0)**

Table 3: Average (0-20 dB) word accuracy in percent*.

	TEST SET A	TEST SET B	TEST SET C
MFCC-D-A	62.7 (62.3)	66.9 (67.2)	60.0 (63.4)
MODGDF-D-A	64.0 (63.1)	67.7 (67.0)	62.6 (57.6)
MODGDF-CC-D-A	63.7 (61.2)	66.4 (68.1)	62.3 (58.3)
PS-D-A	63.4 (61.7)	67.0 (66.6)	60.2 (63.3)
CGDF-D-A	63.1 (62.3)	67.8 (67.2)	61.9 (55.1)
ARGDD-D-A	68.3 (67.2)	71.3 (71.2)	69.9 (65.0)
ARGDMF1-D-A	69.7 (69.1)	72.5 (72.8)	71.9 (66.3)
ARGDMF2-D-A	72.1 (71.2)	75.0 (74.9)	73.6 (67.9)
ARGDMF3-D-A	69.6 (68.2)	73.7 (73.8)	69.7 (67.3)
ARGDMF4-D-A	**74.7 (73.4)**	**77.9 (77.4)**	**75.6 (68.9)**
ARGDMF5-D-A	**81.0 (80.0)**	**81.7 (81.1)**	**83.8 (77.9)**

∗ Numbers outside and inside the parentheses correspond to adaptive and fixed pre-emphasis approaches, respectively.

5. CONCLUSION

In this paper we proposed a novel robust phase-based feature extraction algorithm for speech recognition. It consists of four main steps: autoregressive (AR) model extraction, group delay computation, compression, and scale information augmentation. The proposed method, called ARGDMF, was shown to substantially outperform the state-of-the-art in the presence of both additive and convolutional (channel) noises. This work therefore tends to prove that using appropriate phase-based features can yield a relevant representation of speech, with high applicability potentials among which a remarkable robustness.

6. REFERENCES

[1] G. S. Ohm, “Uber die definition des tones, nebst daran geknüfter theorie der sirene und ähnlicher tonbildender vorichtungen,” *Ann. Phys. Chem.,* vol. 59, pp. 513–565, 1843.

[2] H. L. F. von Helmholtz, *On the Sensations of Tone* (English translation by A.J. Ellis), Longmans, Green and Co., London, 1912 (original wsork published 1875).

[3] J. M. Tribolet, “A new phase unwrapping algorithm,” *IEEE Trans. Acoust. Speech, Signal Process*., vol.25, pp. 170–177, 1977.

[4] D. C. Ghiglia, M. D. Pritt, *Two-Dimensional Phase Unwrapping: Theory, Algorithms and Software*, Wiley, New York, 1998.

[5] G. Nico and J. Fortuny, “Using the matrix pencil method to solve phase unwrapping,” *IEEE Trans. Signal Process*., vol. 51, pp. 886–888, 2003.

[6] A. V. Oppenheim, J. S. Lim, G. E. Kopec, and S. C. Pohlig, “Phase in speech and pictures,” in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing,* pp. 632-637, Apr. 1979.

[7] A. V. Oppenheim and J. S. Lim, “The Importance of Phase in Signals,” *Proceedings of the IEEE*, Vol. 69, No. 5, pp. 529-541, May 1981.

[8] D. L. Wang and J. S. Lim, “The unimportance of phase in speech enhancement,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, pp. 679–681, Aug. 1982.

[9] L. Liu, J. He, and G. Palm, “Effects of phase on the perception of intervocalic stop consonants,” *Speech Communication*, Vol. 22, pp. 403-417, 1997.

[10] K. K. Paliwal and L. D. Alsteris, “Usefulness of phase spectrum in human speech perception,” in *proc. of Eurospeech-2003*, Geneva, Switzerland, pp. 2117–2120, 2003.

[11] L. D. Alsteris and K. K. Paliwal, “Importance of window shape for phase-only reconstruction of speech,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing*, Montreal, Canada, pp. 573–576, 2004.

[12] L. D. Alsteris and K. K. Paliwal, “Short-time phase spectrum in speech processing: A review and some experimental results,” *Digital Signal Processing*, Vol. 17, pp. 578-616, May 2007.

[13] E. Loveimi and S. M. Ahadi, "Objective evaluation of phase and magnitude only reconstructed speech: new considerations," in *Proc. 10th International Conference on Information Sciences Signal Processing and their Applications (ISSPA)*, Kuala Lumpur, Malaysia, pp. 117-120, May 2010.

[14] E. Loveimi and S. M. Ahadi, "Objective evaluation of magnitude and phase only spectrum-based reconstruction of the Speech signal," in *Proc. 4th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, Limassol, Cyprus, pp. 1-4, March 2010.

[15] H. A. Murthy and V. R. R. Gadde, “The modified group delay function and its application to phoneme recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Hong Kong, China, vol. 1, pp. 68–71, Apr. 2003.

[16] D. Zhu and K. K. Paliwal, “Product of power spectrum and group delay function for speech recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Montreal, Canada, vol. , pp. 125–128, May 2004.

[17] B. Bozkurt, L. Couvreur, and T. Dutoit, “Chirp group delay analysis of speech signals,” *Speech Commun.*, vol 49, no. 3, pp. 159-176, 2007

[18] E. Loweimi and S. M. Ahadi, “A new group delay-based feature for robust speech recognition,” in *Proc. IEEE Int. Conf. on Multimedia & Expo*, Barcelona, pp. 1-5, July 2011.

[19] E. Loweimi, S. M. Ahadi, and H. Sheikhzadeh, “Phase-only speech reconstruction using short frames,” in *Proc. InterSpeech,* Florence, Italy, pp. 2501-2504, August 2011.

[20] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing (3^{rd} edition)*, Prentice-Hall, Upper Saddle River, NJ, 2010.

[21] J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification,” in *Proc. A Speaker Odyssey, The Speaker Recognition Workshop*, Crete, Greece, pp. 243-248, 2001.

[22] B. Bozkurt, T. Dutoit, L. Couvreur, “Spectral analysis of speech signals using chirp group delay,” *In Progress in NonLinear Speech Processing. Y. Stylianou, M. Faundez-Zanuy, A. Esposito (Eds.),* Springer, pp. 41-58, 2008.

[23] H. G. Hirsch and D. Pearce, “The AURORA experimental framework for the performance evaluation of speech recognition Systems under noisy conditions,” in *Proc. ASR2000*, Paris, France, pp. 181-188, September 2000.

[24] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, The HTK Book Version 3.4., Cambridge University Press, Cambridge, Mass, USA, 2006.

[25] J. Makhoul, R. Viswanathan, “Adaptive preprocessing for linear predictive speech compression systems,” *Journal of Acoustic Society of America,* Vol. 55, pp. 475, 1974.