J Ambient Intell Human Comput (2017) 8:845–859
DOI 10.1007/s12652-017-0482-8

ORIGINAL RESEARCH

Noise robust voice activity detection using joint phase and magnitude based feature enhancement

Khomdet Phapatanaburi[1] · Longbiao Wang[2] · Zeyan Oo[1] · Weifeng Li[3] · Seiichi Nakagawa[4] · Masahiro Iwahashi[1]

Received: 31 May 2016 / Accepted: 30 March 2017 / Published online: 11 April 2017

Abstract Recently, deep neural network (DNN)-based feature enhancement has been proposed for many speech applications. DNN-enhanced features have achieved higher performance than raw features. However, phase information is discarded during most conventional DNN training. In this paper, we propose a DNN-based joint phase- and magnitude -based feature (JPMF) enhancement (JPMF with DNN) and a noise-aware training (NAT)-DNN-based JPMF enhancement (JPMF with NAT-DNN) for noise-robust voice activity detection (VAD). Moreover, to improve the performance of the proposed feature enhancement, a combination of the scores of the proposed phase- and magnitude-based features is also applied. Specifically, mel-frequency cepstral coefficients (MFCCs) and the mel-frequency delta phase (MFDP) are used as magnitude and phase features. The experimental results show that the proposed feature enhancement significantly outperforms the conventional magnitude-based feature enhancement. The proposed JPMF with NAT-DNN method achieves the best relative equal error rate (EER), compared with individual magnitude- and phase-based DNN speech enhancement. Moreover, the combined score of the enhanced MFCC and MFDP using JPMF with NAT-DNN further improves the VAD performance.

✉ Longbiao Wang
longbiao_wang@tju.edu.cn

[1] Nagaoka University of Technology, Nagaoka, Japan

[2] Tianjin Key Laboratory of Cognitive Computing and Application, School of Computer Science and Technology, Tianjin University, Tianjin, China

[3] Graduate School at Shenzhen, Tsinghua University, Shenzhen, China

[4] Toyohashi University of Technology, Toyohashi, Japan

Keywords Deep neural network (DNN) · Phase information · Noise-robust VAD · Feature enhancement

1 Introduction

Voice activity detection (VAD) is an active research topic in the field of speech processing because it is a key factor that influences the performance of practical speech applications (Tanrikulu 1997; Freeman and Cosier 1989; Malah et al. 1999; Enqing et al. 2002). VAD is used in problems that must identify speech and non-speech regions in a given speech signal. In high quality recording conditions, methods based on energy in (Junqua et al. 1991; Tucker 1992) perform well because of the energy difference between speech and non-speech segments. However, it is difficult to distinguish between speech and non-speech segments when the signal is corrupted by noise or a low signal-to-noise ratio (SNR). This is because human speech and non-speech signals have similar energy levels. Thus, energy-based VAD still remains challenging and requires a design with noise robustness.

Recently, machine-learning-based methods (Chang et al. 2006; Wu and Zhang 2011) have been proven to be effective at VAD. These methods have a strong theoretical basis that guarantees their performance under noisy conditions. To achieve state-of-the-art performance on VAD tasks, a VAD classifier and feature were investigated. There is a large amount of existing research on VAD classifiers. Kinnunen and Chernenko (2007) proposed a method to construct a support vector machine (SVM)-VAD that is based on a single frame classifier. In an evaluation on bus stop and lab noise, the SVM-VAD outperformed a short-term energy-based method, long-term spectral divergence method (Tong et al. 2006), and Gaussian mixture model

(GMM)-based VAD. This is because the SVM-VAD captures speech-relevant information effectively. Zhang and Wu (2013a) proposed a deep belief network-based VAD. They showed that this VAD outperformed traditional VAD (e.g., G.729B (Benyassine et al. 1997) and Sohn VAD (Ying et al. 2011)) because of its robustness to noise, due to its multi-frame-based classifier. In Zhang and Wang (2016), a DNN-based classifier achieved the state-of-the-art performance on a VAD task. Therefore, SVM and DNN-based VAD were considered in our experiment. In addition, feature extraction is important for machine-learning-based VAD. The mel-frequency cepstral coefficient (MFCC) (Davis and Mermelstein 1980) is a feature derived from the magnitude spectrum that has been proven to be an effective speech feature because it captures the most relevant information for speech. Zou et al. (2014) proposed an SVM-based VAD using MFCC features. Their results showed that MFCC provides high performance for VAD. Ryant et al. (2013) proposed DNN-based VAD using MFCC. Their result showed that a DNN classifier-based on MFCC outperformed GMM-based VAD. However, a conventional feature maybe not effective under noisy conditions because of the corruption of the speech.

To improve conventional features, feature enhancements have attracted attention in many speech processing tasks (Ueda et al. 2015; Wang et al. 2014; Xia and Bao 2013; Lu et al. 2013; Ren et al. 2016). This is because of the better classifier performance obtained using the enhanced features. Deep neural networks (DNNs), which have been improved by the introduction of restricted Boltzmann machine (RBM)-based pretraining (Hinton and Salakhutdinov 2006), have become popular for feature enhancement. Zhang and Wu (2013b) proposed a DNN-based feature enhancement (called a denoising autoencoder) that we here call DNN-based feature enhancement for noise-robust VAD. Compared with traditional feature-based VAD, the VAD using a DNN-based enhanced feature improves performance because the nonlinear mapping function can predict clean features from corrupted features, hence making the VAD decision better. (Ueda et al. 2015) also applied DNN-based feature enhancement to speaker identification. Compared with the traditional method, the enhanced feature provided better speaker identification accuracy. However, the DNN-based feature enhancement has a weakness when evaluated on previously unseen data.

While DNN-based feature enhancement has been applied to many speech processing tasks, noise-aware training (NAT) (Xu et al. 2015) was introduced to solve the problem of poor DNN-based feature enhancement performance on unseen test data in the speech enhancement task. In (Xu et al. 2014), NAT-DNN-based feature enhancement obtained significantly better performance than conventional DNN-based feature enhancement because it could better predict clean features from corrupted features owing to the addition of noise information during DNN training. Although NAT-DNN-based feature enhancement can provide better performance than DNN-based feature enhancement, we observe that phase information is discarded during feature enhancement training because of the inflexibility of phase computation. Therefore, NAT-DNN-based feature enhancement may be improved by phase-aware training.

In this paper, two conventional DNN-based feature enhancements (DNN- and NAT-DNN-based feature enhancement) are used as baselines for enhancing individual magnitude- and phase-based features for noise-robust VAD. Although both DNN and NAT-DNN-based feature enhancement may achieve good noise-robust VAD performance, they have a weakness in that phase information is discarded during feature enhancement training. To overcome this weakness, we propose adding phase-aware training into DNN- and NAT-DNN-based feature enhancement, which we call DNN-based joint phase and magnitude feature (JPMF) enhancement (JPMF with DNN), and NAT-DNN-based JPMF enhancement (JPMF with NAT-DNN). Moreover, we apply a combined score of the magnitude- and phase-based features to improve the VAD performance.

The remainder of this paper is organized as follows: Sect. 2 describes DNN-based feature enhancement and NAT-DNN-based feature enhancement. Section 3 introduces the proposed JPMF enhancement. Section 4 briefly describes the magnitude and phase-based feature extraction for DNN input. The score combination approach is described in Sect. 5. The VAD experiments using several feature enhancement approaches are evaluated in Section 5. Finally, Sect. 7 summarizes the paper and describes future work.

2 Conventional DNN-based feature enhancement

2.1 DNN-based feature enhancement

Neural networks (NNs) are universal mapping functions that can be used for both classification and regression problems (Xu et al. 2015). Many researches have used NNs for feature enhancement for a quite some time. An NN with more than one hidden layer is usually called a deep NN (DNN). Recently, DNN has been improved because of the introduction of a pretraining algorithm (Hinton et al. 2006) based on the RBM. This is why the deep structure of a DNN enables a much more efficient representation of many nonlinear transformations. In the past several years, DNNs have been utilized in many speech processing tasks such as acoustic modeling and feature enhancement. In this paper, we use DNN to enhance the MFCC magnitude feature.

Our aim is to utilize the flexibility of a DNN to model the highly nonlinear and complicated mapping from a distorted MFCC to the underlying clean MFCC. Note that we also apply DNN to map distorted phase features, the mel-frequency delta phase (MFDP), to clean them, and the basic concepts of both methods are the same. Hence, we use a unified description for both features here.

The structure of the conventional DNN-based feature enhancement is shown in Fig 1a. On the left of the DNN in the figure, a sequence of feature vectors is generated from a noisy feature. To enhance the features, we use the MFCC as a magnitude feature and MFDP as a phase feature. To predict the clean feature of current feature (shown as gray in the figure), a sequence of features around the current vector is fed into the DNN. This allows the DNN to use context information to predict the clean feature vector. After the nonlinear transformation of the hidden layers, a linear output layer is used to predict the clean feature vector for the current frame.

To train the DNN for feature enhancement, parallel data comprising the clean and corrupted features of the same speech signal are required. The clean and corrupted feature vector sequences must be aligned accurately at frame level. We use clean and multi-condition data for training the DNN. The objective of training is to minimize mean square error (MSE) function between the output feature and the target features (Xiao et al. 2014).

$$E_r = \frac{1}{N}\sum_{\eta=1}^{N} ||\hat{X}((Y_{\eta-\tau}^{\eta+\tau}), W, b) - X_\eta||_2^2. \quad (1)$$

Here, E_r is the MSE, $\hat{X}((Y_{\eta-\tau}^{\eta+\tau}), W, b)$ and X_η denote the estimated and reference normalized feature at sample index η, respectively, N represents the mini-batch size, $Y_{\eta-\tau}^{\eta+\tau}$ is the input feature, which is spliced at $\pm\tau$ context frames, W denotes the weight matrices, and b indicates the bias vector. The DNN parameters are then estimated iteratively

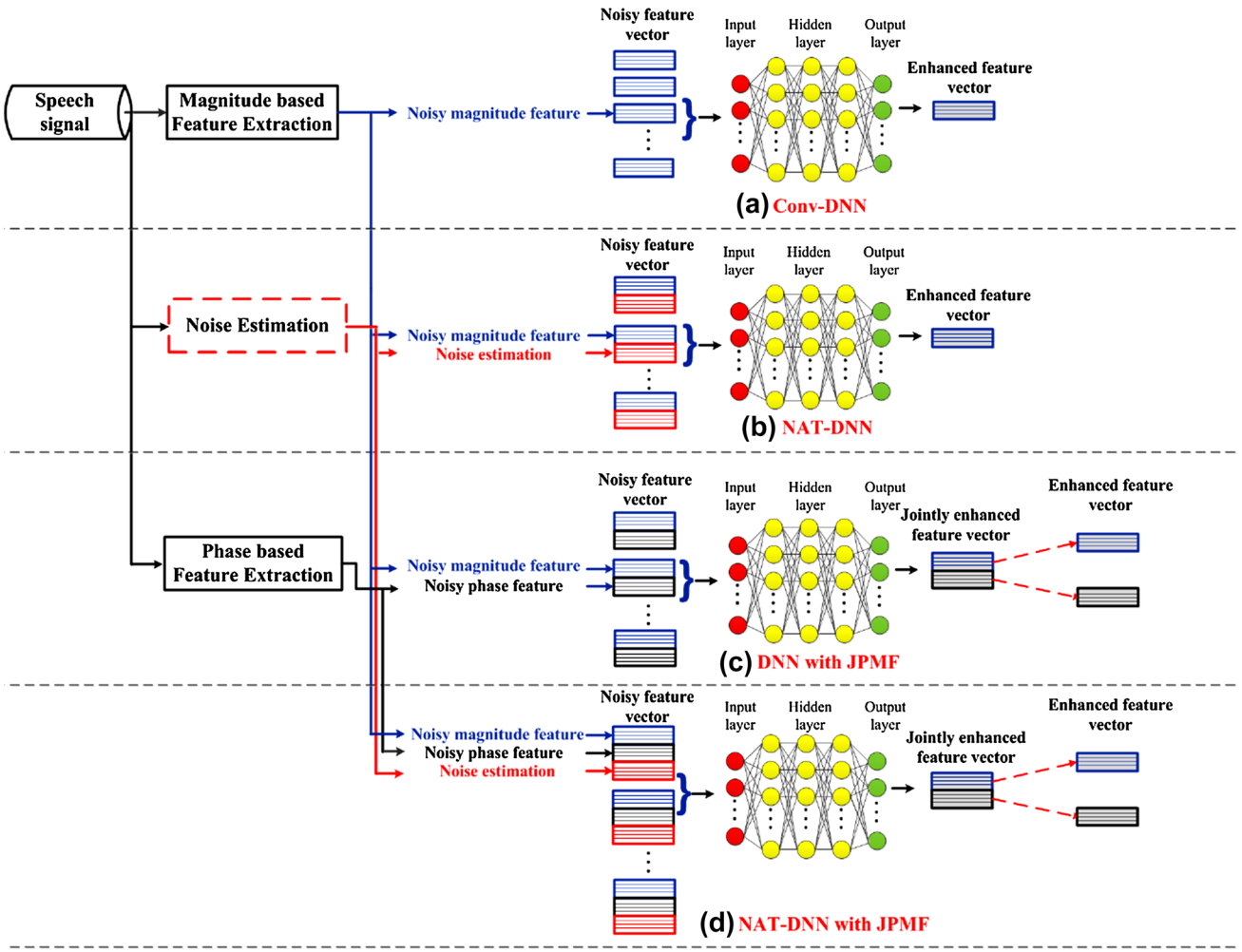

Fig. 1 Structure of each feature enhancement type: *a* conventional DNN-based feature enhancement, *b* NAT-DNN-based feature enhancement, *c* JPMF with DNN, and *d* JPMF with NAT-DNN

by stochastic gradient decent (Lu et al. 2013) using the updated equation below.

$$\Delta(W_{i+1}, b_{i+1}) = -\lambda \frac{\partial E_r}{\partial (W_i, b_i)} - \kappa \lambda (W_i, b_i) + \omega \Delta (W_i, b_i). \tag{2}$$

Here, i denotes the number of update iterations, λ indicate the learning rate, κ is the weight decay coefficient, and ω is the momentum coefficient. This supervised training step is often called fine-tuning. Before the MSE step, the DNN is initialized by a pretrained RBM, which uses unsupervised training, and hence only a corrupted version of the speech is required. When the DNN is trained, it is expected to handle corrupted features well.

2.2 NAT-DNN-based feature enhancement

In conventional DNN-based feature enhancement, the training is off-line because only a single magnitude feature is used for the regression function (Xu et al. 2015), as shown in Fig. 1a. Although the mapping function can effectively deal with a previously seen noisy condition, an evaluation of mismatched noise types might affect the generalization capabilities of a non-speech segment owing to unseen noisy speech whose distortion characteristics are not similar to those of the training data. To solve this problem, NAT-DNN is applied to enable noise awareness. In this process, DNN is fed with a feature augmented with its corresponding estimation of the noise using a conventional minimum MSE (MMSE)-based noise estimation (Hendriks 2010). Hence, the DNN can use additional on-line noise information to improve the prediction of the clean feature. The input vector of the DNN with the noise estimate is constructed as follows:

$$\hat{Y}_{\eta-\tau}^{\eta+\tau} = [A_{\eta-\tau}, \hat{Z}_{\eta-\tau}, \ldots, A_{\eta}, \hat{Z}_{\eta}, \ldots, A_{\eta+\tau}, \hat{Z}_{\eta+\tau}] \tag{3}$$

where A_η represents the magnitude-based feature vector of the current noisy speech frame η, where the window size is $2 * \tau + 1$, and $\hat{Z}_\eta$ is the noise estimation based on MMSE (Tong et al. 2006). In this procedure, the DNN is trained from the parallel data of the reference feature consisting of magnitude-based feature samples (MFCCs), like the traditional DNN and noise corrupted feature input vector $\hat{Y}_{\eta-\tau}^{\eta+\tau}$ of Eq. (3). The features are aligned at frame level. After training, the trained network can predict the underlying clean features when given noisy features, as shown in Fig. 1b.

3 Proposed JPMF enhancement

In the methods described in the previous section, phase information is discarded during most of the feature enhancement training because of the inflexibility of phase computation. In this section, we propose integrating phase-aware training into two conventional DNN feature enhancements. By applying phase-aware training, the DNN-based feature enhancement is augmented with phase information during training, which is the proposed JPMF with the DNN method. Similarly, NAT-DNN-based feature enhancement can also be augmented with phase information during training, and this is the proposed JPMF with NAT-DNN method. The additional phase-aware training is expected to achieve better performance when compared with the conventional DNN feature enhancements described in Sect. 2.

3.1 JPMF with DNN

The DNN-based feature enhancement in Sect. 2.1 is only trained with magnitude information, hence phase information is missing during training. A complex spectrum includes the magnitude spectrum and phase spectrum. In (Williamson et al. 2016a, b), the authors proposed a monaural speech separation in the complex domain. The experimental results show that complex traditional ratio masking outperforms ratio masking in the magnitude domain. That is, the results indicate that jointly enhancing the real and imaginary components can be better than enhancing the magnitude and phase independently. Joint enhancement of the magnitude and phase can improve speech quality, and is expected to enhance features well for VAD. To improve DNN-based feature enhancement, we propose JPMF with the DNN, which uses both the magnitude and phase information in one NN, which is expected to provide more accurate prediction than the DNN-based feature enhancement. In the training process, the input vector of the DNN is augmented using phase information as follows:

$$\bar{Y}_{\eta-\tau}^{\eta+\tau} = [P_{\eta-\tau}, A_{\eta-\tau}, \ldots, P_{\eta}, A_{\eta}, \ldots, P_{\eta+\tau}, A_{\eta+\tau}] \tag{4}$$

where P_η represents the phase-based feature vector of the current noisy speech frame η, where the window size is $2 * \tau + 1$. The DNN is trained on parallel data consisting of the reference feature with a clean JPMF and the input vector of the corrupted feature $\bar{Y}_{\eta-\tau}^{\eta+\tau}$. After training, the enhanced phase and magnitude features were derived separately from the jointly enhanced phase and magnitude information predicted by the trained network. Fig. 1c briefly shows the concept of JPMF enhancement.

3.2 JPMF with NAT-DNN

The NAT-DNN-based feature enhancement described in Sect. 2.2 was introduced to solve the problem of DNN-based feature enhancement when the testing and training data do not match (Xu et al. 2014; Seltzer et al. 2013). However, phase information is discarded during the

NAT-DNN-based feature enhancement training because of the complexity of the phase computation. Therefore, we introduce phase-aware training into traditional NAT-DNN-based feature enhancement. This is the proposed JPMF with the NAT-DNN method that uses magnitude information, phase information, and noise estimation in the DNN training. This method is expected to achieve more accurate prediction than the NAT-DNN-based feature enhancement. In the training process, the input vector of the NAT-DNN is augmented using phase information as follows:

$$\check{Y}_{\eta-\tau}^{\eta+\tau} = [P_{\eta-\tau}, A_{\eta-\tau}, \hat{Z}_{\eta-\tau}, \ldots, P_{\eta}, A_{\eta}, \hat{Z}_{\eta}, \ldots, P_{\eta+\tau}, A_{\eta+\tau}, \hat{Z}_{\eta+\tau}] \tag{5}$$

Figure 1d briefly shows the concept of the JPMF. The DNN is trained from the parallel data of the reference feature with the clean feature of Eq. 5, $\check{Y}_{\eta-\tau}^{\eta+\tau}$, and the input vector of the corrupted JPMF feature. After training, the enhanced phase and magnitude features are derived separately from the jointly enhanced features of the phase and magnitude information predicted by the trained network.

4 Magnitude and phase-based feature

In this work, we use two feature extraction methods to utilize both magnitude-based MFCC and phase-based MFDP.

4.1 MFCC

MFCC is a popular magnitude-based feature for speech processing, including VAD. We used MFCC as a magnitude-based feature for the DNN input.

4.2 MFDP

MFDP was proposed by McCowan and Dean (2011). It can be computed from the phase spectrum difference of the short-time discrete Fourier transform (STFT) between neighboring frames. The STFT of an input speech signal sequence is formulated as

$$X_m(k) = \sum_{n=-\infty}^{\infty} w(n - mD)x(n)e^{-jw_k n} \tag{6}$$

where m is the frame index, $w(n)$ is a causal window of length T (i.e., zero-valued outside the range $0 \le n \le T-1$), D is the number of samples between successive analysis frames (the step size, where $D \le T$), and $w_k = \frac{2\pi k}{L}$, where L is the number of analysis frequencies being considered in the discrete Fourier transform (with $L \ge T$). The short-time delta phase spectrum $\triangle\phi_m(k)$ is designed to avoid some of the phase unwrapping problems and is given as the time-derivative of the short-time phase spectrum as follows,

$$\triangle\phi_m(k) = arg[X_m(k)X_{m-1}(k)^* e^{-jw_k D}] \tag{7}$$

where $(.)^*$ indicates the complex conjugate. We then take advantage of the delta phase spectrum based on (McCowan and Dean 2011) to yield MFDP features by applying the mel filter bank to the absolute delta phase spectrum, followed by taking the logarithm of the filter bank energies and performing a discrete cosine transform to obtain the MFDP feature. Here, we use MFDP as phase information, using a rectangular window with a length of 25 ms instead of the common length of 256 ms because the longer length includes excess speech and non-speech segments in each window.

5 Score combination

In the previous section, both magnitude and phase feature are exploited in our method. Therefore, we propose a combination of the phase and magnitude feature scores to take advantage of the different benefits of these features. This technique is briefly described in this section. A flowchart of the VAD system is shown in Fig. 2. The SVM or DNN is used as a typical two-class classification for VAD. The decision about whether a given segment is speech or non-speech is based on the difference in probability that the segment is speech or non-speech.

$$\wedge(O) = p(O|\lambda_{speech}) - p(O|\lambda_{non\text{-}speech}), \tag{8}$$

where O is the feature vector of the input speech and λ_{speech} and $\lambda_{non\text{-}speech}$ are the models (SVM or DNN) of the speech and non-speech segments, respectively. Here, MFCC and MFDP are both used as feature vector. To combine the scores, the probabilities obtained from the different features are combined by the following equation.

$$p(O_{comp}|\lambda_j) = \alpha p(O_{MFCC}|\lambda_j) + (1-\alpha)p(O_{MFDP}|\lambda_j). \tag{9}$$

$$\alpha = \frac{p(O_{MFCC}|\lambda_j)}{p(O_{MFCC}|\lambda_j) + p(O_{MFDP}|\lambda_j)}. \tag{10}$$

where $p(O_{MFCC}|\lambda_j)$ and $p(O_{MFDP}|\lambda_j)$ are the probabilities based on MFCC and MFDP. Here, $j \in \{1, 2\}$ corresponds to class of speech or non-speech. In addition, α denotes the weighting coefficient.

6 Experiment

6.1 Experimental setup

Our experiments were conducted on the CENSREC-1-C database (Kitaoka et al. 2009). The speech data were

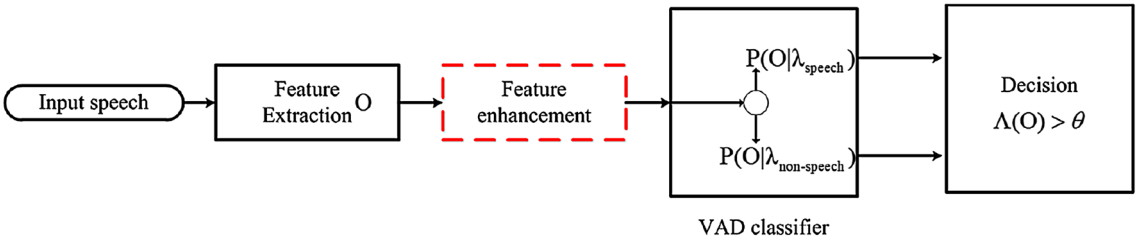

Fig. 2 Flowchart of the VAD system

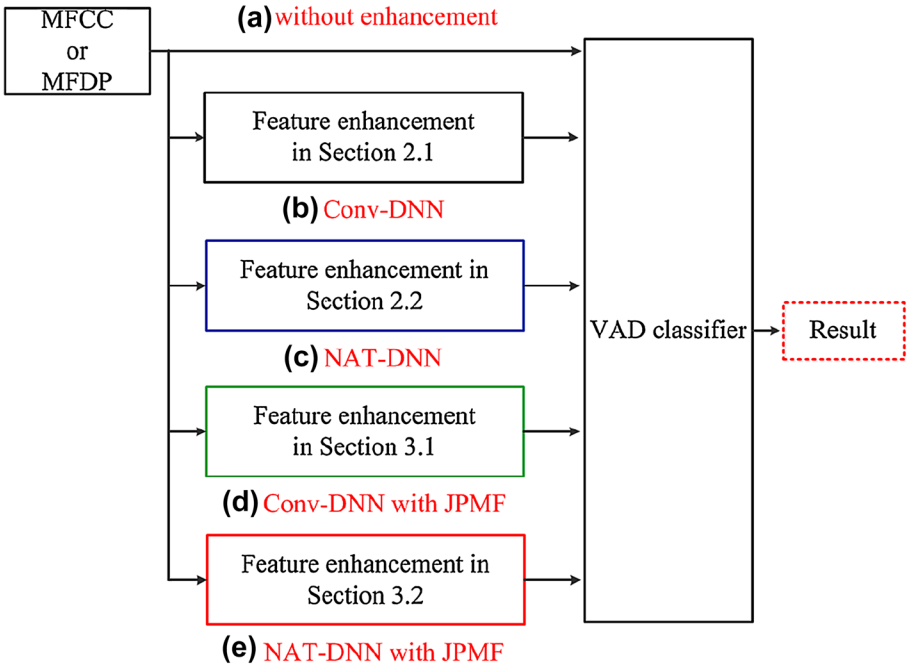

Fig. 3 Overview of the VAD system based on different feature enhancement approaches: *a* raw feature-based approach, *b* conventional DNN-based feature enhancement, *c* NAT-DNN-based feature enhancement, *d* JPMF with DNN, and *e* JPMF with NAT-DNN

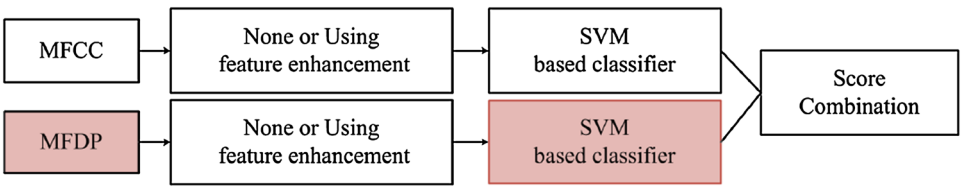

Fig. 4 Overview of the VAD system with score combination

sampled at 16 kHz and finally downsampled to 8 kHz. The details of recording condition, utterances, and speaking style are the same as in CENSREC-1(AURORA-2J). To create the noisy speech, the samples were corrupted by two noise sets A and B, as shown in Table 1. Each noise set includes four different noise environments with SNRs from −5 to 20 dB in increments of 5 dB.

Figure 3 shows an overview a VAD system using the different kinds of feature enhancements introduced in Sects. 2 and 3. The VAD system is fed with either MFCC or MFDP features. In Fig. 4, the MFCC and MFDP features are separately enhanced with the same type of enhancement method, as shown in Fig. 3 then their scores are obtained from the VAD classifier, which are then combines as described in Sect. 5. For the VAD detector, there are two classifier methods including SVM and DNN. Both classifier methods were trained using set A of the CENSREC-1-C

Table 1 Noise environment in the CENSREC-1-C database

Set	Additive noises
A	Subway, babble, car, exhibition
B	Restaurant, street, airport, station

database (artificially noisy speech segments with 19.39 h were derived from subway, babble, car, and exhibition noise) and tested with set B of the CENSREC-1-C database (artificially noisy speech segments with 19.39 h were derived from restaurant, street, airport, and station noise). For testing, the equal error rate (EER) is used as a measure of VAD performance. The result was evaluated using a broad range of SNR levels that were divided into three groups: high SNR, medium SNR, and low SNR. High SNR speech files were only corrupted by a small amount of noise and have SNR values of 20 and 15 dB. Medium SNR files have SNR values of 10 and 5 dB. To test the worst case scenario, the low SNR group was evaluated with SNR values of 0 and −5 dB.

To train the DNN, the multi-condition speech data of set A of the CENSREC-1-C database was used. Both MFCC and MFDP under the analysis conditions shown in Table 2 were tested. The input features consisted of seven spliced frames. A sigmoid type hidden layer was used for all layers except the input layer, where a linear hidden unit was used. To train the model for the feature enhancement approach we performed unsupervised RBM pretraining before supervised fine-tuning. To speed up the training, we performed RBM pretraining first. The Kaldi toolkit (Povey and Ghoshal 2011) was used for the pretraining task. The layers were trained in a layer-wise greedy fashion to maximize the likelihood over the training sample. The pretraining only requires a corrupted version of the utterance. For the back propagation to train the DNN, parallel data consisting of clean and distorted versions of the same utterance were used. The objective of this training is to minimize the MSE between the features. A stochastic gradient decent algorithm was used to improve the MSE error function. In the fine-tuning stage, the learning rate was 0.01, the weight decay coefficient was 0.5, and the momentum was 0.5.

6.2 Experimental result

To evaluate the proposed JPMF enhancements for noise-robust VAD, this section presents the results compared with two conventional DNN-based feature enhancements.

Table 2 Analysis conditions for MFCC and MFDP

	MFCC	MFDP
Frame length	25 ms	
Frame shift	10 ms	
FFT size	512 point	
Dimensions	39 (13 MFCCs, 13 △s, 13△△ s)	39 (13 MFDPs, 13 △s, and 13△△ s)

6.2.1 Results for the SVM classifier

In this subsection, SVM was used as a VAD to consider the performance of a single frame-based classifier which was based on (Kinnunen and Chernenko 2007). To rapidly optimize the support vectors, we used the publicly available LIBLINEAR tool (Fan et al. 2008), which considers linear kernels for our experiment.

First, the results of VAD using unenhanced features (raw feature-based VAD) shown in Fig. 3a are first compared. The EERs are shown in Table 3 for MFCC and Table 4 for MFDP (rows highlighted by gray). Next, the VAD using different DNN of feature enhancement configurations (shown in Fig. 3b–e) were computed to achieve better performance. The DNNs of the enhanced features are trained using the parameters described in the subsection. The number of layers was one or three, and the number of nodes in each hidden layer was varied from 512 to 2048. The results of each layer and node were specified using 30 fine-tuning iterations. The EER results of the DNN for each feature enhancement method are shown in Fig. 5 and show performance for MFCC (Fig. 5a), and MFDP (Fig. 5b). The results show that feature enhancement using a DNN with three layers did not perform well according to our expectations when unseen noise was used in the VAD evaluation. This might be because training data are not sufficient. However, when the data were limited to seen noise, feature-enhanced VAD with all DNN configurations achieved better performance for both MFCC and MFDP than the VAD based on raw features (Tables 3, 4). We selected the configuration with the best results for later experiments from Fig. 5.

Table 3 shows the equal error rates (EERs) of SVM classifier using a magnitude-based feature (MFCC). By applying the feature enhancement with noise-robust VAD, the EERs were improved from the raw MFCC-based VAD in (Zou et al. 2014). Similarly, the EERs were also improved from the raw MFCC-based VAD by applying the NAT-DNN-based feature enhancement. This is because of the enhanced MFCC using DNN. Moreover, the results show that the proposed feature enhancement, which uses both magnitude and phase information in one NN, provides better performance than conventional DNN-based feature enhancements. Using JPMF with DNN, the EERs were better than those of the DNN-based feature enhancement. In same way, the EERs of JPMF with the NAT-DNN were better than those of the NAT-DNN-based feature enhancement. This is because these methods better predict features, as the training input contains phase information to make DNN more efficient.

Table 4 shows the EERs of SVM classifier using the phase-based feature (MFDP). The abbreviations of the methods are the same as those in Table 3.

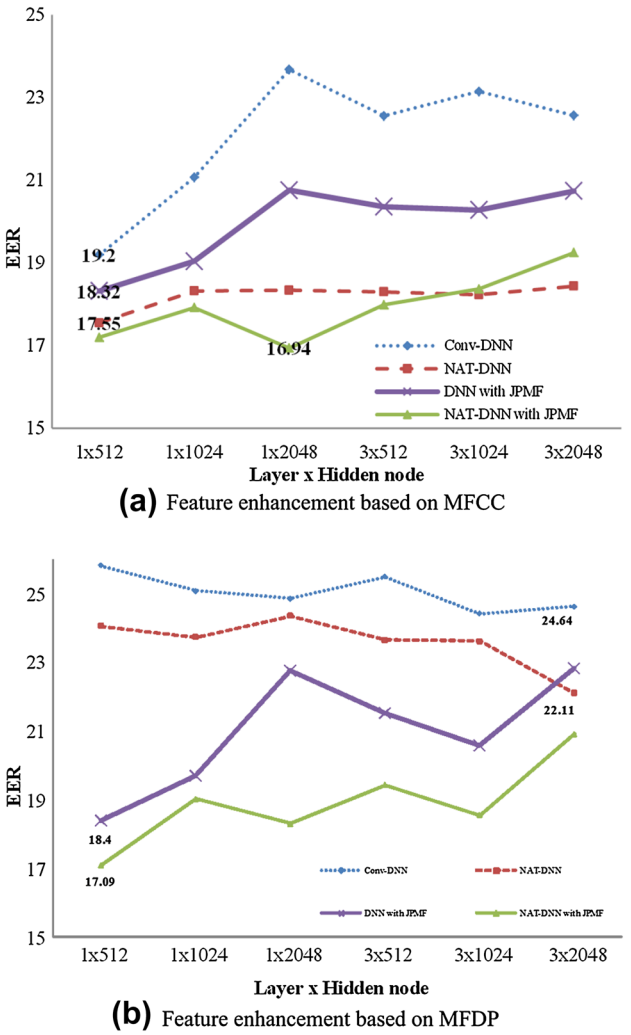

Fig. 5 EERs of each enhancement system shown in Fig. 3b, e: **a** using MFCC and **b** using MFDP

By applying the DNN-based feature enhancement and the NAT-DNN-based feature enhancement, the EERs were improved compared with those of the raw MFDP. Therefore, we can confirm that the DNN enhancement was also effective for phase features. Moreover, applying JPMF with DNN and JPMF with NAT-DNN have the same tendency as for the magnitude feature. This is because the DNN can use both magnitude and phase information for the enhancement, and hence more accurate clean features can be estimated.

6.2.2 Result of DNN classifier

In this section, the DNN was used as VAD detector to consider the effect of a multi-frame-based classifier. SignalGraph (Xiao 2016) was used to train the DNN. The DNN has one layer containing 512 neurons. The input

Table 3 EER (%) of the SVM classifier based on MFCC: None is the result of the system in Fig. 3a, Conv-DNN is the result of the system in Fig. 3b, NAT-DNN is result of the system in Fig. 3c, DNN with JPMF is the result of the system in Fig. 3d, and NAT-DNN with JPMF is the result of the system in Fig. 3e

Enhanced method	Restaurant			Street			Airport			Station			Avg
	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	All SNR
None	16.47	22.86	35.03	13.31	21.25	33.72	16.34	22.86	34.23	11.26	19.73	32.71	23.31
Conv-DNN	8.55	18.59	34.41	9.41	17.67	31.67	10.00	17.39	32.49	5.75	13.56	30.93	19.20
NAT-DNN	**5.19**	15.79	33.01	**5.48**	18.21	33.34	**7.39**	16.16	31.20	3.32	10.66	30.80	17.55
DNN with JPMF	7.21	18.13	34.20	7.41	**16.00**	**30.78**	9.53	17.23	32.20	4.39	12.46	30.31	18.32
NAT-DNN with JPMF	5.62	**15.26**	**31.95**	5.64	17.29	33.32	7.78	**14.78**	**29.64**	3.25	**9.48**	**29.31**	**16.94**

Avg is short for average
Bold value corresponding to best result under each conditions

Table 4 EER (%) of the SVM classifier based on MFDP

Enhanced method	Restaurant			Street			Airport			Station			Avg
	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	All SNR
None	22.61	35.22	44.22	16.3	25.25	39.53	19.74	29.89	42.34	16.95	27.40	40.60	30.00
Conv-DNN	14.49	23.54	38.44	12.64	21.77	36.82	14.22	22.05	36.13	15.24	23.16	34.54	24.42
NAT-DNN	12.95	21.56	35.91	11.28	21.58	35.46	11.94	19.05	31.89	12.81	19.05	31.89	22.11
DNN with JPMF	**6.20**	14.47	31.55	10.58	**19.05**	**32.64**	8.07	14.87	30.3	6.57	14.12	32.34	18.4
NAT-DNN with JPMF	5.33	**13.47**	**31.62**	**7.49**	19.79	32.86	**6.12**	**13.61**	**29.06**	**4.25**	**10.97**	**30.49**	**17.09**

Avg is short for average

Bold value corresponding to best result under each conditions

feature for the DNN contains nine frames, and cross entropy was used. The learning rate started from 0.1 and was changed to 1 for the second epoch. It then decayed by a factor of 0.5 each time the cross entropy on a cross validation set between two consecutive epochs increased. The features used were the same as those used by the SVM classifier.

Before using feature enhancements, we investigated the DNN configurations with 1, 2 and 3 layers, each using raw MFCC as the input (Ryant et al. 2013). The results are shown in Table 5. DNN-VAD-based on more hidden layers did not perform according to our expectations. It might not be effective to do this if we simply consider VAD as a binary-class classification problem with the noisy speech and the background noise as the two classes. Therefore, we selected the DNN-VAD-based on one hidden layer with feature enhancements.

Table 6 shows the EERs of the DNN classifier using MFCC. When applying the most feature enhancement, the EER results did not perform to our expectations. This is because multi-frame-based classification requires significantly enhanced features. However, JPMF with NAT-DNN could provide better performance than the other features because of its significantly enhanced feature.

Table 7 shows the EERs of the DNN classifier using MFDP. When feature enhancement is applied, the results have the same tendencies as the SVM classifier. This is because the DNNs significantly contribute to the performance of the phase-based feature.

6.2.3 Result of score combination

This section reports the results of combining the MFCC and MFDP scores. Table 8 shows the results of score combination based on the SVM, and Table 9 shows the results of score combination based on the DNN. We can see that the VAD based on the combined score outperformed the systems using individual feature. This is because it takes advantage of the combination of different decisions of the systems.

6.2.4 Analytic illustration of noise suppression

To better visualize the enhanced magnitude and phase-based feature described in the previous sections, this section displays the spectrogram of the enhanced MFCC and MFDP features and their scores.

Figure 6 shows MFCC feature spectrograms of an utterance example corrupted by stationary noise at SNR = 0. The spectrograms of noisy MFCC, clean MFCC, and MFCCs enhanced by the conventional DNN, NAT-DNN,

Table 5 EER (%) of DNN classifiers with different numbers of layers

# Number layer	Restaurant			Street			Airport			Station			Avg
	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	All SNR
1	3.57	11.14	32.50	3.40	9.67	27.35	5.54	11.48	26.60	2.88	8.21	27.75	**14.17**
2	3.78	11.30	32.58	3.44	9.75	28.89	5.28	11.11	26.13	2.88	8.21	27.66	14.25
3	3.94	11.33	32.63	3.39	9.84	28.95	5.41	11.35	26.42	2.88	8.02	27.82	14.33

Bold value corresponding to best result under each conditions

Table 6 EER (%) of the DNN classifier based on MFCC

Enhanced method	Restaurant			Street			Airport			Station			Ave
	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	All SNR
None	3.57	11.14	32.5	3.4	**9.67**	27.35	5.54	11.48	**26.6**	2.88	8.21	**27.75**	14.17
Conv-DNN	3.95	11.79	31.77	4.03	8.98	24.87	5.88	12.37	28.27	3.57	9.43	28.82	14.48
NAT-DNN	**2.69**	8.92	30.64	2.84	11.61	28.76	**4.08**	11.96	28.46	**2.49**	8.39	30.5	14.28
DNN with JPMF	3.51	10.75	30.63	3.19	8.99	**24.86**	5.75	12.49	29.11	3.24	9.38	28.69	14.21
NAT-DNN with JPMF	2.78	**8.91**	**28.52**	**2.71**	11.25	28.37	4.49	**11.24**	26.93	2.51	**7.62**	28.59	**13.66**

Ave is short for average

Bold value corresponding to best result under each conditions

Table 7 EER (%) of the DNN classifier based on MFDP: Avg is short for average

Enhanced method	Restaurant			Street			Airport			Station			Avg
	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	All SNR
None	14.20	23.9	39.08	12.17	22.11	37.69	12.90	21.58	36.28	13.92	22.49	35.24	24.29
Conv-DNN	14.11	23.51	38.34	11.45	20.7	35.89	13.30	21.28	35.49	13.95	21.60	34.49	23.67
NAT-DNN	11.96	20.44	36.78	10.25	21.55	35.22	11.44	18.67	33.89	12.81	20.07	37.18	22.52
DNN with JPMF	5.37	14.22	32.61	7.37	**16.10**	**30.64**	7.88	14.86	30.38	5.38	13.55	31.82	17.51
NAT-DNN with JPMF	**3.58**	**10.85**	**26.42**	**4.25**	16.73	31.25	**5.18**	**12.69**	**29.55**	**3.28**	**9.33**	**31.79**	**15.41**

Bold value corresponding to best result under each conditions

Table 8 EER (%) of SVM classifier based score combination of MFCC and MFDP

Enhanced method	Restaurant			Street			Airport			Station			**Avg**
	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	All SNR
None	12.14	19.99	33.33	8.68	15.62	32.12	11.86	18.51	31.15	7.53	13.50	30.22	19.55
Conv-DNN	7.06	15.82	33.06	5.69	15.24	30.75	8.57	15.34	30.52	5.08	12.32	30.31	17.48
NAT-DNN	4.53	14.59	33.09	**4.65**	14.06	33.28	7.00	14.93	30.78	3.29	10.49	30.84	16.79
DNN with JPMF	5.9	14.23	31.59	7.19	**14.04**	**30.51**	8.04	14.73	30.11	4.28	12.1	30.41	16.93
NAT-DNN with JPMF	**4.41**	**13.01**	**31.57**	6.11	14.58	30.98	**5.97**	**13.47**	**29.17**	**2.95**	**9.16**	**29.82**	**15.93**

Bold value corresponding to best result under each conditions

Table 9 EER (%) of DNN classifier based score combination of MFCC and MFDP

Enhanced method	Restaurant			Street			Airport			Station			Avg
	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	All SNR
None	3.43	9.93	31.1	3.11	8.55	25.97	5.51	11.31	**25.77**	2.85	8.07	**27.68**	13.61
Conv-DNN	3.76	10.81	30.86	3.99	8.13	26.50	5.91	12.28	27.12	3.47	9.17	28.16	14.18
NAT-DNN	**2.62**	8.92	29.49	2.84	11.61	28.76	3.97	11.84	27.86	2.45	7.80	30.00	14.01
DNN with JPMF	3.42	10.44	30.13	3.28	**8.07**	**24.04**	5.75	12.29	29.00	3.18	9.18	28.66	13.95
NAT-DNN with JPMF	2.77	**8.87**	**28.41**	**2.7**	11.25	26.95	**4.52**	**11.2**	26.81	**2.43**	**7.44**	28.42	**13.48**

Bold value corresponding to best result under each conditions

JPMF with DNN (the first proposed method), and JPMF with NAT-DNN (the second proposed method) are shown in Fig. 6b–g. Comparing Fig. 6b with Figs. 6d–f, the spectrograms using feature enhancement provide better speech/non-speech segment boundaries than those of the raw feature. This is because of DNN enhancement. To observe how the proposed method works, the spectrogram of Fig. 6d can be compared with that of Fig. 6f. It is clear that the enhanced MFCC using JPMF with the DNN provides a better boundary between the speech segment than that using the conventional DNN. The same tendency can be found for NAT-DNN. This is because of the phase-aware training. Hence, we confirmed that introducing JPMF in both conventional DNN and NAT-DNN training improves noisy MFCC features.

Figure 7 shows the MFDP spectrograms of an utterance example corrupted by stationary noise at SNR = 0. The spectrograms of noisy MFDP, clean MFDP, and MFDP enhanced by the conventional DNN, NAT-DNN, JPMF with the DNN (the first proposed method), and JPMF with NAT-DNN (the second proposed method) are shown in Figs. 7b–g. Comparing Fig. 7b with Fig. 7d–f, spectrograms using feature enhancement provide better speech/non-speech segment boundaries than those using the raw feature. Again, this is because of DNN enhancement. To observe the proposed method, the spectrograms of Fig. 7d with Fig. 7f can be compared. Clearly, the enhanced MFCC using JPMF with DNN provide better boundaries between speech segments than those using conventional DNN and the same tendency can be found for NAT-DNN. This is due to introducing the magnitude information during DNN training.

7 Conclusions and future work

In this paper, we proposed a DNN-based JPMF enhancement called JPMF with DNN and a NAT-DNN-based JPMF enhancement called JPMF with NAT-DNN for noise-robust VAD. Moreover, to improve performance of feature enhancement, a combination of the scores of the phase- and magnitude-based features was also applied. MFCCs and MFDP were used as magnitude and phase features. The experimental results show that the proposed feature enhancement significantly outperforms the conventional magnitude-based feature enhancements (DNN and NAT-DNN-based feature enhancements) under both SVM and DNN-based VAD. Furthermore, a combined score of the enhanced MFCC and MFDP features (using JPMF with NAT-DNN) further improved the VAD performance.

In the future, we plan to extend the generation capabilities of the DNN to enable it to handle non-stationary noise conditions. We will try to use other efficient noise estimator

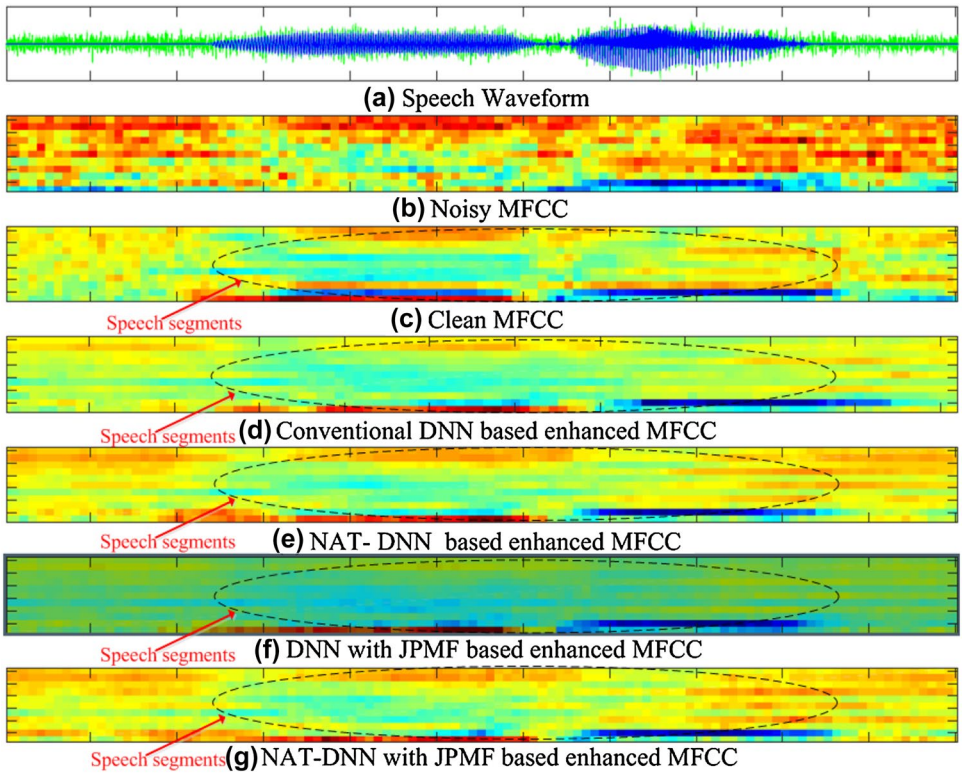

Fig. 6 Spectrograms and VAD score of each enhancement method based on the MFCC feature: **a** the speech waveform, where the *blue line* is clean speech and the *green line* is stationary noise at SNR = 0, **b** noisy MFCC, **c** clean MFCC, **d** DNN-based enhanced MFCC, **e** NAT-DNN-based enhanced MFCC, **f** enhanced MFCC using JPMF with DNN, and **g** enhanced MFCC using JPMF with NAT-DNN

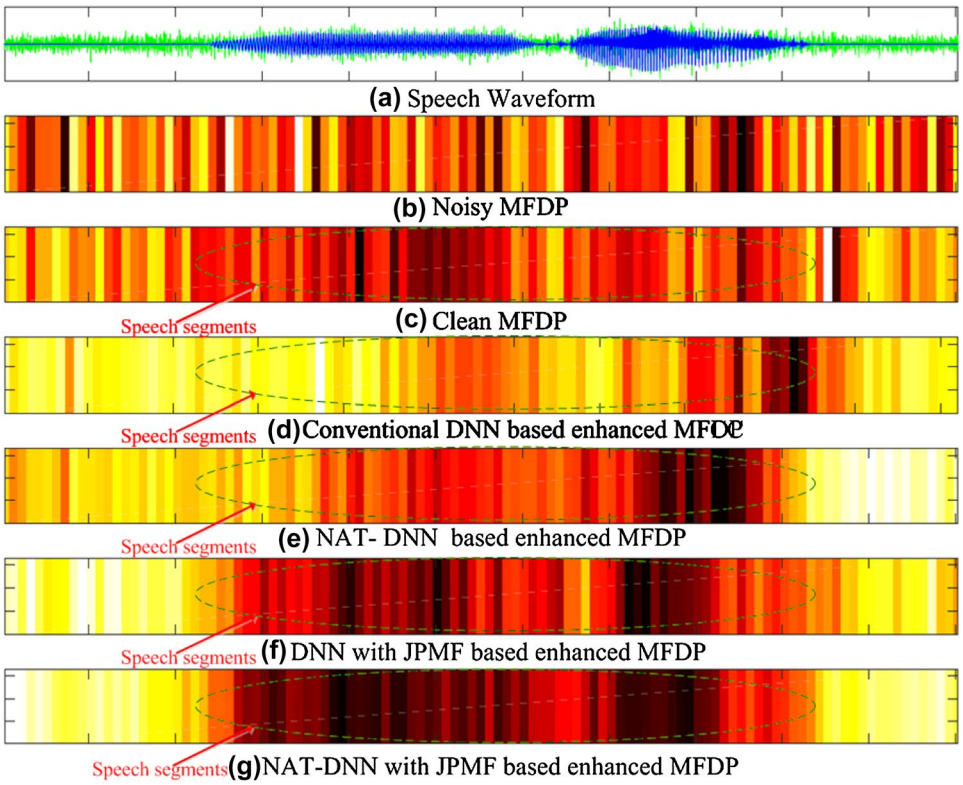

Fig. 7 Spectrograms and VAD scores of each enhancement method based on the MFDP feature: **a** the speech waveform, where the *blue line* is clean speech and the *green line* is stationary noise at SNR = 0, **b** noisy MFDP, **c** clean MFDP, **d** DNN-based enhanced MFDP, **e** NAT-DNN-based enhanced MFDP, **f** enhanced MFDP using JPMF with DNN, and **g** enhanced MFDP using JPMF with NAT-DNN

to estimate non-stationary noise. Moreover, we will investigate and combine other magnitude-based and phase-based spectral features such as linear prediction cepstral coefficients, power-normalized cepstral coefficients, and relative phase information (Kim and Stern 2012; Nakagawa et al. 2012; Wang et al. 2010; Wang et al 2015).

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

References

Benyassine A, Shlomot E, Su H-Y, Massaloux D, Lamblin C, Petit J-P (1997) Itu-t recommendation g. 729 annex b: a silence compression scheme for use with g. 729 optimized for v. 70 digital simultaneous voice and data applications. IEEE Commun Mag 35(9):64–73

Chang J-H, Kim NS, Mitra SK (2006) Voice activity detection based on multiple statistical models. IEEE Trans Signal Process 54(6):1965–1976

Davis SB, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 28:357–366

Enqing D, Heming Z, YongLi L (2002) Low bit and variable rate speech coding using local cosine transform. In: TENCON'02. Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, pp 423–426

Fan RE, Chang KW, Hsieh CJ (2008) LIBLINEAR: a library for large linear classification. J Mach Learn Res 9:1871–1874

Freeman D, Cosier G (1989) The voice activity detector for the Pan-European digital cellular mobile telephone service. In: 1989 international conference on acoustics, speech, and signal processing, 1989. ICASSP-89. pp 369–372

Hendriks RC (2010) MMSE based noise PSD tracking with low complexity. 2010 IEEE international conference on acoustics speech and signal processing (ICASSP), pp 4266–4269

Hinton G, Osindero S, Teh Y (2006) A fast learning algorithm for deep belief nets. Neural Comput 18:1527–1554

Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313:504–507

Junqua J, Reaves B, Mak B (1991) A study of endpoint detection algorithms in adverse conditions: incidence on a DTW and HMM recognizer. In: Second European conference on speech communication and technology

Kim C, Stern RM (2012) Power-normalized cepstral coefficients (PNCC) for robust speech recognition. In: 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 4101–4104

Kinnunen T, Chernenko E (2007) Voice activity detection using MFCC features and support vector machine. In: Conf. on Speech and Computer (SPECOM07), Moscow, Russia, pp 556–561

Kitaoka N, Yamada T, Tsuge S (2009) CENSREC-1-C: an evaluation framework for voice activity detection under noisy environments. Acoust Sci Technol 30:363–371

Lu X, Tsao Y, Matsuda S, Hori C (2013) Speech enhancement based on deep denoising autoencoder. In: INTERSPEECH, pp 436–440

Malah D, Cox RV, Accardi AJ (1999) Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments. In: 1999 IEEE international conference on acoustics, speech, and signal processing, 1999. Proceedings, pp 789–792

McCowan I, Dean D (2011) The delta-phase spectrum with application to voice activity detection and speaker recognition. IEEE Trans Audio Speech Lang Process 19:2026–2038

Nakagawa S, Wang L, Ohtsuka S (2012) Speaker identification and verification by combining MFCC and phase information. IEEE Trans Audio Speech Lang Process 20:1085–1095

Povey D, Ghoshal A (2011) The Kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding (No. EPFL-CONF-192584). IEEE Signal Processing Society

Ren B, Wang L, Lu L, Ueda Y, Kai A (2016) Combination of bottleneck feature extraction and dereverberation for distant-talking speech recognition. Multimed Tools Appl 75(9):5093–5108

Ryant N, Liberman M, Yuan J (2013) Speech activity detection on youtube using deep neural networks. In: INTERSPEECH, pp 728–731

Seltzer ML, Yu D, Wang Y (2013) An investigation of deep neural networks for noise robust speech recognition. In: 2013 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 7398–7402

Tanrikulu O (1997) Residual echo signal in critically sampled subband acoustic echo cancellers based on IIR and FIR filter banks. IEEE Trans Signal Process 45:901–912

Tong R, Ma B, Lee KA, You C (2006) The IIR NIST 2006 Speaker Recognition System: Fusion of Acoustic and Tokenization Features. In: Presentation in 5th Int. Symp. on Chinese Spoken Language Processing, ISCSLP

Tucker R (1992) Voice activity detection using a periodicity measure. IEE Proc Commu Speech Vis I:377–380

Ueda Y, Wang L, Kai A, Ren B (2015) Environment-dependent denoising autoencoder for distant-talking speech recognition. EURASIP J Adv Signal Process 2015(92):1–11

Wang L, Minami K, Yamamoto K, Nakagawa S (2010) Speaker identification by combining MFCC and phase information in noisy environments. In: 2010 IEEE international conference on acoustics speech and signal processing (ICASSP), pp 4502–4505

Wang L, Ren B, Ueda Y (2014) Denoising autoencoder and environment adaptation for distant-talking speech recognition with asynchronous speech recording. In: asia-pacific signal and information processing association, 2014 annual summit and conference (APSIPA), pp 1–5

Wang L, Yoshida Y, Kawakami Y, Nakagawa S (2015) Relative phase information for detecting human speech and spoofed speech. In: INTERSPEECH, pp 2092–2096

Williamson DS, Wang Y, Wang D (2016a) Complex ratio masking for joint enhancement of magnitude and phase. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 5220–5224

Williamson DS, Wang Y, Wang D (2016b) Complex ratio masking for monaural speech separation. IEEE/ACM Trans Audio Speech Lang Process 24(3):483–492

Wu J, Zhang X (2011) Efficient multiple kernel support vector machine based voice activity detection. IEEE Signal Process Lett 18:466–469

Xia B, Bao C (2013) Speech enhancement with weighted denoising auto-encoder. In: INTERSPEECH, pp 3444–3448

Xiao, X. (2016). SignalGraph. https://github.com/singaxiong/SignalGraph

Xiao X, Zhao S, Nguyen DHH (2014) The NTU-ADSC systems for reverberation challenge 2014. In: Proc, REVERB challenge workshop

Xu Y, Du J, Dai L, Lee C (2015) A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans Audio Speech Lang Process 23:7–19

Xu Y, Du J, Dai LR, Lee CH (2014) Dynamic noise aware training for speech enhancement based on deep neural networks. In: INTERSPEECH, pp 2670–2674

Ying D, Yan Y, Dang J, Soong FK (2011) Voice activity detection based on an unsupervised learning framework. IEEE Trans Audio Speech Lang Process 19(8):2624–2633

Zhang X-L, Wang D (2016) Boosting contextual information for deep neural network based voice activity detection. IEEE/ACM Trans Audio Speech Lang Process 24(2):252–264

Zhang XL, Wu J (2013a) Deep belief networks based voice activity detection. IEEE Trans Audio Speech Lang Process 21:697–710

Zhang XL, Wu J (2013b) Denoising deep neural networks based voice activity detection. In: 2013 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 853–857

Zou YX, Zheng WQ, Shi W, Liu H (2014) Improved voice activity detection based on support vector machine with high separable speech feature vectors. In: 2014 19th international conference on digital signal processing (DSP), pp 763–767