PHASE AWARE DEEP NEURAL NETWORK FOR NOISE ROBUST VOICE ACTIVITY DETECTION

Longbiao Wang[1,*], *Khomdet Phapatanaburi*[2], *Zeyan Oo*[2], *Seiichi Nakagawa*[3],
Masahiro Iwahashi[2], *Jianwu Dang*[1,4,*]

[1]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, China
[2]Nagaoka University of Technology, Japan
[3]Toyohashi University of Technology, Japan
[4]Japan Advanced Institute of Science and Technology, Japan

longbiao_wang@tju.edu.cn, s147009@stn.nagaokaut.ac.jp, nakagawa@tut.ac.jp, jdang@jaist.ac.jp

ABSTRACT

Phase information is ignored for almost all voice activity detection (VAD). To exploit full information in the original signal, this paper proposes a deep neural network (DNN) using magnitude and phase information (that is, phase aware DNN) to achieve better VAD performance. Mel-frequency cepstral coefficient (MFCC), power-normalized cepstral coefficients (PNCC), instantaneous frequency derivative (IF), baseband phase difference (BPD) and modified group delay cepstral coefficient (MGDCC) are used as magnitude and phase information. The proposed methods were evaluated using CENSREC-1-C database under noise condition. The results show that the phase aware DNN significantly outperforms the DNN using only magnitude information. For DNN-based classifier, the equal error rate (EER) was reduced from 23.70% of MFCC, to 20.43% of joint dual magnitude and single phase features (augmenting PNCC, MGDCC and IF), to 19.92% of joint dual phase and single magnitude feature features (augmenting PNCC, MGDCC and BPD). By combining joint dual magnitude and single phase features with joint dual phase and single magnitude features, the EER was reduced to 19.44%.

Index Terms— VAD, phase information, DNN, magnitude information

1. INTRODUCTION

Voice activity detection (VAD) is an active research topic in the field of speech processing because VAD is one key factor that influence the performance of practical speech application, such as speech recognition systems and speech communication systems [1, 2, 3, 4]. Machine-learning based VAD [5, 6] is receiving more and more attention. There are highly competitive to without machine-learning based VAD in the following merit. First, their theoretical bases can grantee the performance of VAD under low signal-to-noise ratio (SNRs). Second, they can merge the advantages of several features by concatenating the acoustic feature better than conventional VAD (feature combination). Final, they can be combined naturally by score combination. In this paper, we apply deep neural network (DNN) based VAD [7] as machine-learning based VAD, to determine speech segment or non speech segment in speech sample.

To classify speech and non speech segment, many features have been considered for machine-learning based VAD. In [8], perceptual linear predictive coefficients (PLP) was proposed to distinguish speech and non-speech segment from speech sample. In [9], Mel-frequency cepstral coefficient (MFCC) was also proposed to distinguish speech and non-speech segment. In [10], concatenating feature from PLP, MFCC, pitch, discrete Fourier transform (DFT), and amplitude modulation spectrograms (AMS) was applied to distinguish speech and non-speech segment. These features have been proven to be powerful for VAD. However, they may have weakness due to missing phase information, which is the half of information present in original signal. DNN-based classifier achieves the best performance for VAD, but almost DNN-based VAD only as magnitude information. The phase information has been proven to be important for many speech processing tasks [11, 12, 13]. Therefore, this study proposes a phase aware DNN which jointly uses magnitude and phase information for noise robust VAD to determine speech segment or non speech segment.

The joint use of magnitude and phase feature was motivated to receive a set of practical features for robust noise VAD under low SNRs, without loss of phase information or magnitude information. To generate jointly magnitude and phase feature, we use both magnitude and phase features in our experiment. In magnitude based features, MFCC and power-normalized cepstral coefficients (PNCC) [14], that provide better result than MFCC for noise robust speech recognition, are used. For phase based features, instanta-

[*] Corresponding author

978-1-5090-6067-2/17/$31.00 ©2017 IEEE

neous frequency derivative (IF) [15] derived from the derivative of the phase along time axis, baseband phase difference (BPD) [16] derived from difference of baseband short time Fourier transform, and modified group delay cepstral coefficient (MGDCC) [17] derived from negative derivative of the Fourier transform phase are applied. These features will be variously concatenated in serial. They are expected to achieve a better performance than single magnitude based feature.

The remainder of this paper is organized as follows: The conventional and proposed DNN-based VAD approaches are described in Section 2. Section 3 presents magnitude and phase based feature extraction used to product jointly magnitude and phase feature. The experimental setup and result are reported in Section 4, and Section 5 presents our conclusions.

2. PHASE AWARE DNN

The flowchart of the VAD system is shown in Figure 1. In this study, a deep neural network (DNN) is used as speech/non-speech detector. The DNN output refers to the posterior probabilities of two classes, including probability of speech segment, ϱ_{speech} and non-speech segment, $\varrho_{non-speech}$. The decision about whether speech is speech or non-speech segment is based on the posterior probability differences:

$$\wedge(I) = P(\varrho_{speech}|I) - P(\varrho_{non-speech}|I) \tag{1}$$

where I is the feature vector of input speech. Power-Normalized cepstral coefficients (PNCC),instantaneous frequency derivative (IF), baseband phase difference (BPD) and modified group delay (MGDCC) described in Section 3 are used. For conventional DNN-based VAD, only magnitude based feature is used as the feature vector of input speech as follows,

$$I = F_{mag} \tag{2}$$

where F_{mag} only uses magnitude based feature. In this paper, MFCC or/and PNCC is/are used. The structure of magnitude based DNN is shown in Figure 2(a). In [7] , conventional DNN-VAD was successfully applied to magnitude feature (MFCC). However, MFCC only contains the magnitude information of the speech, therefore, the speech/non-speech classifier might be incomplete.

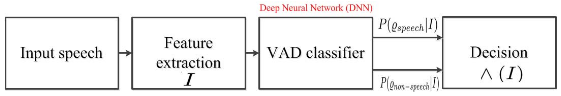

Fig. 1. Flowchart of VAD system.

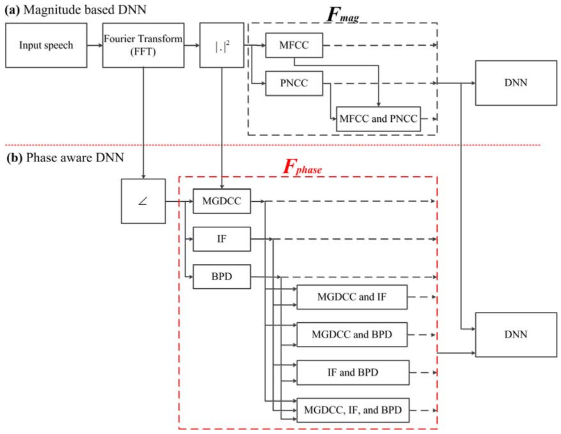

Fig. 2. A block diagram of phase aware DNN.

Recently, the phase information has been proven to be important for many speech processing tasks [11, 18]. In previous work[13], phase feature (MGDCC) augmented with corresponding MFCC could improve the performance of the DNN training as a regression task. This can be seen in improvement in performance in simultaneous enhancement of amplitude and phase feature. With this in mind, we expect that phase information can also improve performance of the DNN training as speech/non-speech classifier. Therefore, this study proposes a phase aware DNN which jointly uses magnitude and phase information for noise robust VAD to determine speech segment or non speech segment. The structure of phase aware DNN is shown in Figure 2(b). The feature vector of input speech, covering magnitude and phase information, is used as follows,

$$I = [F_{mag}, F_{phase}] \tag{3}$$

where F_{phase} is phase based features. MGDCC, IF, BPD, dual phase feature (augmenting MGDCC and IF, MGDCC and BPD, or of IF and BPD), or triple phase features (augmenting MGDCC, IF, BPD) is used to be augmented with magnitude based feature.

3. MAGNITUDE AND PHASE BASED FEATURE

Five features are used in this work. They include two magnitude-based features, namely Mel-frequency cepstral coefficients (MFCC) and Power-Normalized cepstral coefficients (PNCC); and three phase-based features, namely instantaneous frequency derivative (IF), baseband phase difference (BPD), and modified group delay (MGDCC).

- **MFCC** [9] is the most popular feature for speech processing including voice activity detection. We used MFCC as an amplitude feature for the VAD input.
- **PNCC** is another feature based on magnitude information. It has been developed to enhance the robustness of speech recognition systems under noisy condition. The major innovations of this feature when comparing with MFCC are the use of a power-law nonlinearity that replaces the tra-

Table 1. Analysis conditions for MFCC, PNCC, IF, BPD and MGDCC.

	Dimension	Frame length	Frame shift	FFT size
MFCC	39 (13 MFCCs, 13 $\triangle$s, and 13$\triangle\triangle$s)	25 ms	10 ms	256 point
PNCC	39 (13 PNCCs, 13 $\triangle$s, and 13$\triangle\triangle$s)			
IF	127			
BPD	127			
MGDCC	36 (12 MGDCCs, 12 $\triangle$s, and 12$\triangle\triangle$s)			

ditional log nonlinearity used in MFCC coefficients, a noise-suppression algorithm based on asymmetric filtering that suppress background excitation, and a module that accomplishes temporal masking, which in detail is found in [14]. So, this method may be a useful feature for VAD when speech is corrupted by noise.

- **IF** instantaneous frequency [15] is a phase feature designed to provide clearer phase pattern than original phase spectrum that hardly displays any patterns. The computational algorithm is based on derivative of the phase along time axis. Therefore, it can capture the temporal information of phase. Unlike the original phase spectrum that has the problem called phase wrapping, there are better patterns in the IF spectrum, making it possible to be used as a feature.
- **BPD** baseband phase difference [16] is a phase feature extracted from baseband STFT which is different from IF processing using the phase difference between two successive segments, which can also yield significant phase information.
- **MGDCC** modified group delay [17] is a representation of filter phase response, which is defined as the negative derivative of the Fourier transform phase. It is designed to obtain better phase performance than group delay. Two factors, α and γ, are used for control the dynamic range of the modified group delay. Here, we perform the same setting as recommended in [19].

4. EXPERIMENT

Our experiments were conducted on CENSREC-1-C database [20]. The speech data is sampled at 16 kHz, and finally downsampled to 8 kHz. The details of recording condition, utterances, and speaking style are the same in CENSREC-1(AURORA-2J) [21]. As for training data, 104 clean speech data (52 males and 52 females) per one noise environment, which constructed by concatenating several utterances spoken by one speaker (the number of utterances in concatenated speech data is nine or ten), were used to create artificial noisy speech. Each artificial noisy speech is obtained by corrupting each speech data with one of 4 noise types (Subway, Babble, Car, Exhibition) at one of six noise levels of SNR, i.e., 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, and -5 dB. For the test data, 104 speech data per one noise types were corrupted by three unseen noise type, namely, Restaurant, Airport, and Station at three SNR levels: 5 dB, 0 dB, and -5 dB.

4.1. Baseline setup

We applied support vector machine (SVM) -based VAD using the concept based on [9]. The SVM is a binary classifier, which models the decision boundary between the two classes as a separating hyperplane. To rapidly optimize the support vectors, we used the publicly available LIBLINEAR tool [22] which considers linear kernels. For SVM, a grid-search on C and γ using cross-validation is used. Various pairs of (C, γ) values are tried and the one with the best cross-validation accuracy is picked. The regularization parameter is searched from the exponential grid $\{C = 2^{-5}, 2^{-3}, ..., 2^{15}, \gamma = 2^{15}, 2^{-13}, ...2^{3}\}$.

SignalGraph [23] was used to train the DNN. The DNN has one, two, three layers each of which contains 512 neurons. The N-layers of DNN are denoted as DNN-LN. The input feature for DNN contains 9 frames, Cross entropy (CE) was used for a learning criterion. The learning rate was started from 0.1 and changed to 1 for the second epoch, then it was decayed by a factor of 0.5 when the cross entropy on a cross-validation set between two consecutive epochs increases. 39-dimensional MFCC (plus deltas and double deltas) was used for training SVM and DNN.. The equal error rate (EER) is used as measure of VAD result.

Table 2 shows the results of baseline SVM and DNN-VAD. The highlighted contents are the best performance on the averaged noise scenario.

From Table 2, DNN-based VADs outperformed SVM-VAD because DNNs have high feature representation ability to distinguish speech and non-speech accurately. However, DNN-VAD based on more one hidden layer did not perform according to our expectation. It seem that increasing the number of hidden layer for conventional DNN could to yield consistent performance gain for unseen noise types which the same tendency can be found in [10, 24, 25]. This might be due to the different characteristics of unseen noises or the weird architecture of DNN with only two-dimensional output which could not handle the diversified training data well.

4.2. Experimental result of individual feature

In this subsection, we used DNN-VAD using 1 layer, which is the best result from previous subsection, to investigate five types of the feature described in Section 3.

Table 3 shows result of each types of extraction feature. The spectrograms of each feature is shown in Figure 3, where (a) waveform,(b) clean MFCC, (c) clean PNCC, (d) clean IF, (e) clean BPD, and (f) clean MGDCC. The figure on the left is corresponding feature on the right under station noise at 5 dB. Comparing (a-f), the spectrogram of PNCC shows the

Table 2. Performance (EER%) comparison of SVM and DNN classifier using MFCC

Classifier	Restaurant			Airport			Station			Ave. of all noise types			Ave.
	5	0	-5	5	0	-5	5	0	-5	5	0	-5	ALL
SVM	27.36	30.35	38.46	23.95	29.69	34.86	20.84	26.12	34.57	24.05	28.72	35.96	29.58
DNN-L1	14.69	26.69	38.31	14.12	21.15	32.04	10.84	21.84	33.65	13.22	23.23	34.67	**23.70**
DNN L2	14.05	24.04	35.81	16.09	23.89	33.92	12.38	22.57	34.96	14.17	23.50	34.90	24.19
DNN-L3	14.12	23.91	35.81	16.09	24.15	33.95	12.24	22.19	34.81	14.15	23.42	34.86	24.14

Table 3. Performance (EER%) comparison of DNN-VAD using individual feature

	Restaurant			Airport			Station			Ave. of all noise types			Ave.
	5	0	-5	5	0	-5	5	0	-5	5	0	-5	ALL
MFCC	14.69	26.69	38.31	14.12	21.15	32.04	10.84	21.84	33.65	13.22	23.23	34.67	23.70
PNCC	14.35	24.53	36.88	13.51	20.14	31.23	9.07	17.13	27.04	12.31	20.60	31.72	**21.54**
IF	24.53	31.65	39.06	24.04	30.17	38.47	22.60	30.60	39.17	23.72	30.81	38.90	31.14
BPD	19.78	27.41	36.13	22.72	30.37	38.66	23.28	31.32	40.60	21.93	29.70	38.46	30.03
MGDCC	24.84	32.62	41.34	21.45	26.45	34.93	16.57	24.37	33.83	20.95	27.81	36.70	28.49

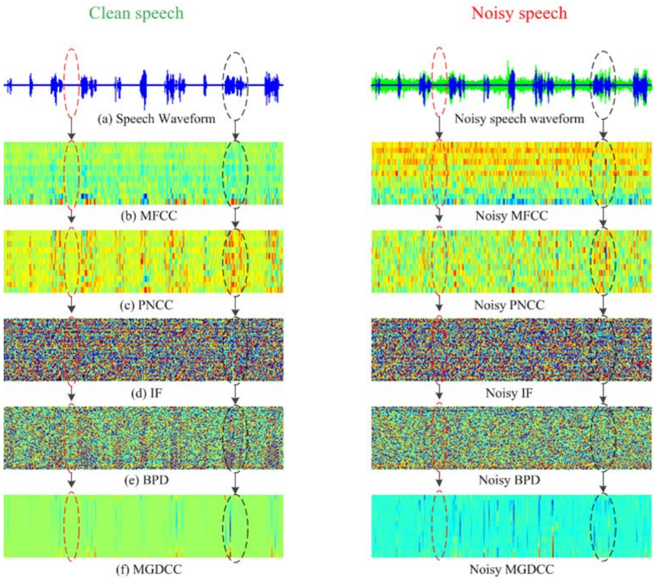

Fig. 3. Spectrograms of five types of feature.

distribution of values with greater variance than all other features. This can be seen in speech segments (highlighted by black dashed lines) and non-speech segments (highlighted by red dashed lines).

From Table 3, we can observe that PNCC outperformed all other feature due to strong of asymmetric noise suppression (ANS) and temporal masking in PNCC processing leading to clear spectrogram compared to other features. In this result, phase based feature worked worse than magnitude based feature in both MFCC and PNCC, but it had some ability of speech and non-speech as observed in Figure 3. Therefore, it might be useful to use phase information to augment magnitude based feature.

4.3. Experimental result of proposed method

According to [25], deep neural network based voice activity detection (VAD) has been proved to be powerful in fusing the advantages of multiple features, especially based on magnitude information. However, fusing the advantages of joint magnitude and phase features information has not been well investigated. In this subsection, we take multiple features contains magnitude and phase information into our experiment. Table 4 shows the result of each multiple features.

From Table 4, it can be seen that multi information based feature provided better performance than individual feature for robust noise VAD. This is due to combining the advantage of multiple features. In feature derived from two s, augmenting PNCC with MGDCC outperformed all of other multiple feature because of exploiting the advantage of robust noise magnitude and phase information. Moreover, the result showed that the feature augmented by dual magnitude features with single feature or augmented by dual phase features with single magnitude features gave better performance by feature augmented by dual feature. This is due to more clear information. However, for multiple feature derived four features, it could not perform our expected result. This might be because additional IF or BPD make feature complicated.

4.4. Score combination

In this subsection, score combination is proposed to exploit the complementary characteristics of these two feature sets. The score ratio (that is differences of speech and non-speech segments) from different kind of joint magnitude and phase features, based on the combination of three features which obtained the best result from the previous section, are combined linearly by following equation.

Table 4. Performance (EER%) comparison of DNN-VAD using different multiple features

	Restaurant			**Airport**			**Station**			**Ave. of all noise types**			**Ave.**
	5	**0**	**-5**	**5**	**0**	**-5**	**5**	**0**	**-5**	**5**	**0**	**-5**	**ALL**
MFCC-PNCC	13.18	24.08	35.69	12.97	19.31	29.38	8.40	16.93	26.70	11.51	20.10	30.59	20.74
MFCC-IF	14.93	26.08	37.15	13.88	20.65	31.18	11.17	21.35	32.49	13.33	22.69	33.61	23.21
MFCC-BPD	15.59	26.31	37.66	13.78	20.31	30.56	10.98	20.66	32.41	13.45	22.43	33.54	23.14
MFCC-MGDCC	14.18	26.18	37.91	13.16	19.53	30.03	10.45	20.15	30.92	12.60	21.95	32.95	22.50
PNCC-IF	18.99	28.34	38.01	18.27	25.99	35.00	12.55	21.64	31.92	16.60	25.32	34.98	25.63
PNCC-BPD	13.47	23.77	35.79	13.12	20.18	31.06	8.87	17.00	27.37	11.82	20.32	31.41	21.18
PNCC-MGDCC	13.18	23.94	35.85	13.21	19.37	30.53	8.26	16.03	25.77	11.55	19.78	30.72	**20.68**
IF-BPD	21.17	28.66	37.12	24.24	31.72	40.40	23.36	31.97	40.81	22.93	30.78	39.44	31.05
IF-MGDCC	21.17	28.67	37.12	24.24	31.72	40.40	23.36	31.97	40.81	22.93	30.78	39.44	31.05
BPD-MGDCC	14.74	25.60	37.81	14.96	22.83	36.07	15.53	25.53	38.72	15.08	24.66	37.53	25.75
PNCC-MGDCC-MFCC	12.72	23.07	34.88	12.89	18.51	28.65	7.73	15.95	24.88	11.11	19.18	29.47	**19.92**
PNCC-MGDCC-IF	12.82	22.45	35.21	13.05	19.24	30.36	8.37	16.23	26.14	11.41	19.31	30.57	20.43
PNCC-MGDCC-BPD	12.92	22.55	35.25	13.09	19.07	30.29	8.33	16.27	26.13	11.45	19.30	30.55	20.43
MFCC-PNCC-MGDCC-IF	12.94	23.00	34.40	12.86	18.73	29.08	8.13	16.44	26.21	11.31	19.39	29.9	20.20
MFCC-PNCC-MGDCC-BPD	13.46	23.88	35.4	12.68	18.53	29.37	8.83	17.50	26.80	11.65	19.97	30.52	20.72
PNCC-MGDCC-BPD-IF	13.05	22.54	34.46	13.69	20.04	30.41	8.89	16.64	26.86	11.88	19.74	30.58	20.73

$$\wedge_{comb}(I) = (1-\beta)\wedge(I_{dmsp}) - (\beta)\wedge(I_{dpsm}), \qquad (4)$$

$$\beta = \frac{\wedge(I_{dpsm})}{\wedge(I_{dpsm)} + \wedge(I_{dmsp})}.$$

where β is a weighing coefficient, $\wedge(I_{dmsp})$ and $\wedge(I_{dpsm})$ denote the score ratio of joint dual magnitude and single phase feature, and of joint dual phase and single magnitude feature, respectively. Table 5 lists the result of score combination.

From Table 5, It can be observed that the score combination of joint magnitude and phased feature outperformed individual phase aware DNN-based VAD system. This is because of combination of complementary characteristics of different features.

5. CONCLUSION

In this paper, we proposed a deep neural network (DNN) using magnitude and phase information (that is, phase aware DNN) to achieve better VAD performance. The proposed methods were evaluated using CENSREC-1-C database under various noise conditions. The results showed that the phase aware DNN significantly outperforms the DNN using only magnitude information. For DNN-based classifier, the equal error rate (EER) was reduced from 23.70% of MFCC, to 20.43% of joint dual magnitude and single phase features (augmenting PNCC, MGDCC and IF), to 19.92% of joint dual phase and single magnitude feature features (augmenting PNCC, MGDCC and BPD). By combining joint dual magnitude and single phase features with joint dual phase and single magnitude features, the EER was reduced to 19.44%.Therefore, we confirmed the effectiveness of phase aware DNN trained by full information in original signal under both phase and magnitude information for noise robust VAD.

In our future work, we will compare joint phase and magnitude feature with feature proposed by [25, 26]. Moreover, we will try to implement and LDA based dimensional reduction for noise robust VAD.

6. ACKNOWLEDGEMENTS

This work was partially supported by the National Basic Research Program of China (No. 2013CB329301), the National Natural Science Foundation of China (No. 61233009) and JSPS KAKENHI Grant Number 16K12461.

7. REFERENCES

[1] D.K. Freeman et al., “The voice activity detector for the pan-european digital cellular mobile telephone service,” in *Proc. ICASSP*, IEEE, 1989, pp. 369–372.

[2] D. Malah, R.V. Cox and A.J Accardi, “Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments,” in *Proc. ICASSP*, IEEE, 1999, vol. 2, pp. 789–792.

[3] V. Hautamäki et al., “Improving speaker verification by periodicity based voice activity detection,” in *Proc. SPECOM*, 2007, pp. 645–650.

[4] H. Sun, B. Ma and H. Li, “Frame selection of interview channel for nist speaker recognition evaluation,” in *Proc. ISCSLP*, IEEE, 2010, pp. 305–308.

[5] J.H. Chang, N.S. Kim and S.K. Mitra, “Voice activity detection based on multiple statistical models,” *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 1965–1976, 2006.

Table 5. Performance (EER%) of score combination

	Restaurant			**Airport**			**Station**			**Ave. of all noise types**			**Ave.**
	5	**0**	**-5**	**5**	**0**	**-5**	**5**	**0**	**-5**	**5**	**0**	**-5**	**ALL**
PNCC-MGDCC-MFCC + PNCC-MGDCC-IF	12.15	21.70	34.12	12.74	18.12	28.45	7.62	15.66	24.43	10.84	18.50	29.00	**19.44**
PNCC-MGDCC-MFCC + PNCC-MGDCC-BPD	12.23	21.78	34.36	12.75	18.29	28.72	7.52	15.60	24.28	10.83	18.55	29.12	19.50

[6] J. Wu and X.L. Zhang, "Efficient multiple kernel support vector machine based voice activity detection," *IEEE Signal Processing Letters*, pp. 466–469, 2011.

[7] N. Ryant, M. Liberman and J. Yuan, "Speech activity detection on youtube using deep neural networks," in *Proc. Interspeech*, 2013, pp. 728–731.

[8] F. Eyben et al., "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies," in *Proc. ICASSP*, IEEE, 2013, pp. 483–487.

[9] T. Kinnunen et al., "Voice activity detection using mfcc features and support vector machine," in *Proc. SPECOM*, 2007, vol. 2, pp. 556–561.

[10] X.L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.

[11] 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.

[12] P. Mowlaee, R. Saeidi and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, 2016.

[13] 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 *Proc. Interspeech*, 2016, pp. 2204–2208.

[14] C. Kim and R.M. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," in *Proc. ICASSP*, IEEE, 2012, pp. 4101–4104.

[15] L.D. Alsteris and K.K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, pp. 578–616, 2007.

[16] M. Krawczyk and T. Gerkmann, "Stft phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1931–1940, 2014.

[17] B. Yegnanarayana and H.A. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Transactions on signal processing*, pp. 2281–2289, 1992.

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

[19] 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.

[20] N. Kitaoka et al., "Censrec-1-c: An evaluation framework for voice activity detection under noisy environments," *Acoustical Science and Technology*, vol. 30, no. 5, pp. 363–371, 2009.

[21] S. Nakamura et al., "Aurora-2j: An evaluation framework for japanese noisy speech recognition," *IEICE transactions on information and systems*, vol. 88, no. 3, pp. 535–544, 2005.

[22] R.E. Fan et al., "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[23] X. Xiao, "`SignalGraph`: Date created: 16 jun 2016," `https://github.com/singaxiong/SignalGraph`, 2016.

[24] Q. Wang et al., "A universal vad based on jointly trained deep neural networks," in *Proc. Interspeech*, 2015, pp. 2282–2286.

[25] X. Zhang and J. Wu, "Denoising deep neural networks based voice activity detection," in *Proc. ICASSP*, IEEE, 2013, pp. 853–857.

[26] X. Zhang and D. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Proc. Interspeech*, 2014, pp. 1534–1538.