University of Sheffield Engineering Symposium (USES)
ISSN (Online) XXXX-XXXX

Volume I, Issue I (XXXXX 2014), PP.

Compression of Model-based Group Delay Function for Robust Speech Recognition

Erfan Loweimi, Jon Barker and Thomas Hain

Speech and Hearing Research Group (SPandH), Department of Computer Science, University of Sheffield

Abstract

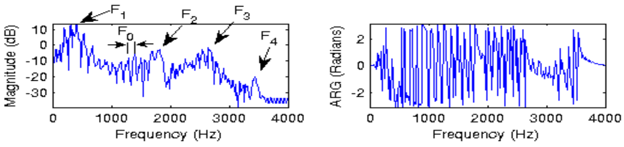
In this paper, we improve the performance of the ARGDMF [3] feature by adding a nonlinear filtering block. ARGDMF is a group delay-based feature consists of four main parts, namely autoregressive (AR) model extraction, group delay function (GDF) calculation, compression, and scale information augmentation. The main problem with the GDF is its spiky nature which is solved by coupling the GDF with an all-pole model. 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. The novelty of this paper is in adding a filtering block to compression process to make it more efficient. This filter aims at elevating the performance of the ARGDMF via a more optimum dynamic range and formants sharpness adjustment. The feature was evaluated on Aurora 2 database. In the presence of both additive and convolutional noises, the proposed method noticeably outperforms the MFCCs and other phase-based features, without remarkable increase in computational load.

Keywords Robust speech recognition, phase spectrum, group delay, compression, scale information

1. Introduction

The Fourier analysis plays a major role in signal processing. It returns a complex-valued function of frequency which can be represented in polar coordinates in terms of magnitude and phase spectra. For speech signals, the magnitude spectrum is believed to carry the most important information while the phase spectrum is thought to play a marginal role [1] because of lacking perceptually important information and noise-like shape which limits its physical interpretation and mathematical modeling.

Most of the modeling techniques try to capture either the trend or extrema which exists in the data. Such mathematical clues are related closely to the physical properties of the underlying process which generates the data. For instance, in case of the speech signal, the magnitude spectrum trend and its extrema closely pertain to the speech production system characteristics. The phase spectrum, however, behaves ambiguously as in Figure 1.



(a) (b)

Figure 1. (a) Magnitude spectrum, (b) phase spectrum.

In an extensive set of experiments [2], we tried to reinvestigate the information content of the phase spectrum from a perceptual point of view. The speech signal was reconstructed from the phase-only and the magnitude-only spectra and the information content of each spectrum was estimated based on the distance of the reconstructed signal from the original signal. The more similarity, the higher the information. For gauging the similarity a perceptually motivate measure was employed. The results showed that contrary to the prevailing belief, there is notable information in the phase.

The next step is to take advantage of such information for practical applications. To do so, a novel phase-based front-end was proposed in [3]. The proposed features afford significantly better performance under noisy condition than the MFCC (baseline). On average, up to 15% higher recognition rates were attained (absolute). In this paper we add a nonlinear filtering block to the proposed method in [13] in order to improve the recognition rates. This filter adjusts the dynamic range and bandwidth of the formants.

The rest of this paper is structured as follows. In Section 2, the block diagram of the proposed feature extraction method is presented and explained. Section 3 includes the results and discussion and Section 4 concludes the paper.

2. Workflow of the Proposed Method

Figure 2 illustrates the block diagram of the suggested algorithm. As seen, the feature extraction starts with pre-emphasis and windowing. For pre-mephasis, a single-zero FIR high-pass filter with a zero at $r(1)/r(0)$ was used ($r(k)$ denotes autocorrelation sequence and k is the lag). For windowing, a Chebyshev window with 30 dB dynamic range was applied. In [2], it is shown that this window maximizes the quality of the phase-only reconstructed speech.

The next step is signal modeling. An autoregressive (AR) model was extracted from each frame through LPC method. The AR model provides a reasonable estimate of the vocal tract characteristics. Next, the group delay of the parametric model (AR) was computed. The main problem with applying the group delay in speech processing is its spiky nature due to the zeros introduced by the excitation component of the speech which are close to the unit circle.

MODGDF [4] and CGDF [5] are two proposed solutions to this issue. None of them provide notable improvement over MFCCs, however. Coupling the GDF with the AR model will alleviate the aforementioned problem. It also paves the way for taking advantage of the high resolution property of GDF. Figure 3 illustrates the power spectrum and different variants of the GDF in clean and noisy (10 dB) conditions.

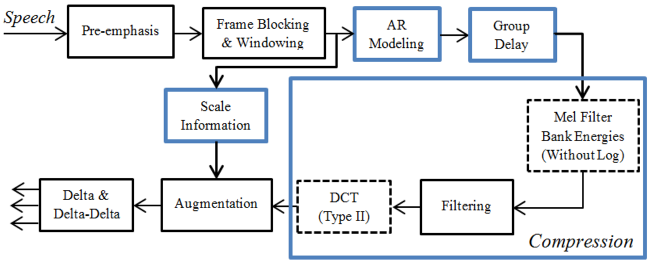
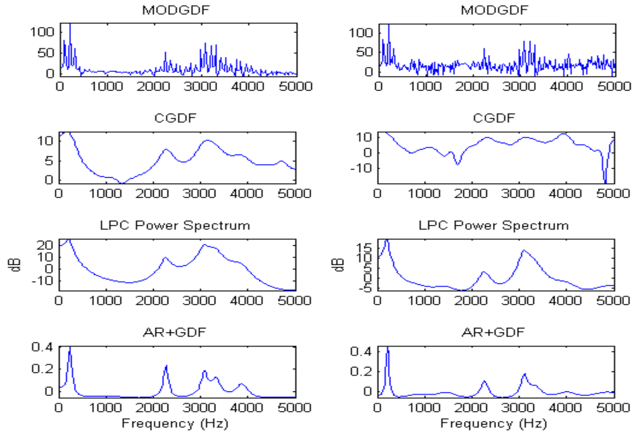


Figure 2. Block diagram of the proposed method.



(a) (b)

Figure 3. Various group delay-based representations. (a) clean signal, (b) noisy signal (additive white Gaussian noise, SNR: 10 dB).

The next step is compression which aims at representing each frame with fewer samples and also limiting the dynamic range while preserving all significant information. To do so, the mel filterbank is employed. As well, we take the DCT of the energies of the filterbank outputs and retain the first 12 coefficients. In contrast to MFCC, logarithm was not taken from the output energies. There are two main reasons for taking the logarithm: compressing the dynamic range; and converting the multiplication between the source and filter components into the addition. This paves the way for homomorphic processing of the speech and helps in separating the excitation and vocal tract components. The convolution in time domain will be equivalent to the addition in group delay domain not multiplication. In addition, the dynamic range of the group delay coupled with LPC (all-pole) model is limited (Figure (3)). For better adjustment of the dynamic range we added a nonlinear filter as follows

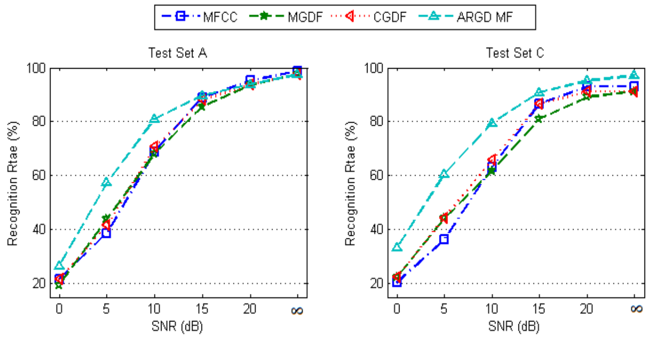
$$ene_2 = ene_1^{\alpha} \tag{1}$$

where ene_1 and ene_2 denote the input (energy) and output of the filtering block, respectively and α is compression factor. By decreasing this coefficient the bandwidth of the spectrum peaks (formants) increases and the dynamic range reduces. Our simulation results indicate that the optimum value for α over Aurora 2 [6] database is 0.85. Finally, the feature vector is augmented with the scale information [3] which is computed by Hilbert transform relations. The proposed method is called ARGDMF.

3. Experimental Results and Discussion

The performance of ARGDMF was assessed on the Aurora 2 database [6]. Aurora2 includes three test sets which A and B include additive noises while the C test set contains both additive and convolutional distortions. We have used clean-data training in all our experiments and HMMs were trained with HTK [7]. Table 1 and Figure 4 show the results.

As seen, the proposed method both on average and at each SNR returns significantly better performance than the other techniques. The reason backs to the capability of this feature in coping with the noise. Figure 3 shows the spectrum distortion due to the noise in ARGD is less than the other methods which indicates its higher robustness. As seen in Table 1 the proposed modification improves the results without introducing noticeable computational overhead.



(a) (b)

Figure 4. Recognition rates as a function of SNR. 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 C. To avoid clatter we have just plotted one of the ARGDMF variants.

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

	TEST SET A	TEST SET B	TEST SET C
MFCC-D-A	62.7	66.9	60.0
MODGDF-D-A	64.0	67.7	62.6
CGDF-D-A	63.1	67.8	61.9
ARGDMF1*-D-A	74.7	77.9	75.6
ARGDMF2*-D-A	**75.5**	**78.6**	**76.4**

* ARGDMF1: without filtering, ARGDMF2: with Filtering.

4. Conclusion

In this paper we have added a nonlinear filter to the parameterization algorithm proposed in [3]. This block aims at improving the performance of the ARGDMF [3] by further adjustment of the sharpness of the resonance frequencies (formants) and compression of the dynamic range. Looking for more optimum filters could further elevate the performance of the proposed feature and increase its robustness against additive and convolutional noises.

References

1. Schroeder MR. Computer Speech: Recognition, Compression, Synthesis. Springer Series in Information Sciences. Springer; 2004.
2. Loweimi E, Ahadi SM, Sheikhzadeh H. Phase-Only Speech Reconstruction Using Very Short Frames. In: INTERSPEECH. ISCA; 2011. p. 2501–2504.
3. Loweimi E, Ahadi SM, Drugman T. A new phase-based feature representation for robust speech recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International conference on; 2013. p. 7155–7159.
4. Hegde RM, Murthy HA, Gadde VRR. Significance of the Modified Group Delay Feature in Speech Recognition. Audio, Speech, and Language Processing, IEEE Transactions on. 2007;15(1):190–202.
5. Bozkurt B, Couvreur L, Dutoit T. Chirp group delay analysis of speech signals. Speech Communication. 2007;49(3):159 – 176.
6. Pearce D, Hirsch HG. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: INTERSPEECH; 2000. p. 29-32.
7. Young SJ, Evermann G, Gales MJF, Hain T, Kershaw D, Moore G, et al. The HTK Book, version 3.4. Cambridge, UK: Cambridge University Engineering Department; 2006.