On the Importance of Pre-emphasis and Window Shape in Phase-based Speech Recognition

Erfan Loweimi[1], Seyed Mohammad Ahadi[1], Thomas Drugman[2], and Samira Loveymi[3]

[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
[3] Computer Engineering Department, Buali Sina University, Hamedan, Iran

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

Abstract. This paper aims at investigating the potentials of the phase spectrum in automatic speech recognition (ASR). We show that speech phase spectrum could potentially provide features with high discriminability and robustness. Out of such belief and to realize a higher portion of the phase spectrum potentials, we propose two simple amendments in two common blocks in feature extraction, namely pre-emphasis and windowing, without changing the workflow of the algorithms. Recognition tests over Aurora 2 indicate up to 11.2% and 14.7% performance improvement in average in the presence of both additive and convolutional noises for phase-based MODGDF and CGDF features, respectively. It proves the high potentials of the phase spectrum in robust ASR.

Keywords: Phase spectrum, speech recognition, feature extraction, discriminability, robustness, pre-emphasis, window shape.

1 Introduction

There is a general belief among the signal processing researchers that phase spectrum does not play a significant role in speech processing. Taking a glance on different areas of this field shows that only the magnitude spectrum is put under the center of attention. Phase spectrum is either directly transferred to the output without any processing (e.g. in speech enhancement) or discarded immediately after taking Fourier transform (e.g. in feature extraction for speech recognition).

Looking for the reasons behind the aversion toward speech phase spectrum, three issues could be found. In 19th century Ohm [1] and Helmholtz [2] stated that human ear performs Fourier analysis and only the magnitude spectrum is utilized in perception process. It implies that human ear is phase deaf. This misleading historical consideration, to some extent, biased the researchers against the sounds phase spectrum.

The second problem with phase spectrum, which seems to be the main one, is phase wrapping. It overwhelmingly complicates the interpreting and consequently

processing of the phase spectrum and creates a chaotic and noise-like shape lacking any meaningful trend or extrema points while the magnitude spectrum is much more understandable and matches well with our psychoacoustical knowledge.

The third problem is that it has been shown that speech phase spectrum is informative only in long frames while in short frames (20 to 40 ms) it does not carry notable deal of information [3]-[6]. Based on the current paradigms in signal processing, the non-stationary signals should be decomposed into short frames in which the stationarity assumption is held. As a result, working with long frame lengths does not make sense. Incidentally, this trend was remained unreasoned for about three decades.

Nonetheless, a number of phase (group delay)-based features such as modified group delay function (MODGDF) [7] and chirp group delay function (CGDF) [8] were proposed for automatic speech recognition (ASR). The recognition rates of these methods are comparable with MFCC in the presence of additive noise. However, channel noise distortion may highly degrade their performance. A missing point is that if this is really a fact that the phase spectrum is not informative in short frames, why are the recognition rates of the phase-based features comparable with those of the magnitude-based ones? This point also remained unaddressed and unreasoned.

In [9], we justified the two aforementioned questions and showed that, in contrast to the prevailing belief, speech phase spectrum is highly informative, even in short frames. This finding implies that much unexplored potential exists in the speech phase spectrum. In this paper, we aim at dealing with the possible capabilities of the phase spectrum in extracting strong features for ASR. We will show that this spectrum could potentially provide features with high discrimination abilities and robustness. After proving this point, we will propose two modifications in pre-emphasis and windowing stages, without changing the main workflow of the MODGDF and CGDF, aiming at realizing a higher portion of the phase spectrum potentials. Notable recognition rate improvements supports the idea that speech phase spectrum is worth more in ASR than what has been thought of it.

The rest of this paper is organized as follows. In Section 2 we will investigate the possible capabilities of the phase spectrum in speech recognition. In Section 3 two modifications which could provide more efficient usage of the phase spectrum information will be discussed. Section 4 includes the simulation results as well as their analysis and Section 5 concludes the paper.

2 Potentials of Speech Phase Spectrum in Feature Extraction

Usefulness of the phase spectrum in feature extraction, in the first degree, depends on both its information content and noise sensitivity in the short frame lengths. High information content and low noise sensitivity could potentially lead to features with high discriminability and robustness, respectively. The second concern is ambiguities in the behavior of this spectrum due to the phase wrapping because it complicates understanding and modeling of it. To some extent, this problem could be alleviated while working with GDF, since it can be computed without encountering the wrapping problem. It can also provide an estimate of the power spectrum which is an un-

derstandable and important function. However, GDF's spiky nature is an issue. MODGDF and CGDF are two possible solutions for dealing with this problem.

For checking potentials of the phase spectrum in providing features with high discriminability, we should determine its information content in short frame lengths. The information content of phase (or magnitude) spectrum could be evaluated by reconstructing the signal only from that spectrum. The quality and/or intelligibility of the reconstructed signal can be considered as an indicator of such information.

In [9], we have investigated this issue and shown that speech phase spectrum, even in short frame lengths, could be highly informative. In fact, we have shown that the quality and/or intelligibility of the phase-only reconstructed speech in all frame lengths, including short ones (16 and 32 ms), could be very high. It is an evidence for the capabilities of the speech phase spectrum in developing features with high discriminability. Moreover, the high recognition rates of the phase-based methods in the clean/matched condition could be justified considering this point.

The second issue, which is really challenging, is robustness. Although most of the features perform well in the clean/matched conditions, reduction of SNR leads to rapid degradation of their performance. In [10], we have investigated the sensitivity of the phase and magnitude spectra to (additive) noise and have shown that for speech signal decomposed into frame lengths of 32 ms, replacing the phase spectrum of noisy signal in 0 dB SNR, with its clean version could improve the quality up to 0.8 in PESQ scale. Similarly, substituting the magnitude spectrum with its clean version in the same situation could elevate the quality up to 2.1 in PESQ scale.

This observation may be interpreted in two ways. First, the quality of the signal primarily pertains to the magnitude spectrum and the phase spectrum's relation with the quality of the signal and consequently its importance is not as high as the magnitude spectrum. Second, the phase spectrum is less deviated from its clean version after contaminating the signal with noise whereas the magnitude spectrum is more sensitive to such disturbances. The first justification does not appear to be true since we have already shown that even in short frame lengths speech phase spectrum is highly informative. It seems that the second idea is the right case. This supports the high capabilities of the phase spectrum in providing more robust features in comparison with the magnitude spectrum due to its lower noise-sensitivity.

3 Possible Improvement for Phase-based Features

Despite the aforementioned potentials of the speech phase spectrum, the features which are extracted from it such as MODGDF [7] and CGDF [8] do not show eye-capturing performance. Although their discrimination abilities seem to be high, their robustness is not remarkable. However, based on the arguments presented in the previous section, it appears logical to expect to reach better recognition rates for the phase-based features. In other words, phase spectrum seems to have something more than what is captured by these features.

The neglected and important point which should be noted is that due to the predominant role of the magnitude spectrum in speech processing, common stages of

feature extraction algorithms such as pre-emphasis and windowing are based on the properties of the magnitude spectrum, not the phase spectrum. In this section, we will show that modification and adjustment of these two ostensibly simple blocks could lead to more efficient realization of the phase spectrum capabilities.

3.1 Pre-emphasis and Phase Spectrum

Generally, pre-emphasis is performed for flattening the magnitude spectrum and balancing the high and low frequency components. The point is that this task is defined based on the magnitude spectrum properties. However, the power spectrum which is estimated by the GDF, as depicted in Figure 1, is relatively flat. Therefore, pre-emphasis appears not to be a much needed block in phase-based speech processing. Nevertheless, since the magnitude-based paradigms are prevailed in speech processing, even in the case of phase-based features, pre-emphasis is used, without any modification. As illustrated in Figure 1, the group delay-based estimations of the power spectrum are relatively flat (far less negative slope) and pre-emphasis does not show any particular balancing influence. Hamming window is applied in this stage.

3.2 Window Shape and Phase Spectrum

Generally windowing is performed for getting a better smear-leakage trade-off. In [9], we have investigated the effect of 13 windows on the quality/intelligibility of the phase and magnitude-only reconstructed speech over different frames. In case of magnitude-only signal reconstruction, Hamming window results in the maximum quality. However, this window does not seem to be a good option for working with the phase spectrum since the quality of the phase-only reconstructed signal after applying it was quite poor. We observed that Chebyshev window with dynamic range of 25 to 35 dB results in the maximum quality for the phase-only reconstructed signal.

Changing the window from Hamming to Chebyshev (25 dB) in frame length of 32 ms, improved the quality of the phase-only reconstructed speech up to 1.4 in PESQ scale [9] which is quite significant and proves the impact of the windowing in working with phase spectrum. However, despite the notable influence of the window shape and unsuitability of the Hamming window, in all of the phase-based features extraction methods, such as MODGDF and CGDF, this window is applied. It appears that utilizing more suitable windows could be considered as a factor which may help in reaching more effective realization of the speech phase spectrum potentials.

Figures 2 and 3 depict the magnitude spectrum, MODGDF, and CGDF after applying rectangular and Chebyshev windows (30 dB), respectively, with and without pre-emphasis. As seen, in case of Chebyshev window, more distortion is introduced and only around the formant frequencies notable activities could be observed. It could have both positive and negative outcomes. In clean condition the introduced distortion may negatively affect the performance. On the other hand, it potentially could help in better retaining of the formants frequencies structure and also to some degrees alleviates the noise influence on the power spectrum. As a result, it could lead to more robust features. Our recognition test results interestingly verify these points.

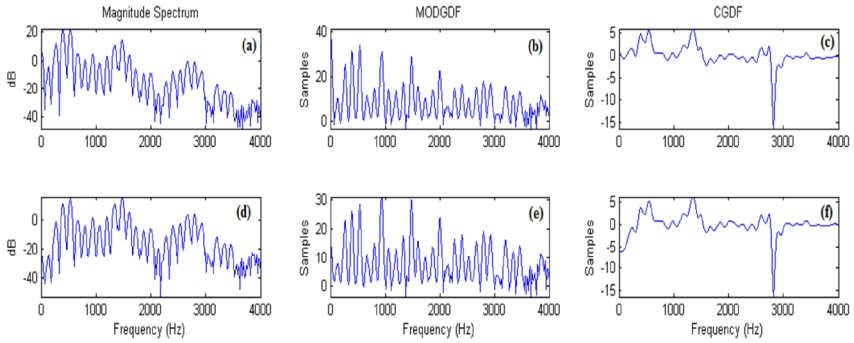

Fig. 1. Influence of application of pre-emphasis and Hamming window on the magnitude spectrum, MODGDF, and CGDF. (a), (b), (c) without pre-emphasis, (d), (e), and (f) with pre-emphasis (0.97).

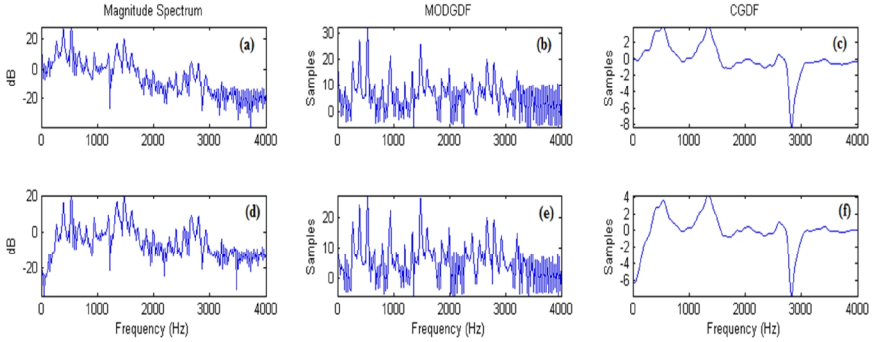

Fig. 2. Influence of application of pre-emphasis and Rectangular window on the magnitude spectrum, MODGDF, and CGDF. (a), (b), (c) without pre-emphasis, (d), (e), and (f) with pre-emphasis (0.97).

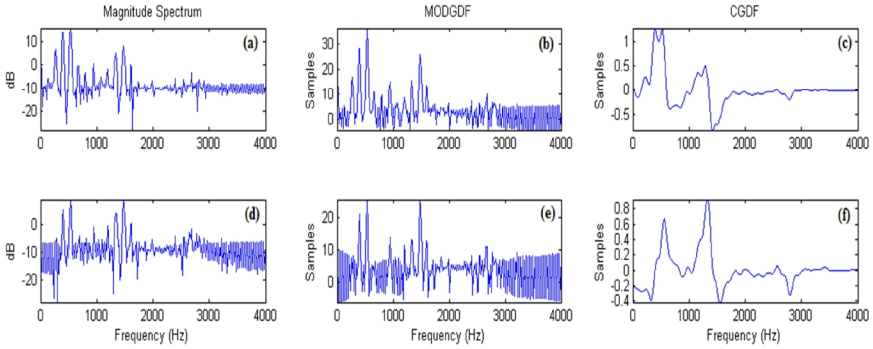

Fig. 3. Influence of application of pre-emphasis and Chebyshev window on the magnitude spectrum, MODGDF, and CGDF. (a), (b), (c) without pre-emphasis, (d), (e), and (f) with pre-emphasis (0.97).

4 Experimental Evaluation

4.1 Dataset and Feature Extraction Setting

Performance of the features is assessed on the Aurora 2 database [11]. It includes three test sets (A, B, and C) with SNRs varying from -5 dB to 20 dB by steps of 5 dB. A and B test sets include additive noises while speech signals in C test set are contaminated with both additive and channel distortions. We used the clean-data training in all our experiments and standard training of HMMs carried out with HTK [12].

For feature extraction techniques, we have used the default parameters reported in their respective publications. For investigating the effect of pre-emphasis, we have checked the effect of applying 0.97 ($a_{0.97}$), zero ($a_{0.0}$), and $r(1)/r(0)$ ($a_{r1/r0}$) [13] (where $r(n)$ is autocorrelation of the signal) as pre-emphasis coefficient. For examining the influence of window, Hamming, Rectangular, and Chebyshev (30 dB) windows are applied. Feature vector consists of 36 elements including 12 static coefficients as well as Δ and $\Delta\Delta$ components. CMN is also performed in all cases. Table 1 shows the average of the recognition rates of 20, 15, 10, 5, and 0 dB SNRs in percent.

4.2 Results and Discussion

As seen in Table 1 and Figure 4, modifying the pre-emphasis and window can notably affect the performance of both phase and magnitude-based features. Results should be compared with those of applying 0.97 as pre-emphasis coefficient and Hamming window which are displayed in italic and underlined form in Table 1.

For phase-based features (MODGDF and CGDF) both pre-emphasis and window shape seem to be influential. As previously mentioned, phase-based features return poor results in the presence of the convolutional noises. However, by applying adaptive pre-emphasis (r_1/r_0) as well as appropriate window this shortcoming is highly alleviated and their performance in the presence of both additive and channel noise notably improves. As seen, over the C test set, performance is elevated up to 11.2% and 14.7% for MODGDF and CGDF, respectively. This is quite remarkable and proves that there is much potential in the phase spectrum which has remained unused. For realizing it, we should rethink some prevalent facts and paradigms, however.

Contrary to MODGDF and CGDF, pre-emphasis is not very influential in case of MFCC and discarding this block even seems to be a better choice, especially in the presence of channel noise. For the C test set in which the high frequency components are strengthened to some extent by MIRS filter [11], further amplifying these components by pre-emphasis does not appear to be an appropriate action. That is why discarding this block in this case leads to better results. On the other hand, changing the window has a noticeable positive effect on the performance of MFCC. Chebyshev window (30 dB) clearly improves the performance of MFCC and results in interesting recognition rates. In addition, applying Rectangular window without performing pre-emphasis leads to good results and practically, it is a more economical choice.

Another interesting point is that based on our previous study [9] in which the Hamming window results in maximum quality in the magnitude-only speech recon-

struction, we were expecting maximum recognition rates for magnitude-based features using this window. However, as seen in Table 1, this is not the case. It shows that the required smear-leakage trade-off which is expected to be provided by the window, not only depends on whether we are working with the magnitude or phase spectrum, but also depends on the task. In speech reconstruction all the information either related to vocal tract or excitation component are important whereas in recognition only a specific part of the speech data are significant. So, the window which more helps in capturing the required information leads to higher performance.

Figure 4 depicts comparison between the recognition rates of the traditional scenario (0.97 + Hamming) with those of our proposed scenario ($\frac{r1}{r0}$ + Chebyshev (30 dB)) versus SNR for different test sets. As seen, pre-emphasis and window shape could notably affect the robustness of both phase-based and magnitude-based features. It should be noted that in clean condition traditional scenario slightly works better. However, by decreasing the SNR, the benefits of our modifications stand out. The last point is that these blocks do not provide any new information and only pave the way for employing the phase and magnitude spectra information in a more efficient way.

Table 1. Average (0-20 dB) word accuracy in percent.

		Test Set A			Test Set B			Test Set C		
		$a_{0.0}$	$a_{0.97}$	$a_{r1/r0}$	$a_{0.0}$	$a_{0.97}$	$a_{r1/r0}$	$a_{0.0}$	$a_{0.97}$	$a_{r1/r0}$
MFCC	*Ham.*	62.8	*62.3*	62.7	67.3	*67.2*	66.9	64.8	*63.4*	60.0
	Rect.	68.9	65.2	65.4	71.7	68.4	69.6	71.9	69.6	67.3
	Ch(30dB)	69.6	70.3	**71.1**	68.2	73.0	**73.1**	**75.6**	68.3	73.4
Max Improvement (%)		**+ 8.8%**			**+ 5.9%**			**+12.2%**		
MODGDF	*Ham.*	59.2	*63.1*	64.0	60.4	*67.0*	67.7	52.3	*57.6*	62.6
	Rect.	64.6	66.0	66.4	65.4	69.6	69.8	64.5	62.6	67.4
	Ch(30dB)	64.0	68.1	**69.9**	62.6	70.1	**71.2**	67.1	62.3	**68.8**
Max Improvement (%)		**+ 6.8%**			**+ 4.2%**			**+ 11.2%**		
CGDF	*Ham.*	66.1	*62.3*	63.1	67.4	*67.2*	67.8	68.5	*55.1*	61.9
	Rect.	67.5	63.0	64.2	68.4	68.3	68.9	69.6	55.8	62.5
	Ch(30dB)	66.4	67.2	**69.0**	65.6	71.1	**72.1**	**69.8**	58.1	66.9
Max Improvement (%)		**+ 6.7**			**+ 4.9%**			**+ 14.7%**		

5 Conclusion

The main target of this paper was investigating the potentials of the phase spectrum in speech recognition. We showed that phase spectrum could result in features with high discriminability and robustness. Out of such idea, we proposed some modifications in two common blocks of feature extraction algorithms aiming at reaching more efficient realization of the phase spectrum potentials. The recognition tests results indicated that pre-emphasis and windowing could notably affect the performance of the phase-based (and also magnitude-based) features. Looking for novel models which explicate the phase spectrum behavior is a broad avenue for future researches.

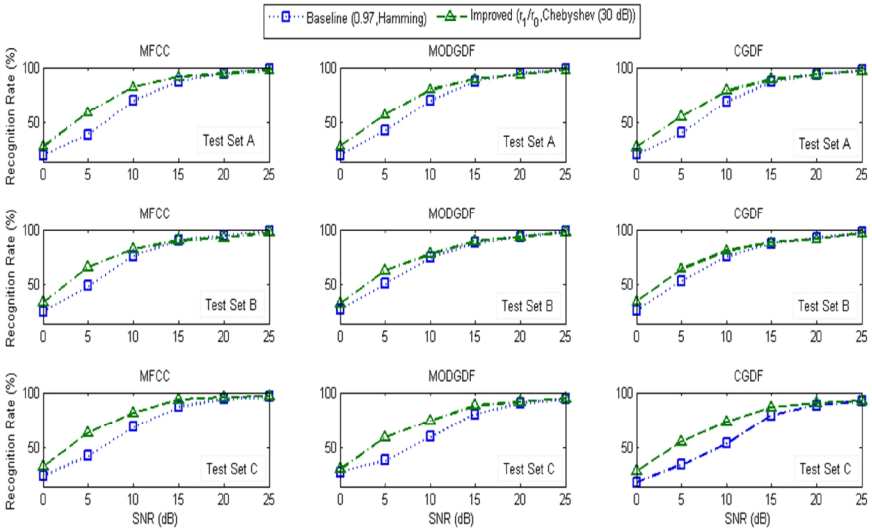

Fig. 4. Performance improvement after modifying pre-emphasis and window shape vs. SNR.

References

1. G. S. Ohm, Uber die Definition des Tones, nebst daran geknupfter Theorie der Sirene und ahnlicher tonbildender Vorrichtungen, Ann. Phys. Chem. 59 (1843) 513–565.
2. H. L. F. von Helmholtz, On the Sensations of Tone (English translation by A.J. Ellis), Longmans, Green and Co., London, 1912 (original work published 1875).
3. A. V. Oppenheim, J.S. Lim, The importance of phase in signals, Proc. IEEE 69 (1981) 529–541.
4. D. L. Wang, J.S. Lim, The unimportance of phase in speech enhancement, IEEE Trans. Acoust. Speech Signal Process. ASSP-30 (4) (1982) 679–681.
5. L. Liu, J. He, G. Palm, Effects of phase on the perception of intervocalic stop consonants, Speech Commun. 22 (4) (1997) 403–417.
6. K. K. Paliwal and L. D. Alsteris, Usefulness of phase spectrum in human speech perception, in: proc. of Eurospeech, September 2003, pp. 2117–2120.
7. H. A. Murthy, V. Gadde, The modified group delay function and its application to phoneme recognition, in: Proc. ICASSP, April 2003, pp. 68–71.
8. B. Bozkurt, L. Couvreur, and T. Dutoit, "Chirp group delay analysis of speech signals," *Speech Commun.*, vol 49, no. 3, pp. 159-176, 2007.
9. E. Loweimi, S. M. Ahadi, and H. Sheikhzadeh, Phase-only speech reconstruction using short frames, in: Proc. InterSpeech, 2011, Florence, Italy.
10. E. Loweimi, S. M. Ahadi, and S. Loveymi, On the importance of phase and magnitude spectra in speech enhancement, in: Proc. ICEE , Tehran, Iran, May 2011.
11. 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, Sep. 2000.
12. 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.
13. J. Makhoul, R. Viswanathan, Adaptive preprocessing for linear predictive speech compression systems, Journal of Acoustic Society of America, 55 (1974), 475.