On a Novel Speech Representation Using Multitapered Modified Group Delay Function

Narendra K. C., Student Member, IEEE, R. Kumaraswamy, Senior Member, IEEE, and Sanjeev Gurugopinath, Member, IEEE

Abstract—In this paper, a novel multitaper modified group delay function-based representation for speech signals is proposed. With a set of phoneme-based experiments, it is shown that the proposed method performs better that an existing multitaper magnitude (MT-MAG) estimation technique, in terms of variance and MSE, both in spectral- and cepstral-domains. In particular, the performance of MT-MOGDF is found to be the best with the Thomson tapers. Additionally, the utility of the MT-MOGDF technique is highlighted in a speaker recognition experimental setup, where an improvement of around 20% compared to the next-best technique is obtained. Moreover, the computational requirements of the proposed technique is comparable to that of MT-MAG. The proposed feature can be used in for many speech-related applications; in particular, it is best suited among those that require information of speaker and speech.

Index Terms—Multitaper methods, modified group delay functions, spectrum estimation, low-variance techniques, speaker recognition.

I. INTRODUCTION

Developing a representation for speech signals is challenging as it contains manifold information about speaker, speech, language, accent, emotion, etc. While the speaker information is normally represented by the fine harmonic structure in the spectrum, the speech information is embedded in the envelope of the spectrum. Spectral-domain representations are widely employed in many applications, and hence obtaining a good estimate of speech spectrum is a key requirement. A simple procedure is to compute the DFT of the autocorrelation sequence of a given speech signal, which is known as the periodogram estimate. However, the periodogram estimate has a large variance [1], and a better estimate can be obtained by multiplying the frame of the speech signal with a window function that is tapered toward the ends of the signal frame. The variance of the estimated spectrum can be further reduced by averaging multiple realizations of the same random process [2], which in the case of speech corresponds to same utterances spoken by different speakers. Use of multiple orthonormal windows [3] [4], instead of a single window as proposed in the former method, yields lower variance by combining the advantages of windowing and averaging. However, this also reduces the spectral resolution. We term this technique as multitaper magnitude (MT-MAG) based estimation, whose analysis and performance benefits are extensively discussed in [5].

Alternatively, the phase spectrum of a speech signal contains information about the resonances of the vocal tract in its transitions. Towards this end, the group delay function (GDF) [6] was defined, which obtains the transitions in the phase spectrum by computing its negative derivative. It is known that the peaks and valleys in the GDF of speech corresponds to the resonances and anti-resonances of the vocal tract, respectively, similar to that of the magnitude spectrum. Additionally, the GDF accurately captures the Formant information in the speech signal, since the phase spectrum is immune to additive degradations on the speech signal [7]. However, the GDF suffers from spurious peaks that are caused due to the zeros located near the unit circle of the z-plane. To alleviate this, a two-parameter modified group delay function (MOGDF) was proposed in, which used a cepstrally smoothed version of the spectrum is employed to reduce the spuriousness in the GDF.

In this paper, a novel speech representation is proposed by processing the phase spectrum – in terms of the MOGDF – of speech signals multiplied with a set of orthonormal tapers. This new representation, termed as the multitaper MOGDF (MT-MOGDF), has the collective advantage of MOGDF and performance benefits obtained by employing multiple tapers. Through experimental results, we show that the proposed technique has lower variance and lower mean squared error (MSE) as compared to the MT-MAG technique, in both spectral- and cepstral-domains. Also, the high resolution property of MOGDF is retained in this representation, which is useful in several applications. As a case study, an application of speaker recognition with varying number of speakers is considered, and is shown that MT-MOGDF performs the best as compared to all the other techniques listed above, owing to its low variance advantage.

The remainder of the paper is organized as follows. The existing spectral estimation techniques and the proposed multitaper modified group delay representation are given in Sec. II. A detailed study on the performance of the proposed technique in comparison to MT-MAG with bias, variance and mean squared error measures for different types and number of tapers is discussed in Sec. III. Performance improvements due to MT-MOGDF as compared to the other techniques for a speaker recognition application are highlighted in Sec. IV.
II. MULTITAPER MODIFIED GROUP DELAY FUNCTION REPRESENTATION

The GDF provides a meaningful representation of the vocal tract information by processing phase spectrum of the speech signal \([7]\). Consider the spectrum of a speech signal \(s(t)\) given by

\[
S(\omega) = |S(\omega)|e^{j\theta(\omega)}, \quad \omega \in \mathbb{R},
\]

where \(|S(\omega)|\) and \(\theta(\omega)\) are the magnitude and phase spectra of the speech signal, respectively. The GDF is defined as

\[
\tau(\omega) = -\frac{d\theta(\omega)}{d\omega}.
\]

The GDF exhibits near-squared magnitude behaviour at the resonances, but exhibits spurious spikes if the zeros are located near unit circle in the \(z\)-plane. To minimize the effect of these peaks, a modified group delay function (MOGDF) was proposed, which alleviates the spuriousness by scaling down the power spectrum \([8]\), and by radially moving the zeros towards the origin. The MOGDF is defined as

\[
\tau_m(\omega) = \frac{\tau(\omega)}{|\tau(\omega)|^{\alpha}}, \quad 0 < \alpha \leq 1,
\]

where

\[
\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{||\xi(\omega)||^{2\gamma}}, \quad 0 < \gamma \leq 1.
\]

Here, \(X(\omega)\) is the Fourier spectrum of a speech signal frame \(x(n)\), and \(Y(\omega)\) is the Fourier spectrum of \(nx(n)\). The subscripts \(R\) and \(I\) denote the real and imaginary parts of the spectrum. \(\xi(\omega)\) is the cepstrally smoothed spectrum of \(S(\omega)\). Parameters \(\alpha\) and \(\gamma\) are tuned to reduce the spurious spikes in the group delay spectrum.

A simple estimate of the spectrum is obtained by computing the DFT of the autocorrelation sequence of a given speech signal, known as the periodogram estimate. However, the periodogram estimate has a large variance \([1]\), which can be reduced either by multiplying the time-domain speech signal by a tapered window, or by averaging over multiple realizations. The multitaper (MT) technique \([5]\) combines both these advantages. An MT-based periodogram spectral estimate was proposed in \([5]\), which was shown to exhibit smoothness with lower variance. Termined as MT-based magnitude spectrum (MT-MAG), this estimate is given by

\[
\hat{S}(\omega) = \frac{1}{N} \sum_{j=1}^{N} \lambda(j) \sum_{t=0}^{M-1} w_j(t)x(t)e^{-j2\pi ft/N}.
\]

The performance is analyzed in terms of a bias-variance study. The MT-MOGDF estimate of \(\hat{S}(\omega)\) is defined as

\[
\hat{S}_{mg}(\omega) = \sum_{j=1}^{N} \lambda(j)\tau_{jm}(\omega)
\]

where \(\tau_{jm}(\omega)\) is the MOGDF function for the speech frame as given in \([3]\) is multiplied by the \(j^{th}\) taper and \(\lambda(j)\) is the associated weight of the taper. Further, the choice of the tapers based on the number and shape is elaborated in \([5]\). To analyze the performance of the MT-MOGDF, the bias, variance and mean squared error (MSE) of the log spectrum are defined as

\[
B(\hat{S}_{mg}(\omega)) = E[log(\hat{S}_{mg}(\omega))] - log(S_T(\omega)),
\]

\[
V(\hat{S}_{mg}(\omega)) = E[log(\hat{S}_{mg}(\omega)^2)] - E[log(\hat{S}_{mg}(\omega))]^2,
\]

\[
MSE(\hat{S}_{mg}(\omega)) = B(\hat{S}_{mg}(\omega)) + V(\hat{S}_{mg}(\omega)).
\]

Speech signals are modeled as concatenation of basic units in sound perception called phonemes, which are further divided into vowels and non-vowels. The vowels form a significant part of the speech signal as the energy concentration is more in these regions. Spectrum in vowel-like regions of the speech signal has peaks centered around the resonances of the vocal tract. Also, there would be no constrictions along the vocal tract when the vowels are produced. Hence, the Formant locations are significant for spectrum estimation. The proposed MT-MOGDF exploits this key requirement. Among the non-vowels, an extreme type are the sounds of fricatives. These sounds have a flattened spectrum and varies largely across speakers. Therefore, energy-based representations may not be sufficient to model these sounds. It is shown that MT-MOGDF provides a better representation even in this case, with an appropriate choice of taper.

Most applications in speech processing use the features in cepstral-domain as a better representation. Moreover, the variance and bias improvements in spectral-domain are intuitively reflected in the cepstral-domain. Towards this end, a detailed performance study is considered in the spectral-domain, as elaborated in the next section. The cepstrum is computed as

\[
\hat{C} = \text{DCT}\{\log(\hat{S}_{mg}(\omega))\}.
\]

DCT is used for compactness in the representation \([9]\).

III. EXPERIMENTAL RESULTS AND DISCUSSION

The phoneme instances are obtained from DR7 directory of TIMIT database \([10]\). 250 instances of the vowel /a/, and fricative /s/ are collected from different speakers uniformly randomly. Frames of 10 ms with an overlap of 50% is considered. This set of frames serves as a set of multiple realizations of the same underlying random process. These are used to obtain the sample mean of the speech signal, whose spectrum is assumed to be the true mean. A total of \(N_s = 11204\) for /a/, and \(N_s = 15064\) for /s/ instances are extracted.

The MT-MOGDF spectrum is computed as given in \([7]\).

Comparison of MT-MAG and MT-MOGDF estimates for the above setup with different types of tapers, namely, Thomson,
Bias variance over all frequencies. The MT-MOGDF technique captured. Usage of a multipeak taper almost uniformly reduces located in frequencies that are greater than 1 kHz should be observed in MT-MOGDF phase is employed for estimation. A trend of lower variance technique holds good irrespective of whether magnitude or

In other words, the trend dictated by using the multitaper [11] function [6]. Note that the comparison with the periodogram estimate is omitted since MT-MAG is known to outperform the periodogram-based estimate in terms of bias and variance [5]. Note that the spectrum obtained by MT-MAG is smoother, and as \( N \) increases, MT-MOGDF also gives a smoother spectrum. Additionally, information on formant locations is lost in MT-MAG as \( N \) increases, while it is retained in MT-MOGDF.

Next, the performances are studied of MT-MOGDF and MT-MAG in terms of bias and variance measures. The performance comparison of these two methods is considered with Thomson and multipeak tapers in Figs. 2 and 3 respectively. In both cases, it is observed that MT-MOGDF gives a lesser variance, albeit with a larger bias. An important key feature of the multitaper technique, i.e., reduction in variance with \( N \) as governed by (6), is retained even in the case of MT-MOGDF. In other words, the trend dictated by using the multipeak taper [11] technique holds good irrespective of whether magnitude or phase is employed for estimation. A trend of lower variance for a larger band frequency can be observed in MT-MOGDF than the MT-MAG technique. This is important as the Formants located in frequencies that are greater than 1 kHz should be captured. Usage of a multipeak taper almost uniformly reduces variance over all frequencies. The MT-MOGDF technique works especially well with the Thomson taper, because of its ability to retain the Formant structure. This bias-variance tradeoff between the two techniques becomes clear with the notion of MSE, which will be explored next. Note that the MSE depends directly on the variance and square of the bias term, and hence the sign of the bias is also important.

Fig. 2 shows the bias, variance and MSE variation of MT-MOGDF and MT-MAG with Thomson and multipeak tapers, with \( N = 8 \). The MOGDF has a lesser MSE as compared to MT-MAG, even though its bias is larger. This establishes that the variance of spectral estimation takes a priority as opposed to the bias. Also, the variance and MSE values of MT-MOGDF with both tapers are better, and Thomson tapers
give the least variance, across both phonemes. The multipeak tapers provide the best MSE for the vowel phoneme /a/ as expected, since this was proposed to estimate peaked spectra. Similarly, Thomson tapers perform the best in terms of MSE with fricative phoneme /s/.

The performance of both techniques in the cepstral-domain is considered, as given in Fig. 5. Cepstrum representation using MT-MAG and MT-MOGDF is obtained by computing the DCT of the log spectrum as given in (11). As expected, the bias and variance of both estimates are reduced as compared to the spectral-domain. The MT-MOGDF representation gives a lesser variance for nearly the same value of bias, compared to MT-MAG. The improvement of MT-MOGDF over MT-MAG is less significant in cepstral-domain, for the considered phonemes. Even here, MT-MOGDF with Thomson taper is found to have lesser variance. The performance trend in this case is similar to that seen in the spectral-domain.

IV. APPLICATION: SPEAKER RECOGNITION

As an application-study of the proposed technique, a speaker recognition system is considered that uses the proposed MT-MOGDF features. The features chosen for comparison are the cepstral-domain features extracted using the periodogram, MOGDF, MT-MAG and the MT-MOGDF spectrum estimates. The cepstral coefficients are then obtained by multiplying the estimated spectrum with the response of a Mel filter bank, followed by a logarithm operator. An associated GMM-UBM system is trained for a varying number of speakers and the corresponding EER and DCF (according to SRE10) values are tabulated in Tab. I. Among the considered methods, the MT-MOGDF features are most effective, and provide an improvement of around 20% in comparison to the next-best MT-MAG feature set, averaged over the number of speakers. This improvement is attributed due to the reduction in the variance and high resolution properties of MT-MOGDF.

V. CONCLUSION

In this paper, a novel representation for speech signals is proposed using multitaper-modified group delay functions. The proposed feature can be used in many speech applications, and in particular, among those that require information of speaker and speech. With different set of phonemes, it can be shown that the proposed method performs better than the existing multitaper-magnitude (MT-MAG) estimation technique, in terms of variance and MSE, both in spectral- and cepstral-domains. The performance of the proposed technique is in particular found to be the best with Thomson tapers. The computational requirements of the proposed technique is comparable to that of MT-MAG. Additionally, the utility of the proposed method is presented in a speaker recognition setup with an improvement of around 20%.
REFERENCES

[1] B. Yegnanarayana and H. A. Murthy, “Significance of group delay functions in spectrum estimation,” IEEE Transactions on Signal Processing, vol. 40, no. 9, pp. 2281–2289, 1992.

[2] P. Welch, “The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms,” IEEE Transactions on Audio and Electroacoustics, vol. 15, no. 2, pp. 70–73, 1967.

[3] D. J. Thomson, “Spectrum estimation and harmonic analysis,” Proceedings of the IEEE, vol. 70, no. 9, pp. 1055–1096, 1982.

[4] B. Babadi and E. N. Brown, “A review of multitaper spectral analysis,” IEEE Transactions on Biomedical Engineering, vol. 61, no. 5, pp. 1555–1564, 2014.

[5] T. Kinnunen, R. Saeidi, F. Sedlák, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li, “Low-variance multitaper MFCC features: A case study in robust speaker verification,” IEEE transactions on Audio, Speech, and Language Processing, vol. 20, no. 7, pp. 1990–2001, 2012.

[6] B. Yegnanarayana, D. Saikia, and T. Krishnan, “Significance of group delay functions in signal reconstruction from spectral magnitude or phase,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 3, pp. 610–623, 1984.

[7] H. A. Murthy and B. Yegnanarayana, “Group delay functions and its applications in speech technology,” Sadhana, vol. 36, no. 5, pp. 745–782, 2011.

[8] 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.

[9] S. H. K. Parthasarathi, R. Padmanabhan, and H. A. Murthy, “Robustness of group delay representations for noisy speech signals,” International Journal of Speech Technology, vol. 14, no. 4, pp. 361–368, 2011.

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1.1,” NASA STI/Recon Technical Report n, vol. 93, 1993.

[11] M. Hansson and G. Salomonsson, “A multiple window method for estimation of peaked spectra,” IEEE Transactions on Signal Processing, vol. 45, no. 3, pp. 778–781, 1997.