Interspeech 2018
2-6 September 2018, Hyderabad

Multiple Phase Information Combination for Replay Attacks Detection

Dongbo Li[1], *Longbiao Wang*[1,*], *Jianwu Dang*[1,2,*], *Meng Liu*[1], *Zeyan Oo*[3], *Seiichi Nakagawa*[4], *Haotian Guan*[5], *Xiangang Li*[6]

[1]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin, China
[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[3]Nagaoka University of Technology, Nagaoka, Japan
[4]Chubu University, Kasugai, Japan
[5]Intelligent Spoken Language Technology (Tianjin) Co., Ltd., Tianjin, China
[6]AI Labs, Didi Chuxing, Beijing, China

{li_dongbo,longbiao_wang,htguan}@tju.edu.cn, jdang@jaist.ac.jp,
lemon_liumeng@163.com,zeyanoo90@gmail.com, lixiangang@didichuxing.com

Abstract

In recent years, the performance of Automatic Speaker Verification (ASV) systems has been improved significantly. However, they are still affected by different kind of spoofing attacks. In this paper, we propose a method that fused different phase features and amplitude features to detect replay attacks. We apply the mel-scale relative phase feature and source-filter vocal tract feature in phase domain for replay attacks detection. These two phase-based features are combined to get complementary information. In addition to these phase characteristics, constant Q cepstral coefficients (CQCCs) are used. The proposed methods are evaluated using the ASVspoof 2017 challenge database, and Gaussian mixture model was used as the back-end model. The proposed approach achieved 55.6% relative error reduction rate than the conventional magnitude-based feature.

Index Terms: anti-spoofing, replay attacks detection, phase information, feature combination, CQCCs

1. Introduction

Automatic speaker verification (ASV) is a technique to verify that speech belongs to a given speaker [1]. In the current years, many significant advances have been made in the field of speaker recognition and verification [2, 3, 4]. The security and reliability of ASVs are the main aspects of technical challenges. The outside world can deceive the trust of the system by imitating the characteristics of speech, and most authentication systems are easily affected. Spoofing attacks and anti-spoofing attacks have been one of the battlefields in the speech area. At present, more and more researchers are beginning to pay attention to the vulnerability of ASV systems [5, 6, 7]. Deceptive attacks can be separated into four categories: impersonation, replay, text-to-speech (TTS) synthesis and speech conversion [8]. In previous studies, the focus of speech detection was mainly on synthesized speech and converted speech. Countermeasures on replay tracks have not been researched deeply owing to lack of publicly available databases and standardized benchmarks. The ASV Spoofing and Countermeasures (ASVspoof) 2017 provides a platform that focused on detecting replay attacks and provides a reference system for replaying attacks detection, a standard database, and a common evaluation standard whose purpose is to build a common framework. With the beginning of the challenge, replay attack detection has made significant progress [9, 10, 11].

Most of the methods proposed in previous studies have focused on amplitude information, such as Mel-Frequency Cepstral Coefficients feature (MFCCs) [12]. MFCCs are widely used in speech applications, such as automatic speech recognition (ASR), speaker verification (ASV) and language recognition [13, 14, 15]. But the Fourier transform used in MFCCs is not ideal for spoofing detection because the frequency bins are processed similarly, and it ignores the resolution of different frequency bins [9]. A new constant Q cepstral coefficient feature based on the constant Q transform (CQT) was proposed to detect various kinds of spoofing attacks [1, 16]. The difference between with MFCCs is that it processed different frequencies with variable resolution that is higher frequency has higher resolution and lower frequency has lower resolution. It is shown in [5, 8, 17] that CQCCs outperforms many previously reported features by a significant margin against both known and unknown attacks.

Phase-based features have shown effective for synthesized and converted speech detection [12, 18, 19, 20]. The most commonly used phase related feature may be the group delay based feature. In fact, this feature contains phase and amplitude information during the calculation process, which means that the detection result may be disturbed. The relative phase can avoid this problem and it is extracted directly from the Fourier transform of the speech wave [18]. To reduce the phase variation by cutting positions, the phase of a certain base frequency is kept constant, and the phases of other frequencies are estimated relative to this. However, in the related phase studies, filter banks have not been applied to improve the performance of spoofing detection. In this paper, we utilize the mel filter bank for filtering, which corresponds to better resolution at low frequencies and less at high. In addition, for the shortcomings of the traditional group delay method, this experiment uses phase domain source-filter separated vocal tract feature [21] to apply to spoofing detection. This method utilizes the group delay method in the phase-domain to avoid the limitations of the traditional group delay functionc. And this feature was first used for spoofing detection task. We also explore the role of different phase information in the replay attacks detection, and combine these phase features with amplitude feature to achieve the best system performance.

The remainder of this paper is organized as follows. In Section 2, baseline system of feature and decision model is

*Corresponding author

10.21437/Interspeech.2018-2001

described. Section 3 describes phase-based feature extraction methods. The experimental setup and results are shown in Section 4. Finally, conclusions are given in Section 5.

2. Baseline system

In this paper, a Gaussian mixture model (GMM) [22, 23] is used as spoofed speech detector. CQCC is used as baseline feature.

2.1. CQCC feature

Compared with the MFCC feature, the CQCC feature [1] is the variable resolution of the spectrum and the time-frequency representation is very efficient in spoof detection, which is more suitable for the ASVspoof task. This feature is used as a benchmark feature and it has proven to be an effective technology in the ASVspoof system. Hence, in this study, CQCC is adopted as the baseline feature.

The CQCC is an amplitude-based feature which combines Constant Q transform (CQT) and traditional cepstral analysis. The object is to transform geometric space of frequency bins to a linear space using performing a linear frequency scale of the CQT, followed by a uniform resampling and a Discrete Cosine Transform (DCT). Extraction of CQCC features is illustrated in below Figure 1.

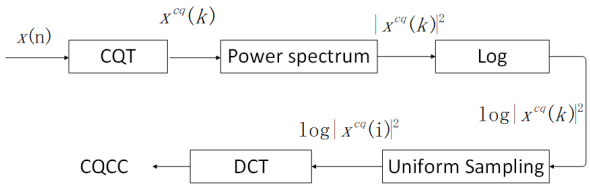

Figure 1: *CQCC feature extraction process.*

2.2. GMM-based spoofed speech detector

Score for decision is derived using the difference between the log-likelihoods of the genuine and spoofed GMMs using the following equation:

$$S = \log(P(X|\theta_g)) - \log(P(X|\theta_s)), \tag{1}$$

where P is the likelihood function, X is the sequence of feature vectors, θ_g and θ_s are parameters for the genuine and spoofed models, respectively. The decision about whether a given segment is genuine speech or spoofed speech is built on score S.

In this work, a variety of features are utilized to complement each other to improve the robustness. To better combine the phase information with the amplitude information, we use the method proposed in [24] that combines the information between the two systems at the score level to obtain the information gain and improves the final result of the combined system. Through the information fusion at the score level, the advantages of both phase and amplitude features can be emphasized.

For two score combination, we have used linear combination proposed in [24].

$$\begin{aligned} L_{comp} &= (1-\alpha)L_1 + \alpha L_2 \\ \alpha &= \frac{\overline{L_1}}{\overline{L_1} + \overline{L_2}} \end{aligned} \tag{2}$$

L_1 and L_2 represent the scores from two independent models. $\overline{L_1}$ and $\overline{L_2}$ denote the averaged L_1 and L_2 over all training data respectively. For example, the combination of CQCC and MGDDC features at the score level, the L_1 and L_2 denote the CQCC score and MGDCC score, respectively.

Furthermore, we use the following formula to calculate score when we use three features to combination.

$$\begin{aligned} L_{comp} &= \alpha L_1 + \beta L_2 + (1-\alpha-\beta)L_3 \\ \alpha &= \frac{\overline{L_1}}{\overline{L_1} + \overline{L_2} + \overline{L_3}}, \beta = \frac{\overline{L_2}}{\overline{L_1} + \overline{L_2} + \overline{L_3}} \end{aligned} \tag{3}$$

3. Phase-based feature extraction

3.1. Modified group delay cepstral coefficient

Most of the phase-related works in speech processing are based on the Modified Group Delay Function. The group delay function (GDF) [25] is defined as the frequency differential of the phase spectrum, that is,

$$\tau_X(\omega) = -\frac{d}{d\omega}\arg[X(\omega)] = -Im\{\frac{d}{d\omega}\log(X(\omega))\}, \tag{4}$$

where $\arg[.]$ and $Im\{.\}$ denote the unwrapped (continuous) phase and imaginary part and ω is angular frequency. Phase unwrapping is not straightforward but the GDF can be computed while avoiding this issue by utilizing real and imaginary parts, The group delay function can also be calculated directly from the speech signal using

$$\tau_X(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2}, \tag{5}$$

where the subscripts R and I denote the real and imaginary parts of the Fourier transform. $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of $x(n)$ and $nx(n)$, respectively. There are many studies reporting that the modified group delay is better than the original group delay [26]. The modified group delay function can be defined as

$$\begin{aligned} \tau(\omega) &= \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{S_c(\omega)^{2\beta}} \\ \tau_m(\omega) &= (\frac{\tau(\omega)}{|\tau(\omega)|})(|\tau(\omega)|)^{\alpha}. \end{aligned} \tag{6}$$

Where $S_c(\omega)$ is the cepstrally smoothed spectrum of $S(\omega)$ and $S(\omega)$ is the squared magnitude $|X(\omega)|^2$ of the signal $x(n)$. The modified group delay function is smoothed to avoid explosion. There are two parameters (α, β) that need to be adjusted to suit different tasks. The cepstral coefficient of the above function was obtained by taking DCT as in [19]. In the experiments we set the values to α = 0.9 $(0 < \alpha \leq 1.0)$ and $\beta = 0.4$ $(0 < \beta \leq 1.0)$.

3.2. Phase domain source-filter separation based vocal tract feature

Speech is a mixed-phase signal, as its complex cepstrum is neither causal nor anti-causal. Therefore, in order to facilitate the phase information in speech processing, it can be divided into two parts, minimum-phase ($MinPh$), $X_{MinPh}(\omega)$, all-pass ($AllP$), $X_{AllP}(\omega)$.

$$\begin{aligned} X(\omega) &= X_{MinPh}(\omega)X_{AllP}(\omega) \\ |X(\omega)| &= |X_{MinPh}(\omega)| \\ arg[X(\omega)] &= arg[X_{MinPh}(\omega)] + arg[X_{AllP}(\omega)], \end{aligned} \tag{7}$$

where $|X(\omega)|$ and $arg[X(\omega)]$ indicate the (short-time) magnitude and unwrapped (continuous) phase spectra, respectively.

$$|X(\omega)| = |X_{VT}(\omega)||X_{Exc}(\omega)| = |X_{MinPh}(\omega)|$$
$$arg[X_{MinPh}(\omega)] = -\frac{1}{2\pi}\log(|X_{MinPh}(\omega)|) * \cot(\frac{\omega}{2}) \quad (8)$$

Since vocal tract ($X_{VT}(\omega)$) and excitation ($X_{Exc}(\omega)$) components are convolved in the time domain, the magnitude spectrum ($|X(\omega)|$) is the product of the corresponding magnitude spectra. Given $|X(\omega)|$ is only linked to the $MinPh$ part. In Eq.8, by replacing the $log|X_{MinPh}(\omega)|$ with $log|X(\omega)|$, $arg[X_{MinPh}(\omega)]$ can be calculated.

Phase characteristics are no longer chaotic and can be understood as a superposition of two components: the vocal tract and excitation parts. The same is shown in Eq.9.

$$\begin{aligned} arg[X_{MinPh}(\omega)] &= -\frac{1}{2\pi}\log(|X_{VT}(\omega)||X_{Exc}(\omega)|) * \cot(\frac{\omega}{2}) \\ &= arg[X_{VT}(\omega)] + arg[X_{Exc}(\omega)] \end{aligned} \quad (9)$$

The above-described method of separating and extracting features by source-filter separation based on the phase domain and it can extract the vocal tract feature and the excitation parts information in the speech. In this work, the experiment used the characteristics of the vocal tract part [21, 27], which had proven to be a good performance in speech recognition and has not yet been applied in replay attacks detection. This feature is known as PBSFVT. The feature extraction method is shown in Figure 2.

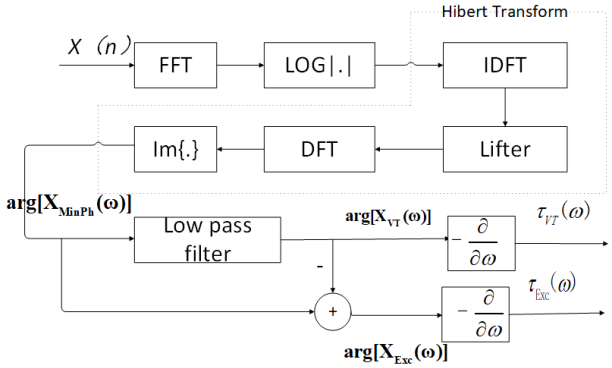

Figure 2: *Phase-based source-filter decomposition [21].*

3.3. Mel-scale relative phase features

The phase changes depending on the clipping position of the input speech even at the same frequency ω. To overcome this problem, the phase of a certain base frequency ω is kept constant, and the phases of other frequencies are estimated relative to this. For example, by setting the base frequency ω to 0, we obtain:

$$X'(\omega) = |X(\omega)| \times e^{j\theta(\omega)} \times e^{j(-\theta(\omega))}, \quad (10)$$

whereas for the other frequency $\omega' = 2\pi f'$, the spectrum becomes:

$$X'(\omega') = |X'(\omega')| \times e^{j\theta(\omega')} \times e^{j\frac{\omega'}{\omega}(-\theta(\omega))}, \quad (11)$$

In this way, the phase can be normalized, and the normalized phase information becomes

$$\tilde{\theta}(\omega') = \theta(\omega') + \frac{\omega'}{\omega}(-\theta(\omega)). \quad (12)$$

After that, we use the method proposed in [28] to process the phase information and change the phase to the coordinates on the unit circle, $\tilde{\theta}$ is converted to $\{\cos\tilde{\theta}, \sin\tilde{\theta}\}$. The relative phase (RP) features were first introduce in [28], and this relative phase feature is used in [18]. Once the process described in [19] is complete, we convert the phase information to Mel scale in this paper. It corresponds to better resolution at low frequencies and less at high.

4. Experiments

4.1. Data and evaluation metric

Source of the ASVspoof 2017 challenge is the RedDots database, the genuine recorded data in the RedDots database as source data for the challenge, and replayed data in the RedDots database as source of spoof replay recordings.

The challenge data are divided into three subsets: training, development and evaluation. Each voice in the training set and the development set has tagged as genuine or spoof. Information regarding phrase ID, speaker, recording environment, recording device and playback device was also available. For the evaluation subset, only the phrase ID was available. The sampling rate of data in this dataset is set to 16 kHz and the sampling precision is 16 bits. Common training set containing 1,508 genuine and 1,508 spoof files was used in our system, and the following systems were trained with only the common training set. In this experiment, we use training data to train the model, utilize the development set to adjust the training parameters, and finally use the evaluation set to evaluate our proposed method. Table 1 shows the detailed data distribution of the ASVspoof 2017 challenge.

Table 1: *ASV-Spoof 2017 Dataset*

Data type	Number of speakers	Utterances	
		Genuine	Spoofed
Train	10	1508	1508
Development	8	760	950
Evaluation	24	1298	12008

Equal Error Rate (EER) is invoked as the performance measures in the ASVspoof 2017 challenge, which is also the evaluation metric in our experiments.

4.2. Experimental setup

For CQCC feature, we use default 96 bins-per-octave and 16 as a number of uniform samples in the first octave. For MGDCC feature, a total of 36 dimensions (12 MDGCC, 12 ΔMGDCC, 12 ΔΔMGDCC) were calculated from the modified group delay function phase spectrum every 10 ms with a window of 25 ms. Thirty-six dimensional PBSFVT feature (12 PBSFVT, 12 ΔPBSFVT, 12 ΔΔPBSFVT) was calculated every 10 ms with a window of 25 ms. Mel-RP feature was calculated every 5 ms with a window of 12.5 ms. A spectrum with 128 components consisting of magnitude and phase was calculated by DFT for every 256 samples. Then 38 static relative phase features (that is, 19 $\cos\tilde{\theta}$ and 19 $\sin\tilde{\theta}$) were extracted using mel-scale filter.

The GMM has two 512-component models which are trained using the EM (Expectation Maximization) algorithm, on genuine and spoof utterances, respectively.

4.3. Experimental results and discussion

First of all, we conducted experiments using the baseline GMM classifier using a different individual type of features based on GMM detector. The detail results are summarized in Table 2.

Table 2: *EERs (%) of spoofing detection performance of individual features*

Feature	Development data	Evaluation data
CQCC	10.35	29.00
MFCC	13.78	34.39
MGDCC	25.93	40.84
PBSFVT	16.18	26.58
RP	19.86	25.68
Mel-RP	10.36	16.03

From Table 2, compared with baseline CQCC features, Mel-RP and PBSFVT features achieved 44.7% and 8.6% relative error reduction rates for evaluation data, respectively. However, the result of the MGDCC feature in this experiment is not as good as the expected. This may be due to the fact that both the amplitude and phase information are included in the MGDCC feature, disrupting the final performance.

In addition, the PBFSVT feature also avoids the defects of MGDCC because the processing of this feature is mainly performed in the phase domain. At the same time, the Mel-RP have been improved compared to the traditional RP features, and the reason is that mel filter banks for traditional RP features have better resolution at low frequencies and less at high.

Table 3: *EER (%) of score combination system*

Feature	Development data	Evaluation data
CQCC+MGDCC	9.03	30.76
CQCC+PBSFVT	9.48	21.77
CQCC+Mel-RP	5.02	13.88
CQCC+MGDCC+Mel-RP	9.94	35.17
CQCC+MGDCC+PBSFVT	9.70	22.36
CQCC+PBSFVT+Mel-RP	5.69	12.88

Next, the score level combinations of different systems were conducted and the results are shown in Table 3. The CQCC+Mel-RP and CQCC+PBSFVT achieved relative error reduction rates of 54.9% and 29.2% compared with CQCC+MGDCC, respectively, This may be due to the complex information of the combination of MGDCC and CQCC handling too much for simple GMM. By comparing different experimental results, the CQCC + Mel-RP + PBSFVT combination system achieved the best performance of 12.88%. The result is obtained using both phase and amplitude characteristics in various resolutions. The CQCC+Mel-RP+PBSFVT combination system selects these phase-based and magnitude-based features to take advantages of all resolution information.

5. Conclusions

In this paper, we proposed a method to fuse different phase features and amplitude features to detect replay attacks. It is the first work to apply the mel filter bank to the relative phase for improving the performance of spoofing detection and the first time for application of PBSFVT to replay attack detection. The experimental results show that the Mel-RP features and PBSFVT features proposed in this paper can effectively improve the performance of the system. Individual Mel-RP feature achieves a relative error reduction rate of 44.7% comparing with the CQCC feature. The combined system CQCC+Mel-RP+PBSFVT achieves a relative error reduction rate of 55.6% comparing with the conventional magnitude-based feature.

In the future we have plan to use deep learning algorithm such as DNN or CNN which have recently provided a good performance on ASV systems.

6. Acknowledgements

The research was supported partially by the National Natural Science Foundation of China (No. 61771333 and No. U1736219), JSPS KAKENHI Grant (No. 16K12461 and No. 16K00297) and Didi Research Collaboration Plan.

7. References

[1] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Speaker Odyssey Workshop, Bilbao, Spain*, vol. 25, 2016, pp. 249–252.

[2] W. Rao, M.-W. Mak, and K.-A. Lee, "Normalization of total variability matrix for i-vector/plda speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4180–4184.

[3] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5115–5119.

[4] N. W. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification." in *Interspeech*, 2013, pp. 925–929.

[5] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[6] Z.-F. Wang, G. Wei, and Q.-H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, vol. 4. IEEE, 2011, pp. 1708–1713.

[7] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.

[8] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014, pp. 1–5.

[9] X. Wang, Y. Xiao, and X. Zhu, "Feature selection based on CQCCs for automatic speaker verification spoofing," *Proc. Interspeech 2017*, pp. 32–36, 2017.

[10] R. Font, J. M. Espın, and M. J. Cano, "Experimental analysis of features for replay attack detection–results on the asvspoof 2017 challenge," *Proc. Interspeech 2017*, pp. 7–11, 2017.

[11] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Gałka, "Audio replay attack detection using high-frequency features," *Proc. Interspeech 2017*, pp. 27–31, 2017.

[12] L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, "Speaker identification by combining MFCC and phase information in noisy environments," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4502–4505.

[13] M. Hébert, "Text-dependent speaker recognition," in *Springer handbook of speech processing*. Springer, 2008, pp. 743–762.

[14] M. J. Alam, P. Kenny, G. Bhattacharya, and T. Stafylakis, "Development of CRIM system for the automatic speaker verification spoofing and countermeasures challenge 2015," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[15] A. Janicki, "Spoofing countermeasure based on analysis of linear prediction error," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[16] J. C. Brown, "Calculation of a constant Q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[17] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[18] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[19] P. Rajan, S. H. K. Parthasarathi, and H. A. Murthy, "Robustness of phase based features for speaker recognition," Idiap, Tech. Rep., 2009.

[20] L. Wang, S. Nakagawa, Z. Zhang, Y. Yoshida, and Y. Kawakami, "Spoofing speech detection using modified relative phase information," *IEEE Journal of selected topics in signal processing*, vol. 11, no. 4, pp. 660–670, 2017.

[21] E. Loweimi, J. Barker, and T. Hain, "Source-filter separation of speech signal in the phase domain," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[22] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[23] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM," *Speech communication*, vol. 49, no. 6, pp. 501–513, 2007.

[24] K. Phapatanaburi, L. Wang, Z. Oo, W. Li, S. Nakagawa, and M. Iwahashi, "Noise robust voice activity detection using joint phase and magnitude based feature enhancement," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, no. 6, pp. 845–859, 2017.

[25] 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.

[26] R. M. Hegde, H. A. Murthy, and G. R. Rao, "Application of the modified group delay function to speaker identification and discrimination," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–517.

[27] E. Loweimi, J. Barker, O. S. Torralba, and T. Hain, "Robust source-filter separation of speech signal in the phase domain," *Proc. Interspeech 2017*, pp. 414–418, 2017.

[28] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 4, pp. 1085–1095, 2012.