SYNTHETIC SPEECH DETECTION USING TEMPORAL MODULATION FEATURE

Zhizheng Wu[1,2], Xiong Xiao[2], Eng Siong Chng[1,2], Haizhou Li[1,2,3]

[1]School of Computer Engineering, Nanyang Technological University (NTU), Singapore
[2]Temasek Laboratories@NTU, Nanyang Technological University (NTU), Singapore
[3]Human Language Technology Department, Institute for Infocomm Research (I^{2}R), Singapore
wuzz@ntu.edu.sg

ABSTRACT

Voice conversion and speaker adaptation techniques present a threat to current state-of-the-art speaker verification systems. To prevent such spoofing attack and enhance the security of speaker verification systems, the development of anti-spoofing techniques to distinguish synthetic and human speech is necessary. In this study, we continue the quest to discriminate synthetic and human speech. Motivated by the facts that current analysis-synthesis techniques operate on frame level and make the frame-by-frame independence assumption, we proposed to adopt magnitude/phase modulation features to detect synthetic speech from human speech. Modulation features derived from magnitude/phase spectrum carry long-term temporal information of speech, and may be able to detect temporal artifacts caused by the frame-by-frame processing in the synthesis of speech signal. From our synthetic speech detection results, the modulation features provide complementary information to magnitude/phase features. The best detection performance is obtained by fusing phase modulation features and phase features, yielding an equal error rate of 0.89%, which is significantly lower than the 1.25% of phase features and 10.98% of MFCC features.

Index Terms— Anti-spoofing attack, synthetic detection, modulation, phase modulation, temporal feature

1. INTRODUCTION

The task of speaker verification is to make a binary decision to accept or reject a claimed identity based on a speech sample. It has many applications in telephone or network access control systems, such as telephone banking and telephone credit cards [1]. The security of speaker verification systems is threatened by two related speech processing techniques, i.e. voice conversion [2] and speaker adapted speech synthesis [3]. In voice conversion, the speech of a source speaker is converted to sound like a target speaker, while speaker adapted speech synthesis can mimic the voice of the target speaker given any text. As both techniques are able to produce a voice to sound like that uttered by the claimed speaker, they present a threat to speaker verification systems.

The research on vulnerability of speaker verification systems against voice conversion and HMM-based speech synthesis systems has received a lot of attentions. In [4, 5, 6], the authors study the vulnerability of speaker verification systems under spoofing attack using synthetic speech from HMM-based speech synthesis system. Adapted HMM-based speech synthesis system is studied in [7], and voice conversion techniques in [8, 9, 10]. These studies on both high quality and telephone quality speech confirm the vulnerability of speaker verification systems. For this reason, the detection of synthetic speech from human speech is an important task to enhance the security of speaker verification systems.

To defend the spoofing attack using synthetic speech, a detection technique to distinguish between synthetic and human speech is necessary. To respond to such a concern, features based on relative phase shift to classify HMM-based synthetic speech from human speech is proposed in [7]. In our previous study [11], motivated by the facts that synthesis filter introduces artifacts in phase spectrum, features derived from cosine-normalized phase and modified group delay function phase spectrum are proposed to discriminating voice converted speech from human speech. Both studies [7, 11] show that phase related features outperform magnitude-based features (e.g. MFCC), confirming that the original phase information is lost in the synthesized/converted speech.

However, all the above features are extracted at frame level, ignoring the distortions in the temporal structure of synthetic speech. Current speech synthesis methods usually synthesize speech signal frame-by-frame. Therefore, besides the short-term spectral distortion produced by reconstruction, the frame-based operation also introduces long-term temporal artifacts in the reconstructed speech signal. Motivated by this fact, we propose to use modulation features to capture speech variation cross frames for detecting synthetic speech. Modulation features capture the long-term temporal information of speech and have been shown to be effective in speech recognition [12], speaker verification [13], and nonnative accent detection [14]. In this paper, we apply the modulation features to the synthetic speech detection task and investigate their interactions with phase-based features [11]. Unlike previous modulation features [12, 13, 14] which are derived from magnitude spectrum only, we also study modulation features derived from phase information of speech.

The paper is organized as follows: corpus design is introduced in section 2. In section 3 and 4, short-term spectral feature extraction and long-term modulation feature extraction are discussed. The score fusion which to combine feature and modulation feature information is presented in section 5. The experimental setups are presented in section 6. We will conclude the paper in section 7.

2. CORPUS DESIGN

Due to the vulnerability of speaker verification systems against both voice conversion and speaker adapted speech synthesis techniques, we decided to develop an anti-spoofing technique to detect both synthesized speech from adapted HMM-based speech synthesis system and voice conversion system. The difference between HMM-based speech synthesis and voice conversion is that: the input for voice conversion is human speech, hence the converted speech can copy

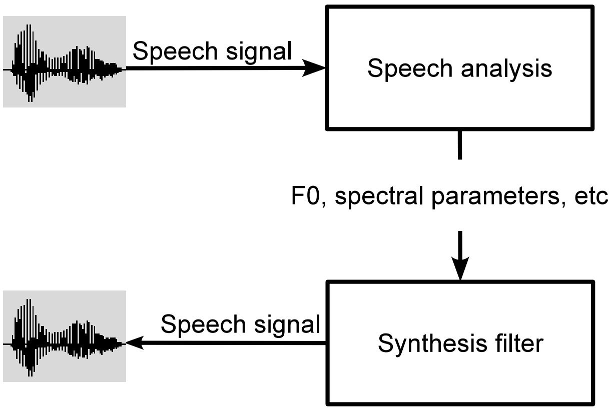

Fig. 1. The analysis-synthesis process to obtain synthetic speech

some human speech information from source speech, such as fundamental frequency and voiced and un-voiced information, while HMM-based speech synthesis can not do this due to that the input is text not speech signal. The common module of the two techniques is that they both adopt the vocoder technique to reconstruct speech signal from speech parameters. To this end, we design the synthetic corpus only employing speech analysis-synthesis technique without any further modification on the features.

Table 1. Statistics of the designed dataset

	Number of utterances	
	Human speech	Synthetic speech
Training set	$4,007$	$4,007$
Development set	$3,131$	$3,131$
Testing set	$15,000$	$15,000$

The Wall Street Journal corpora (WSJ0+WSJ1) are used to generate our synthetic speech detection corpus in this work. The original waveforms are used as human speech. The synthetic speech is obtained by employing the analysis-synthesis process as illustrated in Fig. 1. The speech signal is first analyzed to extract representation parameters, such as fundamental frequency (F0) and spectral parameters. These parameters are then passed directly to a synthesis filter to reconstruct speech signal. The reconstructed speech signal is used as synthetic speech. To produce good quality synthetic speech, we employ STRAIGHT [15], a state-of-the-art speech analysis-synthesis system, to first decompose speech into speech envelope, excitation with fundamental frequency (F0) and aperiodicity envelope, and then reconstruct speech signal from these parameters.

We divided the corpora into three parts: training, development and testing sets. The training and development data consist of $4,007$ utterances, and $3,131$ utterances from WSJ0, respectively. Another 15, 000 utterances from WSJ1 are used as testing data. For each original waveform, a synthetic version is generated. As a result, there are $6,262$ utterances and $30,000$ utterances in the development and testing sets, respectively, including human speech and synthetic speech. The statistics of the dataset are presented in Table 1. We note that there is no speaker overlap between the three datasets.

3. MAGNITUDE AND PHASE SPECTRUM

Given a speech signal, it can be processed by short-time Fourier analysis by assuming the signal to be quasi-stationary within a short period (e.g. 25ms). The short-time Fourier transform of the speech signal $x(n)$ is presented as follows:

$$X(w) = |X(w)|e^{j\phi(w)}, \tag{1}$$

where $|X(w)|$ is the magnitude spectrum and $\phi(w)$ is the phase spectrum. We note that $X(w)$ has two parts: real part $X_R(w)$ and imaginary part $X_I(w)$. The power spectrum is defined to be $|X(w)|^2$. Usually, features (e.g. MFCC) which contain magnitude information can be derived from the power spectrum.

In order to include phase information, the modified group delay function phase spectrum (MGDFPS), which is popular in speech recognition [16, 17], is derived from the Fourier transform spectrum $X(w)$. The modified group delay function phase spectrum is defined as follows:

$$\tau_\rho(w) = \frac{X_R(w)Y_R(w) + Y_I(w)X_I(w)}{|S(w)|^{2\rho}}, \tag{2}$$

$$\tau_{\rho,\gamma}(w) = \frac{\tau_\rho(w)}{|\tau_\rho(w)|}|\tau_\rho(w)|^\gamma, \tag{3}$$

where $X_R(w)$ and $X_I(w)$ are the real and imaginary parts of $X(w)$, respectively; $Y_R(w)$ and $Y_I(w)$ are the real and imaginary parts of the Fourier transform spectrum of $nx(n)$, respectively; $|S(w)|^2$ is the cepstrally smoothed power spectrum corresponding to $X(w)$, ρ and γ are two weighted variables and $\tau_{\rho,\gamma}(w)$ is the MGDFPS. In practice, $|S(w)|^2$ is obtained by applying discrete cosine transform (DCT) on the power spectrum and then pass the first 30 DCT coefficients to inverse discrete cosine transform (IDCT) to reconstruct the trajectory.

3.1. Mel-frequency cepstral coefficients (MFCC)

The Mel-frequency cepstral coefficient (MFCC) is derived from the magnitude spectrum $|X(w)|$ in the following steps:

a) Employ the fast Fourier transform (FFT) to compute the spectrum $X(w)$ of $x(n)$.

b) Compute power spectrum $|X(w)|^2$.

c) Compute filter-bank energies (FBE) by apply a Mel-frequency filter bank to the power spectrum $|X(w)|^2$.

d) Apply discrete cosine transform (DCT) to log-scale FBE to compute the MFCC.

3.2. Modified group delay cepstral coefficients (MGDCC)

We can compute modified group delay cepstral coefficients (MGDCC) from the modified group delay function phase spectrum as follows:

a) Adopt FFT to compute the spectrum $X(w)$ and $Y(w)$ of $x(n)$ and $nx(n)$, respectively.

b) Compute the cepstrally smoothed power spectrum $|S(w)|^2$ of $|X(w)|^2$.

c) Compute MGDFPS using Equation (2) and (3).

d) Obtain filter-bank energies (FBE) by apple a Mel-frequency filter bank to MGDFPS.

e) Apply DCT to FBE to calculate the MGDCC.

In [11], MGDCC has been adopted to detect synthetic speech from human speech, which has achieved better performance than MFCC. In this work, we will use MFCC and MGDCC as baseline features. We note both MFCC and MGDCC are computed at frame level without knowledge of the consecutive frames.

4. MAGNITUDE AND PHASE MODULATION FEATURE EXTRACTION

Both MFCC and MGDCC are derived from power spectrogram and modified group delay function phase spectrogram, respectively, in a frame-by-frame fashion. As a result, they are not good at capturing the correlation between frames, or the temporal characteristics of speech feature trajectories. On the other hand, the frame-based operation in speech analysis and synthesis process as illustrated in Fig. 1 may introduce temporal artifacts. In order to consider the frame dependency and capture the temporal artifacts in the synthetic speech, we proposed to use modulation features to discriminate between synthetic speech and human speech. In this section, we will describe the modulation spectrum extraction procedure and the feature extraction from the modulation spectrum.

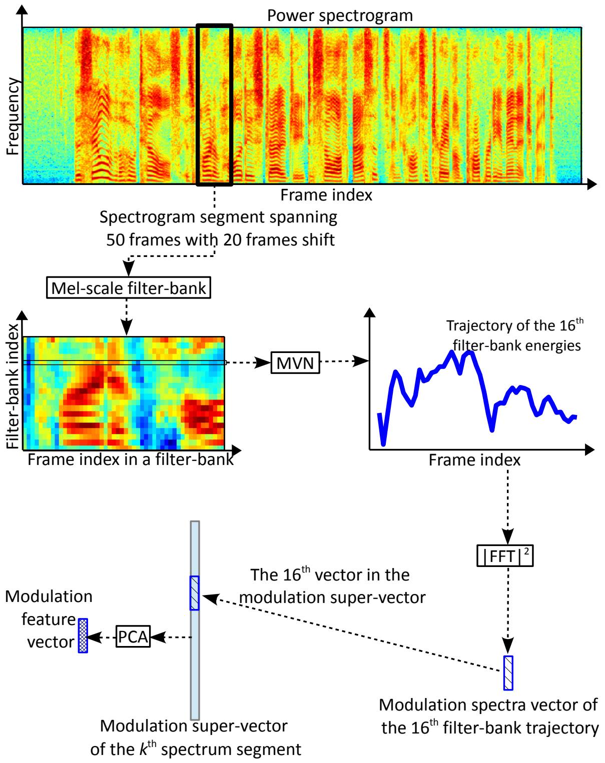
Power spectrogram
Frequency
Frame index
Spectrogram segment spanning 50 frames with 20 frames shift
Mel-scale filter-bank
Filter-bank index
Frame index in a filter-bank
MVN
Trajectory of the 16th filter-bank energies
Frame index
|FFT|²
The 16th vector in the modulation super-vector
Modulation feature vector
PCA
Modulation super-vector of the kth spectrum segment
Modulation spectra vector of the 16th filter-bank trajectory

Fig. 2. Illustration of modulation feature extraction from power spectrogram.

The modulation feature extraction process is illustrated in Fig. 2 and it can be applied on both the power spectrum and modified group delay function phase spectrum. In Fig. 2, a power spectrogram is used for illustration. We first divide the spectrogram into overlapping segments using a 50 frames window with 20 frames shift. Then 20 filter-bank coefficients are obtained from the spectrogram to form a 20 × 50 matrix. After that, mean variance normalization (MVN) is applied to the trajectory of each filter-bank to normalize the mean and variance to zero and one, respectively. To compute modulation spectrum from filter-bank energies, fast Fourier transform (FFT) is applied to the 20 normalized trajectories. Every modulation spectra in the spectrum is concatenated to make up a modulation super-vector as the feature vector.

In this work, the dimensionality of each modulation spectra is 32 as we used a 64-point FFT. Thus, the dimensionality of the modulation super-vector for each spectrum segment is $20 \times 32 = 640$. Due to the high dimensionality and high correlation between modulation spectra of different filter bank trajectories, dimensionality reduction is necessary. Thus, we apply principal component analysis (PCA) on the modulation super-vectors, and 10 projected dimensions with largest variances are used as the features for the synthetic detection task. We call modulation features derived from power spectrogram as magnitude modulation (MM), and features derived from modified group delay function phase spectrogram as phase modulation (PM). As illustrated in Fig. 2, the phase modulation can be extracted by only changing the spectrogram before applying the Mel-scale filter bank.

5. MODEL TRAINING AND SCORE FUSION

In this study, the Gaussian mixture model (GMM) is used to model the feature distributions of the synthetic speech and human speech. The synthetic or human decision is made based on log-likelihood ratio:

$$\mathcal{L}(O) = \log p(O|\lambda_{\text{synthetic}}) - \log p(O|\lambda_{\text{human}}), \tag{4}$$

where O is the feature vector sequence of a test speech signal, $\lambda_{\text{synthetic}}$ and λ_{human} are GMM models for synthetic and human speech, respectively. In our implementation, for MFCC and MGDCC features, 512 Gaussian components are adopted to model the distribution; and 16 Gaussian components for MM and PM features. As modulation features are extracted from spectrogram segments whose window size is 50 frames with 20 frames shift, the number of training modulation feature vectors is one-twentieth of MFCC or MGDCC.

To benefit from both short-term spectral and long-term temporal features, we utilize score fusion method.

$$\mathcal{L}_{\text{combine}}(O) = (1-\alpha)\mathcal{L}_{\text{A}}(O) + \alpha\mathcal{L}_{\text{B}}(O), \tag{5}$$

where $\mathcal{L}_{\text{A}}(O)$ and $\mathcal{L}_{\text{B}}(O)$ are two log-likelihood score adopting two different features, and α is the weighting coefficient to balance the two scores. We will investigate the effect of the combination of the above features.

6. EXPERIMENTS

In this study, MFCC and magnitude modulation features are computed from magnitude spectrogram, and MGDCC and phase modulation features are computed from modified group delay function phase spectrogram. The configurations for computing the magnitude or modified group delay function phase spectrogram are presented in Table 2. The weighting parameters ρ and γ for computing MGDFPS in Equation (2) and (3) are decided using the development data. During testing, we fix $\rho = 0.9$ and $\gamma = 1.8$.

Both MFCC and MGDCC are using 36 dimension features, including 12 dimension static features (no energy feature), and their delta and delta-delta features. The dimensionality of the modulation features is set to 10. Delta or delta-delta coefficients for modulation features are not computed.

Table 4. The detection EER with score fusion of different features

	EER (%)					
α	MFCC+MGDCC	MM+PM	MFCC+MM	MFCC+PM	MGDCC+MM	MGDCC+PM
0.0	10.98	13.71	10.98	10.98	1.25	1.25
0.1	1.64	13.38	9.73	9.83	1.20	1.21
0.2	1.12	**13.33**	8.94	8.73	1.16	1.17
0.3	1.03	13.51	**8.51**	7.85	1.10	1.13
0.4	**1.02**	13.93	8.72	7.30	1.02	1.07
0.5	1.07	14.54	9.43	**7.17**	**0.98**	1.00
0.6	1.14	15.43	10.73	7.42	0.99	0.92
0.7	1.19	16.39	12.60	8.18	1.08	**0.89**
0.8	1.21	17.43	14.79	9.47	1.41	1.00
0.9	1.23	18.29	17.09	11.33	2.97	1.63
1.0	1.25	19.29	19.29	13.71	19.29	13.71

Table 2. Parameter configurations for computing the spectrogram

Parameter configurations	
Pre-emphasis filter	$H(z) = 1 - 0.97z^{-1}$
Window function	Hamming
Window size	400 samples (25ms)
Window shift	160 samples (10ms)
FFT order	512

Table 3. EER results of synthetic detection using different features

Feature	EER (%)
MFCC	10.98
MGDCC	1.25
MM	19.29
PM	13.71

To evaluate the performance of proposed method, we adopt equal error rate (EER) as the evaluation measure. In this study, the EER is the error rate obtained when the percentage of natural speech wrongly classified to synthetic speech is equal to the percentage of synthetic speech wrongly classified to natural speech. The EER can be obtained by having a threshold other than 0 for equation (4). The lower value of EER, the better performance.

We first examine the performance of using different types of features individually. The results are presented in Table 3. It is observed that MGDCC produces much lower EER than MFCC, which confirms that artifacts are introduced in the phase spectrum in the converted speech [11]. This is also confirmed in long-term modulation features. Phase modulation (PM) which is derived from modified group delay function phase spectrogram achieves better performance than magnitude modulation (MM) which is computed from magnitude spectrogram. It is also observed that both magnitude and phase modulation features does not produces good classification performance.

We then check the performance of the log-likelihood score fusion of short-term spectral and long-term modulation features. The results with varying weighting coefficient α are shown in Table 4. For example, MFCC+MGDCC refers to the the log-likelihood score fusion of MFCC and MGDCC features, and MFCC and MGDCC correspond to A and B in equation (5), respectively. From the results, it is observed that MFCC+MM and MFCC+PM reduce the EER of using only MFCC from 10.98 to 8.51 and 7.17, respectively. EER of using only MGDCC is reduced from 1.25 to 0.98 and 0.89 by using MGDCC+MM and MGDCC+PM, respectively. These results show that MM and PM, which are long-term temporal features, have complementary information to the short-term spectral features: MFCC and MGDCC.

We can also find that even though both MFCC and MGDCC obtain lower EER than MM and PM, the combination of MFCC and MGDCC features gets higher EER than the combination of MGDCC and MM or the combination of MGDCC and PM. This shows that long-term temporal features have more complementary information than MFCC feature to MGDCC, as both MFCC and MGDCC are short-term spectral features and can not capture the temporal distortions. The usefulness of modulation features, including magnitude modulation feature and phase modulation feature, confirms that temporal artifacts are introduced during the frame-by-frame operation in speech analysis and synthesis process.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed to combine short-term and long-term temporal modulation features for discriminating synthetic speech and human speech. The experiments results show that the fusion of long-term modulation and short-term spectral features can achieve better performance than using that only MFCC or MGDCC features. In addition to artifacts in short-term spectral, this finding confirms that artifacts in temporal structure are also introduced during the frame-based operation in the analysis-synthesis process as illustrated in Fig. 1. The proposed method can be adopted to detect synthetic speech generated from both HMM-based speech synthesis systems and voice conversion systems which are using vocoder techniques.

We note that the dimensionality reduction and feature selection of the modulation features are important, and they affect the results a lot. As the feature selection from modulation features is not the slope of this paper, we only adopt the simple PCA technique to reduce dimensionality. In addition, although the use of filter-bank energies can reduce the dimensionality of modulation super-vector, the filter-bank operation may result in the loss of some detail information in modulation features. Therefore, we will investigate feature selection and dimensionality reduction techniques to extract robust modulation feature in future work.

8. REFERENCES

[1] Joseph P Campbell Jr, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 2, pp. 131–142, 1998.

[3] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 66–83, 2009.

[4] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi, "On the security of hmm-based speaker verification systems against imposture using synthetic speech," in *Proceedings of the European Conference on Speech Communication and Technology*, 1999, vol. 3, pp. 1223–1226.

[5] T. Masuko, K. Tokuda, and T. Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," in *Proc. ICSLP*, 2000, vol. 2, pp. 302–305.

[6] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an hmm-based speech synthesis system," in *Proc. Eurospeech*, 2001.

[7] P.L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of hmm-based synthetic speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2280–2290, 2012.

[8] Q. Jin, A.R. Toth, A.W. Black, and T. Schultz, "Is voice transformation a threat to speaker identification?," in *ICASSP 2008*. IEEE, 2008, pp. 4845–4848.

[9] Z. Wu, T. Kinnunen, E.S. Chng, H. Li, and E. Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," in *Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012*, 2012.

[10] T. Kinnunen, Z.Z. Wu, K.A. Lee, F. Sedlak, E.S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4401–4404.

[11] Z. Wu, E.S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Interspeech*, 2012.

[12] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, pp. 117–132, 1998.

[13] K.A. Lee T. Kinnunen and H. Li, "Dimension reduction of the modulation spectrogram for speaker verification," in *Odyssey: The Speaker and Language Recognition Workshop*, 2008.

[14] S. Sam, X. Xiao, L. Besacier, E. Castelli, H. Li, and E. S. Chng, "Speech modulation features for robust nonnative speech accent detection," in *Interspeech*, 2011.

[15] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[16] D. Zhu and K.K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *ICASSP 2004*.

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