Interspeech 2018
2-6 September 2018, Hyderabad

Speech Emotion Recognition by Combining Amplitude and Phase Information Using Convolutional Neural Network

Lili Guo[1], *Longbiao Wang*[1,*], *Jianwu Dang*[1,2,*], *Linjuan Zhang*[1], *Haotian Guan*[3], *Xiangang Li*[4]

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

{liliguo, longbiao_wang, linjuanzhang, htguan}@tju.edu.cn, jdang@jaist.ac.jp,
lixiangang@didichuxing.com

Abstract

Previous studies of speech emotion recognition utilize convolutional neural network (CNN) directly on amplitude spectrogram to extract features. CNN combines with bidirectional long short term memory (BLSTM) has become the state-of-the-art model. However, phase information has been ignored in this model. The importance of phase information in speech processing field is gathering attention. In this paper, we propose feature extraction of amplitude spectrogram and phase information using CNN for speech emotion recognition. The modified group delay cepstral coefficient (MGDCC) and relative phase are used as phase information. Firstly, we analyze the influence of phase information on speech emotion recognition. Then we design a CNN-based feature representation using amplitude and phase information. Finally, experiments were conducted on EmoDB to validate the effectiveness of phase information. Integrating amplitude spectrogram with phase information, the relative emotion error recognition rates are reduced by over 33% in comparison with using only amplitude-based feature.

Index Terms: speech emotion recognition, amplitude, phase information, convolutional neural network

1. Introduction

Speech emotion is important to understand users intention in human-computer interaction, so accurately distinguish users emotion can provide great interactivity. However, speech emotion recognition is still a challenging task because we cannot clearly know which features are effective for emotion recognition [1]. In addition, there is no unified way to express emotions, so features should have good robustness for different express ways.

Conventional methods for speech emotion recognition are selecting heuristic features (such pitch, energy, etc.) [2] and then training methods such as support vector machine (SVM) and bidirectional long short term memory (BLSTM) to distinguish emotions [3]. However, it is difficult to select effective features just based on priori knowledge, and it will take much time in selecting features [4]. To solve those problems, convolutional neural network (CNN) was used to extract features [5]. [6] utilized CNN to extract features from amplitude spectrogram, and then SVM was used as the classifier. [7] and [8] proposed a hybrid CNN-BLSTM model directly on amplitude spectrogram, and CNN-BLSTM has become the state-of-the-art approach at present. However, the phase information has been ignored at above speech emotion recognition approaches even in the state-of-the-art method.

As its complicated structure and difficulties in phase wrapping [9], the phase data is ignored in many applications such as emotion recognition. In recent years, the phase information in speech processing field is gathering attention [10]. The most commonly used phase related feature is the group delay based feature [11, 12] which can simply manipulate the phase information. Group delay is defined as the negative derivative of the phase of the Fourier transform of a signal. Hegde et al. proposed modified group delay cepstrral coefficients (MGDCC) which is better than the original group delay [13, 14]. Wang et al. proposed phase normalization method which expresses the phase difference from base-phase value [15, 16, 17, 18, 19], and this is called relative phase which directly extracted from the Fourier transform of the speech wave. A variety of studies have reported the importance of the phase information for different audio processing applications, including speech recognition [14], speech enhancement [20], speaker recognition [21, 22]. However, there are few studies on speech emotion recognition.

Deng et al. [23] exploited phase-based features for whispered speech emotion recognition. They combined Mel-Frequency cepstral coefficients (MFCC) with group delay based features [24]. SVM was used as classifier in this paper. There are some problems exist in this study. On the one hand, they just used the shallow model that cannot extract effective information from phase data, and SVM as a static classifier cannot utilize dynamic information of speech. On the other hand, group delay based phase contains amplitude spectrogram [11, 12, 13], it's difficulty to only analyze the effects of phase data. To explore whether phase data can perform well on deep learning framework, and whether phase data and amplitude spectrogram are complementary, in this paper, we propose feature extraction of amplitude spectrogram and phase information using CNN for speech emotion recognition. The CNN is expected to be able to extract features from both amplitude and phase information simultaneously in one network. BLSTM is used as classifier, which can utilize the context information. In addition, to explore the complementarity between different types of phase data, we adopt MGDCC and relative phase in this work.

The remainder of this paper is organized as follows: We analyze the influence of phase information on speech emotion recognition, and introduces the phase information extraction in Section 2. Our model that combining amplitude and phase information using convolutional neural network is proposed in Section 3. The experimental setup and results are reported in Section 4, and Section 5 presents the conclusions.

[*]Corresponding author

10.21437/Interspeech.2018-2156

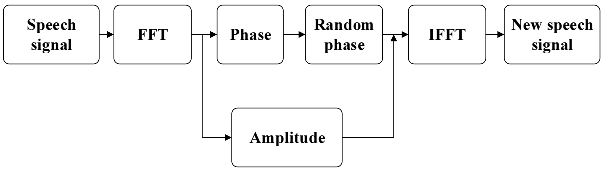
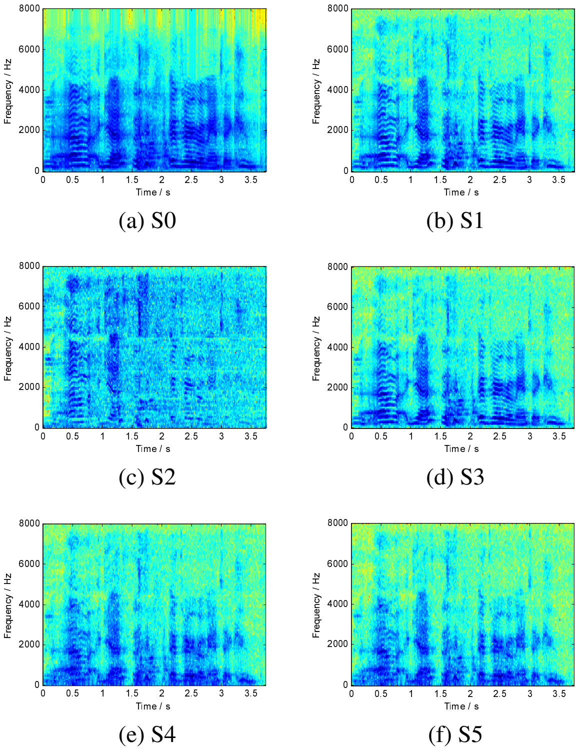


Figure 1: *The process of changing phase data.*

2. Phase information analysis and extraction

2.1. Phase analysis for speech emotion recognition

To analyze the influence of phase data on speech emotion recognition, we use random phase data to replace the original phase data. The detailed procedure is shown as Figure 1. Firstly, we conduct Fourier transform on speech signal to get the phase data and amplitude. Then we use random phase which has the same size as the original phase to replace the original phase data. Finally, the random phase and amplitude are used for inverse Fourier transform, and then gets a new speech signal.



(a) S0

(b) S1

(c) S2

(d) S3

(e) S4

(f) S5

Figure 2: *Spectrogram of different speech signal.*

We use an emotion utterance to make the concrete analysis. As speech signal can be divided into vocal tract and vocal source [25], so linear predictive coding (LPC) is used to divide the speech signal S0 into vocal tract S1 and vocal source S2, as shown in Table 1. Firstly, we use the method as Figure 1 to change vocal source's phase data, and the new vocal source with random phase combines with original vocal tract to form a new speech signal S3. Then the new vocal tract with random phase combines with the original vocal source to form a new speech signal S4. Finally, we use new vocal tract and new vocal source to get the new speech signal S5.

Table 1: *Explanation of different speech signal.*

Vocal tract	Vocal source	Speech signal
Original	-	S1
-	Original	S2
Original	Change phase	S3
Change phase	Original	S4
Change phase	Change phase	S5

As using deep learning to extract features from spectrogram is the most commonly used method, and spectrogram contains useful information to distinguish emotion, so we give their spectrograms in Figure 2. We can see that Figure 2(b) vocal tract contains more information than Figure 2(c) vocal source which likes noise. Figure 2(d) still contains clear harmonic, which indicates that changing vocal source's phase data has a marginal effect on spectrogram. But when we changing the phase data of vocal tract, harmonic of Figure 2(e) and Figure 2(f) is very vague. The fundamental frequency of Figure 2(e) and Figure 2(f) is significantly different from the original speech signal S0, and fundamental frequency is useful information to distinguish emotion. Accordingly, we can draw a conclusion that phase data is important to for acoustic property of speech sound.

2.2. Phase information extraction

In this paper, we use two kinds of phase information that MGDCC and relative phase.

2.2.1. Modified group delay

The spectrum $X(\omega)$ of a signal is obtained by DFT from an input speech signal $x(n)$, and as formula (1):

$$X(\omega) = |X(\omega)|\, e^{j\theta(\omega)}, \tag{1}$$

where $|X(\omega)|$ is the amplitude and $\theta(\omega)$ is the phase at frequency ω. However, the phase values range from $-\pi$ to π, and the phase likes a noise, which is called phase wrapping. To overcome this problem, many phase processing methods are proposed. The group delay feature is the most popular method to manipulate phase information.

Group delay is defined as the negative derivative of the Fourier transform phase for frequency, and it can avoid phase wrapping problem, that is,

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

The group delay function can also be calculated directly from the speech spectrum using following formula:

$$\tau_x(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2}, \tag{3}$$

here, $X(\omega)$ is the Fourier transform of the signal $x(n)$, $Y(\omega)$ is the Fourier transform of $nx(n)$, subscripts R and I denote the real and imaginary parts of the Fourier transform.

Hegde et al. proposed modified group delay function, and there are many studies reporting that modified group delay is better than the original group delay. The modified group delay function can be defined as follows:

$$\tau_m(\omega) = \left(\frac{\tau(\omega)}{|\tau(\omega)|}\right)\left(|\tau\,(\omega)|\right)^{\alpha}, \tag{4}$$

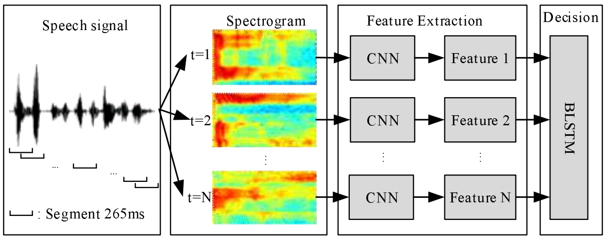


Figure 3: *Structure of CNN-based method on amplitude.*

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

where $S(\omega)$ is the cepstrally smoothed $X(\omega)$, and the range of α and γ are $(0 < \alpha < 1)$, $(0 < \gamma < 1)$.

2.2.2. *Relative phase*

The original phase information changes depending on the clipping position of the input speech even at the same frequency. To overcome this problem, Wang et al. [19] proposed the relative phase which the phase of a certain base frequency ω is kept constant, and the phases of other frequencies are estimated relative to this. Such as, if setting the base frequency ω to 0, we obtain the following formula:

$$X^{'}(\omega) = |X(\omega)| \times e^{j\theta(\omega)} \times e^{j(-\theta(\omega))}, \tag{6}$$

for the other frequency $\omega^{'} = 2\pi f^{'}$, the spectrum becomes

$$X^{'}(\omega^{'}) = |X^{'}(\omega^{'})| \times e^{j\theta(\omega^{'})} \times e^{j\frac{\omega^{'}}{\omega}(-\theta(\omega))}. \tag{7}$$

Finally, the phase information can be normalized, and the normalized phase information as follows:

$$\tilde{\theta}(\omega') = \theta(\omega') + \frac{\omega^{'}}{\omega}(-\theta(\omega)). \tag{8}$$

3. CNN-based feature extraction using phase information

3.1. Conventional CNN-based method using amplitude

In recent years, the most commonly used method for speech emotion recognition is that using CNN on amplitude spectrogram to extract deep features, and then training BSLTM as the decision method. The main idea of BLSTM is utilizing forward direction LSTM and backward LSTM to extract the hidden information in future and past, and the two parts of information forms the final output. BLSTM can utilize the context information which is important in speech processing field [26].

Figure 3 shows the structure of CNN-BLSTM on amplitude spectrogram. Firstly, speech signal is divided into N segments with fixed length. Then it transforms speech signal to amplitude spectrogram by short time Fourier Transform (STFT). For STFT, we use the default values of 256 FFT points, 256 window size and 50% overlap. CNN is used to extract segment-level features from the amplitude spectrogram. For the convenience of training CNN, we transpose the original amplitude matrix. After the transpose, the abscissa becomes frequency, and the ordinate becomes time. Finally, those segment-level features feed to BLSTM to get utterance-level label.

It has become the state-of-the-art method for emotion recognition due to the following reasons. Convolutional neural network (CNN) is adept in extracting local features from raw data [27]. BLSTM can utilize the context information which is important in speech processing field. However, this approach still exists an important problem that the phase information has been ignored, and phase information is gathering attention.

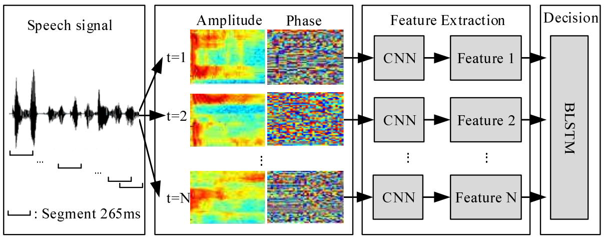


Figure 4: *Structure of CNN-based method on amplitude and phase.*

3.2. CNN-based method using amplitude and phase

From the phase analysis in Section 2.1 we can know that the phase information has important influence on speech emotion recognition. However, phase information contains less (or no) amplitude information, therefore the feature extraction would be combined with amplitude spectrogram. Phase information contains deep relationship with the amplitude spectrogram, and we think CNN could use this relationship to extract more effective features. With this in mind, we propose feature extraction of amplitude spectrogram and phase information using CNN for speech emotion recognition.

Figure 4 shows the whole process of our approach. We use the same method to extract amplitude spectrogram $V1$ as Section 3.1. In addition, for each segment we extract the phase information $V2$ that correlates with the amplitude spectrogram. In this work, we use two types of phase information that MGDCC and relative phase. We combine the amplitude spectrogram with the phase information in one large feature vectors V. The abscissa as frequency and the ordinate as time. The feature vector of t-th segment in i-th utterance is calculated using following formula:

$$V_i^t = [V1_i^t, V2_i^t], \tag{9}$$

where $V1_i^t$ and $V2_i^t$ are the amplitude spectrogram and phase information of t-th segment in i-th utterance, respectively. Then we use CNN to extract deep features from V, and BLSTM is used as the decision method.

4. Experiments

4.1. Experimental setup

We conduct experiments on EmoDB [28] to evaluate our proposed method for speech emotion recognition. EmoDB consists of 535 utterances in German, and all utterances are sampled at 16 KHz with approximately 2-3 seconds long. It contains seven emotions that fear, disgust, happiness, boredom, neutral, sadness and anger with amounts of 69, 46, 71, 81, 79, 62 and 127. We can see it is a small database, so we adopt 10-fold cross-validation in our experiments.

All the features list in Table 2. The first one is the baseline feature with size of 32 × 129 that each segment contains 32 frames and each frame has 129 attributes. The size of relative phase is same as amplitude spectrogram. Relative phase

Table 2: *Feature sizes of one segment.*

ID	Feature	Size
1	Amplitude	32×129
2	Relative phase	32×129
3	MGDCC	32×36
4	Amplitude+relative phase	32×258
5	Amplitude+MGDCC	32×165
6	Amplitude+relative phase+MGDCC	32×294

Table 3: *Weighted and unweighted accuracy for each featrue.*

Feature	WA(%)	UA(%)
Amplitude	87.66	86.66
Relative phase	70.28	68.83
MGDCC	82.80	81.40
Amplitude+relative phase	88.04	87.08
Amplitude+MGDCC	88.79	88.19
Amplitude+relative phase+MGDCC	**91.78**	**91.28**

information is calculated every 8 ms with a window of 16 ms, and the base frequency ω is set to 1000 Hz. In the process of extracting MGDCC, α = 0.1, γ = 0.2 are used. For MGDCCs, a total of 36 dimensions (12 static MGDCCs, 12 ΔMGDCCs and 12 $\Delta\Delta$MGDCCs) are calculated every 8 ms with a window of 16 ms. In this work, firstly, we respectively use amplitude spectrogram, relative phase and MGDCC feature to check their effects for speech emotion recognition, respectively. Finally, we make combinations about the amplitude and phase information, which are our proposed methods.

To choose the optimal structure, we experimented with different numbers of hidden units and layers, learning rate, etc. In the process of training CNN, all segments in one utterance share the label, and we choose cross entropy as the cost function. The structure of CNN contains two convolutional layers and two max-pooling layers. The first convolutional layer uses 32 filters with 5×5 size, and the second convolutional layer uses 64 filters with 5×5 size. The pooling size of the two pooling layers is 2×2. After flatten layer, we adopt a full connected layer with 1024 units. To avoid over-fitting, a dropout layer with 0.5 factor is used before output layer. BLSTM that contains two hidden layers and each layer with 200 units is used.

4.2. Experimental results

Table 3 gives the results for each feature in two common evaluation criteria. Weighted accuracy (WA) is the classification accuracy on the whole test set. Unweighted accuracy (UA) is first calculate the classification accuracy for each emotion and then averaged. From the table we can draw conclusions: 1) The results of using only phase data are acceptable, proving that phase data can perform well on deep learning framework. 2)The results of MGDCC are better than relative phase in this task. We think there are two reasons. Firstly, MGDCC contains amplitude information that is the most commonly used feature in this task. Secondly, MGDCC contains dynamic features (ΔMGDCCs and $\Delta\Delta$MGDCCs) which are important to recognize emotion. 3) The combination of amplitude with relative phase or MGDCC is better than using only amplitude, indicating that the combination of amplitude and phase information is effective. In addition, the combination of amplitude and relative phase significantly outperforms relative phase by 59.76% and 58.55% relative error reduction in WA (70.28% → 88.04%) and UA (68.83% → 87.08%), respectively. However, the combination of amplitude and MGDCC doesn't go up too much compared with MGDCC. We can draw a conclusion that relative phase is more complementary with amplitude than MGDCC. 4) By combining the three features (amplitude, relative phase and MGDCC), the best results are achieved, that is, the relative emotion error recognition rates are reduced by about 33.4% and 34.6% in comparison with using only amplitude feature in WA and UA, respectively. It also outperforms the combination of amplitude and MGDCC by about 26% relative error reduction, indicating that those three features are complementary.

Table 4: *F1 (%) for each emotion. A: Amplitude; RP: Relative phase; MGD: MGDCC; All: A+RP+MGD.*

Emo.	A	RP	MGD	A+ RP	A+ MGD	All
Fea.	91.05	63.16	82.96	87.22	87.22	**94.20**
Dis.	91.96	75.86	82.35	92.13	**95.56**	94.38
Hap.	76.03	54.55	73.91	76.42	78.74	**82.54**
Bor.	85.53	68.51	77.91	88.34	88.89	**90.45**
Neu.	87.34	65.22	79.22	88.89	88.89	**92.02**
Sad.	90.90	81.82	87.10	90.91	89.55	**97.64**
Ang.	89.61	76.92	90.84	90.25	91.51	**91.85**
Aev.	87.49	69.43	82.04	87.74	88.62	**91.87**

To evaluate the effects for each emotion, Table 4 lists F1 results of different features. 1) Integrating amplitude with phase information (relative phase and MGDCC) achieves best performance in most classes of emotions, especially for sadness class (90.90% → 97.64%). 2) When inferring disgust emotion, its result is not the best but still significantly better than amplitude-based feature. The reason should be that disgust class holds the lowest proportion in EmoDB. 3) On average of F1, our approaches (Amplitude+relative phase, Amplitude+MGDCC and Amplitude+relative phase+MGDCC) all get better results than using only amplitude features. The combination of amplitude, relative phase and MGDCC significantly outperforms the baseline feature with 35% relative error reduction.

5. Conclusions

In this work, firstly, we analyzed the influence of phase information on speech emotion recognition, and found that phase information is important to this task. Then we proposed feature extraction of amplitude spectrogram and phase information using CNN. To the best of our knowledge, it is the first work to explore the effective of phase information for speech emotion using deep learning. Experimental results indicate that integrating amplitude spectrogram with phase information significantly outperformed using only amplitude-based feature. In future work, we will make improvements about relative phase such as using filter-bank and applying pitch synchronization.

6. Acknowledgements

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

7. References

[1] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[2] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proceedings of INTERSPEECH*, 2014, pp. 223–227.

[3] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: raising the benchmarks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 5688–5691.

[4] L. Guo, L. Wang, J. Dang, L. Zhang, and H. Guan, "A feature fusion method based on extreme learning machine for speech emotion recognition," in *Proceedings of ICASSP*, 2018, pp. 2666–2670.

[5] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5115–5119.

[6] Z. W. Huang, M. Dong, Q. R. Mao, and Y. Z. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 801–804.

[7] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Signal and Information Processing Association Annual Summit and Conference*, 2016, pp. 1–4.

[8] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proceedings of INTERSPEECH*, 2017, pp. 1089–1093.

[9] B. Yegnanarayana, J. Sreekanth, and A. Rangarajan, "Waveform estimation using group delay processing," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 33, no. 4, pp. 832–836, 1985.

[10] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Interspeech 2014 special session: Phase importance in speech processing applications," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014, pp. 1623–1627.

[11] J. Kua, J. Epps, E. Ambikairajah, and E. H. C. Choi, "Ls regularization of group delay features for speaker recognition," in *Proceedings of INTERSPEECH*, 2009, pp. 2887–2890.

[12] P. Rajan, S. H. K. Parthasarathi, and H. A. Murthy, "Robustness of phase based features for speaker recognition," in *Proceedings of INTERSPEECH*, 2009, pp. 2355–2358.

[13] R. M. Hegde, H. A. Murthy, and G. V. R. Rao, "Application of the modified group delay function to speaker identification and discrimination," in *Proceedings of ICASSP*, 2004, pp. 517–520.

[14] 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, 2006.

[15] S. Nakagawa, K. Asakawa, and L. Wang, "Speaker recognition by combining mfcc and phase information," in *Proceedings of INTERSPEECH*, 2007, pp. 2005–2008.

[16] L. Wang, S. Ohtsuka, and S. Nakagawa, "High improvement of speaker identification and verification by combining mfcc and phase information," in *Proceedings of ICASSP*, 2009, pp. 4529–4532.

[17] L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, "Speaker identification by combining mfcc and phase information in noisy environments," in *Proceedings of ICASSP*, 2010, pp. 4502–4505.

[18] 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.

[19] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Proceedings of INTERSPEECH*, 2015, pp. 2092–2096.

[20] T. Gerkmann, M. Krawczyk-Becker, and J. L. Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, 2015.

[21] Z. Oo, Y. Kawakami, L. Wang, S. Nakagawa, X. Xiao, and M. Iwahashi, "DNN-based amplitude and phase feature enhancement for noise robust speaker identification," in *Proceedings of INTERSPEECH*, 2016, pp. 2204–2208.

[22] 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.

[23] J. Deng, X. Xu, Z. Zhang, S. Frhholz, and B. Schuller, "Exploitation of phase-based features for whispered speech emotion recognition," *IEEE Access*, vol. 4, pp. 4299–4309, 2017.

[24] P. Rajan, T. Kinnunen, C. Hanilci, J. Pohjalainen, and P. Alku, "Using group delay functions from all-pole models for speaker recognition," in *Proceedings of INTERSPEECH*, vol. 5, 2013, pp. 2489–2493.

[25] D. G. Childers and C. F. Wong, "Measuring and modeling vocal source-tract interaction," *IEEE Transactions on Biomedical Engineering*, vol. 41, no. 7, pp. 663–667, 1994.

[26] G. Keren and B. Schuller, "Convolutional rnn: an enhanced model for extracting features from sequential data," in *International Joint Conference on Neural Networks*, 2016, pp. 3412–3419.

[27] J. Donahue, H. Anne, S. Guadarrama *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[28] F. Burkhardt, A. Paeschke, M. Rolfes *et al.*, "A database of german emotional speech," in *Proceedings of INTERSPEECH*, vol. 5, 2005, pp. 1517–1520.