INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Phase based spectro-temporal features for building a robust ASR system

Anirban Dutta, G Ashishkumar, Ch V Rama Rao

National Institute of Technology Meghalaya
anirbandutta118@nitm.ac.in, g.ashishkumar@nitm.ac.in, chvramarao@nitm.ac.in

Abstract

Spectro-temporal feature extraction has shown its robustness in the field of speech recognition. However, these features are derived from magnitude spectrum of the complex Fourier Transform (FT). In this work, we investigate to see if phase information can substitute magnitude based spectro-temporal features. We compared with different state of art phase spectrum and evaluated its performance. The experiments are carried out in different noisy environments. We found Modified Group Delay (MODGD) spectrum to closely resemble the structure of power spectrum. A relative performance difference of 0.03% on average is observed for the MODGD spectro-temporal features compared to the magnitude based features. The analysis showed that phase can indeed carry equivalent or complementary information to magnitude based spectro-temporal features.
Index Terms: phase, Gabor, spectro-temporal, recognition

1. Introduction

Most of the state of art feature extraction methods of ASR system uses the spectral information of speech. The temporal cues are also gaining importance where the characteristic of one time frame depends on the other. Thus, the true feature content is conveyed both by the spectral and temporal cues of speech. Harmonics of sound units, transient from one sound unit to another are some of the features captured by spectral and temporal representations. In this context, the combined spectro-temporal features can capture the different acoustic features of speech. A physiologically and psycho-acoustically motivated filter known as Gabor filters were developed which are known to best extract such features. Previous studies have demonstrated the capability of these features for various speech applications [1]. The fundamental step in the extraction of Gabor based spectro-temporal features is the convolution of the two dimensional Gabor filters with a time-frequency representation of the signal. The spectro-temporal representation should also be robust enough to discard the unwanted speech attributes. The state of art spectro-temporal features uses the image like short-time magnitude spectrum over which Mel filters are applied to form the Mel spectrogram [2]. It is seen that the complex phase spectra is often discarded, even for the state of art spectro-temporal feature extraction methods. The common trend is to remove the phase component even though it is well known that phase is equally important for signal processing [3]. The research to use phase information has been demotivated by earlier studies conducted by [4]. They conducted several experiments with phase and documented the incapability of human ears to distinguish the phase of the incoming signal. This perception remained uninvestigated until [5] reviewed the importance of phase spectrum in human perception through different hearing models. Their work was further carried out by [6] who established that phase indeed carries information and should not be neglected. However, the implementation of signal processing with short term discrete phase spectrum suffers from certain mathematical issues associated with it [7]. This in turn leads to phase wrapping due to which phase information does not convey any meaningful information. Due to the problems of phase wrapping and hence the distorted phase spectrum, phase data is often removed from the samples. Recently, however, there is an upsurge of interest in phase research for different speech technologies. These include speech and speaker recognition, speech enhancement, source separation, speech coding. A detailed review on the recent advancements of phase in speech applications is highlighted in [8]. Since, direct computation from the phase spectrum does not convey any meaningful speech intelligibility; there is a need to transform the short time phase spectrum into a domain which can extract distinct acoustic phase based features. The most useful phase derived representations are the negative derivative of the phase response called the group delay functions. Several modifications of group delay functions are proposed over the last few years to extract meaningful attributes of speech from phase. These phase representations exploit different properties of speech such as resonance, formants, harmonicity etc. It is believed that phase signal processing can provide complementary or equivalent information to magnitude based methods. Whether this is true for spectro-temporal features also; is the study carried out in this paper.

The focus of this work is to examine whether phase intelligibility can be incorporated in spectro-temporal feature extraction methods. If yes, then which representation suits the most in the context of spectro-temporal based features. Moreover, in the very first step of signal processing if the resolution of the unique formant structure is enhanced then the post-processing filtering operation can extract the minute discriminable features of the pattern more adequately. Different phase derived representations are considered here for evaluation. The robustness of each of these representations is analyzed in the context of spectro-temporal features. The rest of the paper is divided into the following sections: section 2 highlights the different phase derived features, section 3 gives the overall spectro-temporal feature extraction process and section 4 briefly outlines the experimental setup. The results are explained in section 5 followed by conclusion in section 6.

2. Phase derived representations

Speech signal is the convolution of the excitation source and the vocal tract characteristics according to the speech production model. These are represented by the envelope and fine structure in the speech spectrum respectively. In the feature extraction process using magnitude spectrum, the goal is to extract the envelope for further processing. The poles of the Z transform of speech correspond to the formants or vocal tract characteristics. In the case of phase spectra using group delay also, the envelope shape needs to be extracted. However, the presence of zeros (arising out of nasals, glottal closure, extent of window lengths) creates unwanted spikes in the spectrum. The effect

http://dx.doi.org/10.21437/Interspeech.2020-2258

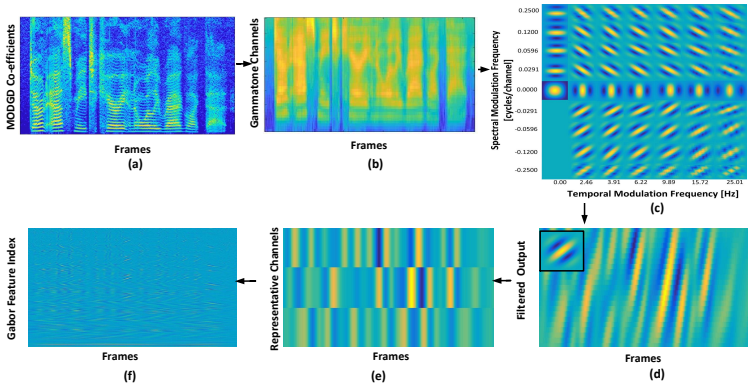


Figure 1: *(a) MODGD phase spectrum (b) Gammatonegram (c) Gabor filter bank (d) Convolution of gammatonegram with Gabor filter (e) Representative channels for one Gabor filter output (f) Final Gabor features*

of zeros needs to be suppressed by means of pushing the zeros radially into the unit circle. Thus, a modification is required in the group delay function for speech signals that can resemble the spectra of a minimum phase signal [9]. Modified group delay, chirp group delay, all pole stabilized parametric models are some of the phase derived modifications which has shown its importance in the speech community over the last few years. A brief theory of each of these methods is outlined here.

2.1. Modified Group Delay (MODGD)

The speech signal represented in terms of its magnitude and phase components of its FT is given by Eq.1.

$$X_n(\omega) = |X_n(\omega)|\, e^{j\theta_n(\omega)} \tag{1}$$

Group delay is defined as the negative derivative of the phase of the complex FT of the signal as given in Eq.2.

$$\tau(\omega) = -\frac{d(\theta_n(\omega))}{d\omega} \tag{2}$$

[10] suggested a modification to this function as given in Eq.3, which they termed MODGD. The basic idea was to smooth the magnitude spectrum by cepstral processing and introduce additional parameters to control the effect of spikes. MODGD is capable of pushing the zeros of the Z transform of speech signal radially inward into the unit circle and emphasize only the contributions from the poles of the vocal tract.

$$MODGD(\omega) = \left(\frac{\tau_p(\omega)}{|\tau_p(\omega)|}\right)(|\tau_p(\omega)|)^{\alpha} \tag{3}$$

$$\text{where, } \tau_p(\omega) = \left(\frac{X_R(\omega)Y_R(\omega)+X_I(\omega)Y_I(\omega)}{S(\omega)^{2\gamma}}\right)$$

Here, $S(\omega)$ is the cepstral smoothed variant of the magnitude spectrum with ρ ans γ as additional parameters. R and I denotes the real and imaginary components. $X(\omega)$ and $Y(\omega)$ are the FT of $x[n]$ and $nx[n]$ respectively.

2.2. Chirp Group Delay (CGD)

An alternative representation of group delay was proposed by [11] known as Chirp Group Delay. It is obtained by computing the negative derivative of the phase on a circle other than the unit circle (chirp) after transforming the signal into its Zero Phase form. The analysis circle in chirp group delay ensures certain distance from zeros of the Z transform of the signal i.e. spike freeness. The Chirp Z transform of the signal is given by Eq.4

$$\widetilde{X}(\omega) = X(z)|_{z=\rho e^{j\omega}} = \left|\widetilde{X}(\omega)\right| e^{j\widetilde{\theta}(\omega)} \tag{4}$$

Here, ρ represents the radius of the analysis circle. The CGD is given as the negative phase derivative of this transformation as given in Eq.5

$$CGD(\omega) = -\frac{d(\widetilde{\theta}(\omega))}{d\omega} \tag{5}$$

2.3. All Pole Group Delay (APGD)

Considering the vocal tract as an All Pole filter, the speech spectrum can be approximated with the help of Linear Prediction analysis [12]. The linear prediction is computed as given in Eq.6

$$H(\omega) = \frac{G}{1 - \sum_{m=1}^{p} a(m)e^{-j\omega m}} \tag{6}$$

Where G is the signal dependent gain and p is th order of the model. The prediction co-efficients $a(m)$ are determined by the method of least square sense. The filter response of $H(\omega)$ consists both of magnitude and phase component. The group delay from all pole model is computed as the negative derivative of the phase response of the filter. The parametric model based on stabilized weighted linear prediction is considered in this work for the Linear Prediction analysis [13]. This method has recently shown its robustness for a variety of speech applications.

3. Spectro-temporal feature extraction

The series of steps followed in the extraction of spectro-temporal features is taken from [14] and is shown in Figure 1. A brief description of each process is elaborated in the following section.

I The first step is to obtain an efficient time-frequency representation of the speech signal. The speech signal is first divided into small segments of 25 ms of frames with an overlap window of 10 ms. The spectrum analysis is done by transforming the quasi-stationary signal using the phase spectra of short-time FT. Here, group delay based phase information (MODGD) is used to obtain the short time phase spectra. A 40 channel gammatone filter bank is applied to this spectral representation per frame. Gammatone filters are based on Equivalent Rectangular Bandwidth (ERB) scale which provide better estimation of the human basilar membrane. They are characterized by smooth magnitude response with asymptotic decreasing skirts that represents the spectro-temporal informations much better. The frequency range of gammatone filters with center frequencies between 130 Hz to 6800 Hz is chosen. Logarithmic scale is applied to the gammatone filter output to form the spectro-temporal representation called gammatonegram.

II In order to extract the spectro-temporal features, two dimensional Gabor filter bank is created using the concept of constant Q analysis. A bank of 59 filters with varying spectral and temporal modulation frequencies is obtained that ensures uniform geometrical spacings with less correlation between the filters. Each Gabor filter is designed to select the patterns that tunes to the pair of spectral and temporal modulation frequencies. Each of these filters are convolved with gammatonegram to represent the spectro-temporal cues that matches the specific Gabor filters. Here, only the real part of the complex Gabor filter bank is used to extract the spectro-temporal features of speech.

III The resulting spectro-temporal features from all the filters of the Gabor filter bank leads to a high dimensional feature vector (59 X 40). For this, a dimension reduction is chosen where the features corresponding to 1000 Hz frequency channel and channels at a quarter of the current filter spectral width is chosen. Finally, the reduced phase based spectro-temporal feature vectors from each of the filters of the Gabor filter bank is concatenated. This creates the final spectro-temporal feature vector used as features for training the DNN network of the acoustic module of hybrid ASR system.

4. Setup of ASR acoustic module

The ASR corpus contains speech waveforms along with their text transcriptions. Forced alignment is used using the GMM-HMM approach to align the speech waveforms with their respective phonetic representations. For supervised learning of Deep Neural Network (DNN) using the spectro-temporal features, the target labels are generated by mono-phone modeling and tri-phone modeling of the standard GMM-HMM baseline [15]. Hybrid DNN-HMM is used to build the acoustic module of the ASR system [16]. The synaptic weights of the network are optimized using Natural Gradient Stochastic Gradient Descent (NG-SGD) algorithm [17]. After the final iterations of training, the model configuration of DNN is used in the hybrid ASR system to build its acoustic module.

5. Results

In this section, we investigate the possibilities of using phase spectrum to validate if it carries any complementary or equivalent information to magnitude based features suitable for ASR feature extraction. To validate this, we use the dataset of

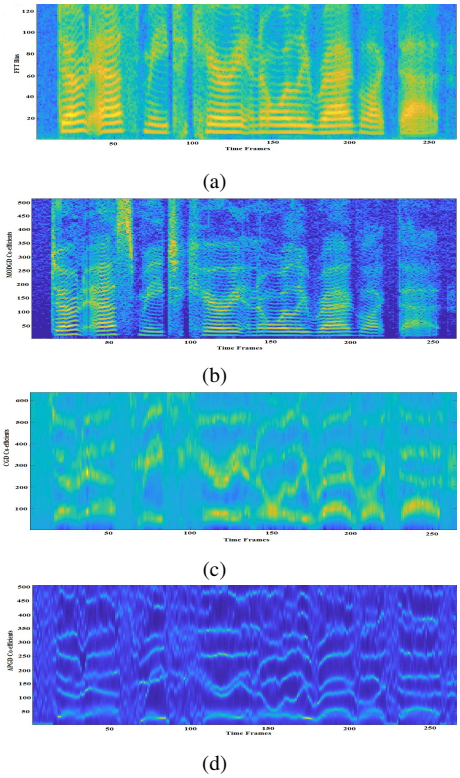
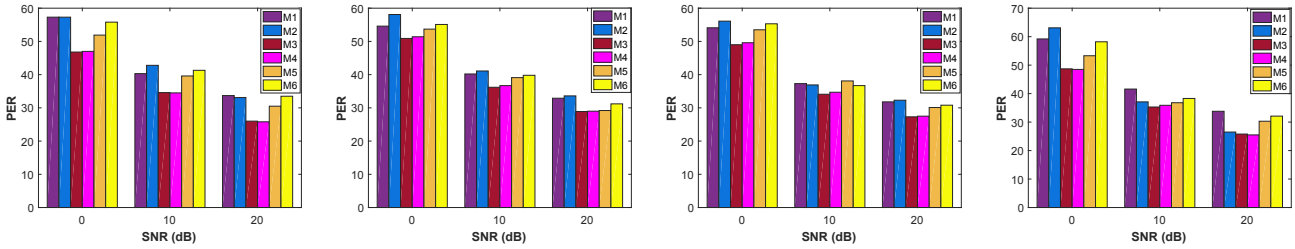


Figure 2: *Spectrogram of noisy speech utterance computed from (a) Magnitude FT (b) MODGD (c) CGD (d) APGD*

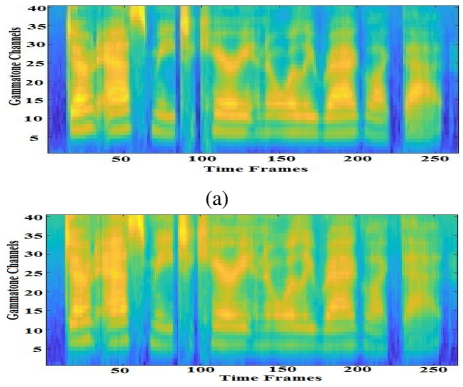
TIMIT [18] which is a phonetically rich resource aimed for speech recognition tasks. The speech utterances are corrupted with different noise sources (babble, white, volvo, factory) taken from NOISEX-92 database [19]. Here, the FaNT tool is used to corrupt the speech samples with different SNRs and down-sampled to 16 KHz. Phone Error Rate (PER) is chosen to measure the performance efficiency of the ASR system with the proposed feature set. Kaldi open source toolkit is used to carry out the experimental verification of the proposal [20]. Here, the aim is to replace the magnitude spectrum with different group delay representations in the spectro-temporal feature extraction process and asses their importance. The optimized set of parameters of gammatonegram based Gabor spectro-temporal features is taken from our previous works which has shown a higher accuracy rate for noisy recognition tasks. The optimal structure for the DNN network is taken from [14], which is found suitable for spectro-temporal feature extraction for ASR task.

We compared the recent group delay based spectral representation pictorially to throw light on the performance measure obtained using these feature sets as shown in Figure 2. The spectrogram representation are plotted in frame by frame fashion. The closest phase spectra resembling the magnitude spectra of Figure 2(a) is the MODGD representation as shown in Figure 2(b). Analysis of group delay representation in noise showed that the MODGD spectrum tends to follow the signal spectrum more better than CGD and APGD representation. The spectral envelope in CGD is seen blurred with less resolution of spectral peaks as evident from Figure 2(c). APGD provides a smoother response of the vocal resonances along with the noise spectrum, thereby masking the minutes cues of spectral information as seen in Figure 2(d). MODGD models the formants



(a) (b) (c) (d)

Figure 4: *Recognition scores obtained using spectral (M1), temporal (M2), magnitude based spectro-temporal (M3) and phase based spectro-temporal (M4,M5,M6) features for different noise sources (a) Babble (b) White (c) Volvo (d) Factory*



(a)

(b)

Figure 3: *Gammatonegram using (a) Magnitude FT (b) MODGD*

when the resonances are more prominent. Thus, the higher formants are less likely to be masked by additive noise. MODGD is the best match among the phase spectrum that represents the resonance of the vocal tract in comparable to the magnitude spectrum.

In order to examine whether MODGD can convey meaningful attribute in spectro-temporal feature extraction, we show the gammatonegram plots for both magnitude and MODGD representations. We find that the spectral observations of both the spectrograms are almost similar as shown in Figure 3. It is seen that the peaks and valleys are better detected and resonant/anti-resonant structures are analyzed more distinctly with higher resolution. This suggest that MODGD has the potential to extract vital features and can be used for spectro-temporal feature extraction. Now, the main investigation is to monitor if they provide enough information and enhance the performance results in ASR framework.

We examine the performance rate of the different spectro-temporal features computed with magnitude and group delay representations. A comparison with state of art spectral, temporal and spectro-temporal features to show the efficiency of the phase based spectro-temporal representation is also highlighted. The notation of the various methodologies chosen here for comparison is shown in Table 1 and the performance evaluation for different noise environments with varying SNR is shown in Figure 4. The features of the group delay function computed from chirp Z transform (M5) and parametric all pole model (M6) do not perform very well as seen from Figure 4. For some SNR values, these features do however provide reasonable

Table 1: *Notations of different features used for comparison*

Sl. No.	Notation	Description
1	M1	MFCC with delta and delta-delta co-efficients
2	M2	RASTA-PLP
3	M3	Spectro-temporal features using Magnitude spectrum
4	M4	Spectro-temporal features using MODGD spectrum
5	M5	Spectro-temporal features using CGD spectrum
6	M6	Spectro-temporal features using APGD spectrum

scores in comparison to spectral (M1) and temporal features (M2). The average performance of magnitude based feature (M3) and MODGD phase based feature (M4) is almost similar across the various noise types considered here. Although the performance obtained with (M3) is better for most of the cases, a relative difference of only 0.03% on average is observed for (M4) as evident from Figure 4. MODGD results outperform the remaining group function mainly due to high intensity amplitude like information which is not seen otherwise in other functions. This shows that phase information can be utilized for spectro-temporal feature extraction also.

6. Conclusion

In this work, a robust group delay based phase spectro-temporal features are used in ASR work. Different phase spectrum are investigated to see whether they provide any useful information in spectro-temporal feature extraction. We found MODGD based spectrum suitable for ASR tasks which showed more harmonic resemblance to the amplitude spectrum. The study established that phase based features are as important as magnitude and should be incorporated in a speech processing task. We believe that the proposed method when combined with other auditory inspired features can become a potential candidate for different speech applications.

7. Acknowledgment

We are thankful to Visvesvaraya PhD Scheme and National Institute of Technology Meghalaya for giving us the grant and the opportunity to carry on with the research.

8. References

[1] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 4134–4151, 2012.

[2] A. M. C. Martinez, S. H. Mallidi, and B. T. Meyer, "On the relevance of auditory-based gabor features for deep learning in robust speech recognition," *Computer Speech & Language*, vol. 45, pp. 21–38, 2017.

[3] A. V. Oppenheim, J. S. Lim, and S. R. Curtis, "Signal synthesis and reconstruction from partial fourier-domain information," *JOSA*, vol. 73, no. 11, pp. 1413–1420, 1983.

[4] H. Von Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Longmans, Green, 1912.

[5] M. R. Schroeder, "Models of hearing," *Proceedings of the IEEE*, vol. 63, no. 9, pp. 1332–1350, 1975.

[6] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, 1981.

[7] 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, no. 3, pp. 578–616, 2007.

[8] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech communication*, vol. 81, pp. 1–29, 2016.

[9] K. M. Murthy and B. Yegnanarayana, "Effectiveness of representation of signals through group delay functions," *Signal Processing*, vol. 17, no. 2, pp. 141–150, 1989.

[10] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 190–202, 2007.

[11] B. Bozkurt, L. Couvreur, and T. Dutoit, "Chirp group delay analysis of speech signals," *Speech communication*, vol. 49, no. 3, pp. 159–176, 2007.

[12] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[13] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 401–411, 2009.

[14] A. Dutta, G. Ashishkumar, and C. V. R. Rao, "Designing of gabor filters for spectro-temporal feature extraction to improve the performance of asr system," *International Journal of Speech Technology*, vol. 22, no. 4, pp. 1085–1097, 2019.

[15] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 215–219.

[16] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, vol. 29, 2012.

[17] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of deep neural networks with natural gradient and parameter averaging," *arXiv preprint arXiv:1410.7455*, 2014.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[19] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.