DNN-BASED VOICE ACTIVITY DETECTION USING AUXILIARY SPEECH MODELS IN NOISY ENVIRONMENTS

Yuuki Tachioka

Denso IT Laboratory
2–15–1 Shibuya, Shibuya-ku, Tokyo, Japan

ABSTRACT

Voice activity detection (VAD) is essential for automatic speech recognition (ASR) in noisy environments. Deep neural network (DNN)-based VAD is more powerful than previous types. In the fields of ASR and speech enhancement, to improve the performance of DNNs, in addition to spectral features, auxiliary features are used because these features are effective for adapting DNNs to a target environment. To improve the performance of DNN-based VAD further, this paper proposes two types of auxiliary feature based on auxiliary speech models. The first is activation of non-negative matrix factorization and the second is acoustic score of ASR acoustic models. These features give auxiliary information to DNNs in the same way as ASR and speech enhancement do. Experimental results for noisy VAD tasks demonstrated that DNN-based methods outperformed one of the most effective conventional methods and that both auxiliary features improved performance, with the second feature being better than the first one.

Index Terms— DNN-based voice activity detection, non-negative matrix factorization, speech recognition, auxiliary features

1. INTRODUCTION

Speech interfaces, such as voice control with distant microphones or automatic speech recognition (ASR) without push-to-talk switches, have become popular. The increase in opportunity of using such interfaces in real environments renders voice activity detection (VAD), especially in highly noisy environments, more important. Conventionally, power-based methods [1] were widely used under the assumption that speech power is greater than noise power. To improve VAD performance in highly noisy environments that cannot assume the above, the likelihood ratio test [2, 3] has become more mainstream. This models spectral characteristics at each frequency bin as Gaussian distributions. Model parameters are calculated from the observed noise without any prior training. Compared with power-based methods that use one-dimensional features, detailed models using multi-dimensional features are robust to noise and can capture spectral patterns, such as harmonic speech structures.

On the other hand, prior training of models improves performance. [4] shows the effectiveness of the use of clean speech models. At the testing time, noisy speech models are composed of a trained clean speech model and a noise model, which is constructed from the observed noise.

After [5] had shown the effectiveness of deep neural networks (DNNs) on ASR tasks, [6, 7] showed its effectiveness on VAD tasks. DNNs use spectrum-derived multi-dimensional features and their model parameters are trained on the noisy speech training data. DNN-based methods have two advantages against conventional methods: flexible models that represent various speech and noise patterns and dimensional compression with non-linear functions although conventional models that can deal with multi-dimensional features use dimensional compression with linear models.

In the case of ASR model adaptation [8, 9], auxiliary features that represent speaker or environmental characteristics, such as i-vectors [10], improve ASR performance. DNN-based speech synthesis also uses auxiliary features, such as speaker codes [11], to change speaker characteristics. These studies show that auxiliary features can adapt DNNs to a target condition.

In addition, in the case of DNN-based speech enhancement (SE), auxiliary features, such as phoneme information obtained from ASR results, can improve SE performance [12, 13]. Phoneme information is helpful because the extent of SE can be changed depending on the phoneme properties. That is, SE carefully deals with phonemes that tend to be mixed with noise and it quickly deals with phonemes that are hardly mixed with noise.

VAD is a problem of discrimination between speech and noise, but solving this problem directly is hard because speech has large varieties. Auxiliary features that limit speech patterns reduce speech diversity. From a different perspective, DNNs adapt to a pre-estimated phoneme at each frame. As well as SE, VAD carefully deals with phonemes that tend to be mixed with noise, such as consonants especially fricatives, and quickly deals with phonemes that are less mixed with noise, such as vowels.

This paper discusses the improvement in the performance of DNN-based VAD by using auxiliary features output from two types of auxiliary speech model: non-negative factorization (NMF) models [14, 15] and ASR acoustic models. An auxiliary feature from the former model is NMF activation and one from the latter model is acoustic score of acoustic models.

Experiments on the data of in-car environments [16] show the effectiveness of our proposed method. There are three objectives for this experiment. The first one is to clarify how DNN-based VAD is better than conventional methods because there are few comparisons of them. The second one, which is the main objective, is to validate the effectiveness of the auxiliary features; and the third one is to show the improvement of ASR due to the improvement of VAD.

Section 2 overviews a likelihood ratio test method, which was one of the state-of-the-art methods before DNN-based VAD was proposed. Section 3 describes conventional DNN-based VAD, which is the baseline of our proposed method. Section 4 proposes two types of auxiliary feature. Section 5 describes a VAD experiment under noisy environments.

2. LIKELIHOOD RATIO TEST

This section overviews an effective conventional method that has been widely used. Sohn's method models spectral characteris-

tics across frequency bins f to detect speech [2]. Short-time Fourier transform (STFT) coefficients of the observed sounds at frame t, $\boldsymbol{X} \in \mathbb{C}^{F\times T}$, are used as an input feature vector $\boldsymbol{X}_t = \{X_{f=1,\dots,F}\} \in \mathbb{C}^F$. When the speech and noise STFT coefficients are $\boldsymbol{S}_t = \{S_{f=1,\dots,F}\}$ and $\boldsymbol{N}_t = \{N_{f=1,\dots,F}\}$, respectively, the observed features are $\boldsymbol{X}_t = \boldsymbol{N}_t$ in non-speech frame L_N and $\boldsymbol{X}_t = \boldsymbol{N}_t + \boldsymbol{S}_t$ in speech frame L_S. Here, in L_N and L_S, the probability density functions of the respective $\boldsymbol{X}_t$ are assumed to be independent Gaussian distributions across frequency bins defined as

$$\begin{aligned} p(\boldsymbol{X}_t|L_N) &= \prod_{f=1}^{F} \frac{1}{\pi\lambda_f^N} e^{-\frac{|X_f|^2}{\lambda_f^N}}, \\ p(\boldsymbol{X}_t|L_S) &= \prod_{f=1}^{F} \frac{1}{\pi[\lambda_f^N + \lambda_f^S]} e^{-\frac{|X_f|^2}{[\lambda_f^N + \lambda_f^S]}}, \end{aligned} \tag{1}$$

where λ_f^N and λ_f^S are the variances of N_f and S_f, respectively. At the fth dimension, the likelihood ratio of speech and non-speech, Λ_f, is described as

$$\Lambda_f(X_f) = \frac{p(X_f|L_S)}{p(X_f|L_N)} = \frac{1}{1+\xi_f} e^{\frac{\gamma_f \xi_f}{1+\xi_f}}, \tag{2}$$

where ξ_f and γ_f are the prior and posterior signal-to-noise ratios:

$$\xi_f = \lambda_f^S/\lambda_f^N,\ \gamma_f = |X_f|^2/\lambda_f^N. \tag{3}$$

If the geometric mean of the log likelihood of Λ is greater than the threshold η, the state at time t is L_S; otherwise, L_N.

$$\log \Lambda(\boldsymbol{X}_t) = \frac{1}{F}\sum_{f=1}^{F} \log\left(\Lambda_f(X_f)\right) \underset{L_N}{\overset{L_S}{\gtrless}} \eta. \tag{4}$$

The noise variance λ_f^N is calculated from the observed noise, and the speech variance λ_f^S can be estimated according to the maximum likelihood criterion. Finally, the discriminant of speech and noise can be obtained as

$$\log \Lambda(\boldsymbol{X}_t) = \frac{1}{F}\sum_{f=1}^{F} (\gamma_f - \log\gamma_f - 1) \underset{L_N}{\overset{L_S}{\gtrless}} \eta. \tag{5}$$

The outputs, $\log\Lambda$, are smoothed by a hidden Markov model (HMM) hangover.

3. DNN-BASED METHOD

After the effectiveness of DNNs for ASR tasks had been proven [5], the effectiveness for VAD was confirmed [6, 7]. [6] uses a simple DNN structure. Acoustic features, which are derived from $\boldsymbol{X}$, $\boldsymbol{X}' \in \mathbb{R}^{F'\times T}$, are input into the network. The output values from two output nodes are $\boldsymbol{y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_t, \dots, \boldsymbol{y}_T] \in \mathbb{R}^{2\times T}$ where $\boldsymbol{y}_t$ is $[y_t(0); y_t(1)]$. $y_t(0)$ is the output of a speech node at time t and $y_t(1)$ is the output of a non-speech node. $\boldsymbol{y}$ is obtained after DNN operation $\mathcal{F}$ is applied to $\boldsymbol{X}'$ as

$$\boldsymbol{y} = \mathcal{F}(\boldsymbol{X}'). \tag{6}$$

During training, a DNN is trained to output a unity from the respective node corresponding to states L_N and L_S of the training data. During testing, the posterior probability of speech can be calculated from the softmax of their outputs. If the softmax of two nodes' outputs, $\mathcal{S}$, exceeds 0.5, the corresponding frame is determined as speech; otherwise noise.

$$\mathcal{S}(\boldsymbol{y}_t) = \mathcal{S}(\mathcal{F}(\boldsymbol{X}'_t)) = \frac{e^{y_t(0)}}{e^{y_t(0)} + e^{y_t(1)}} \underset{L_N}{\overset{L_S}{\gtrless}} 0.5. \tag{7}$$

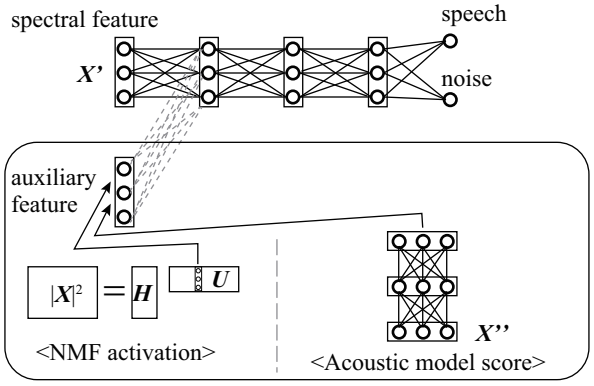

Fig. 1. Proposed DNN-based voice activity detection (VAD) system using auxiliary speech models.

4. AUXILIARY FEATURES OUTPUT FROM AUXILIARY SPEECH MODELS

[12, 13] improved the performance of DNN-based SE by using auxiliary features in addition to the spectral features. Similarly, [8] improved ASR performance by using auxiliary features that represented the speaker characteristics for DNN-based acoustic models. These methods use auxiliary features to adapt DNNs to various target environments. These can be also effective for VAD tasks.

VAD is a problem of discrimination between speech and noise, but it is had to solve this problem directly due to various speech patterns. If auxiliary features limit speech patterns, speech diversity can be reduced. For example, the features that represent phonemes simplify the discrimination between speech and noise into discrimination between a pre-estimated phoneme and noise. These simplifications can improve VAD performance. Fig. 1 shows the overview of the proposed system. We propose two types of auxiliary feature: NMF activation and acoustic model score.

4.1. NMF activation

NMF [14, 15] factorizes an element-wise square of an observation matrix $\boldsymbol{X}$ as

$$|\boldsymbol{X}|^2 \simeq \boldsymbol{H}\boldsymbol{U} = \boldsymbol{H}_s\boldsymbol{U}_s + \boldsymbol{H}_n\boldsymbol{U}_n, \tag{8}$$

where $\boldsymbol{H} \in \mathbb{R}_{\geq 0}^{F\times K}$ is a basis matrix composed of K bases and $\boldsymbol{U} \in \mathbb{R}_{\geq 0}^{K\times T}$ is an activation matrix that indicates an activation of the kth basis at frame t, $U_{k,t}$. When bases are composed of speech bases $\boldsymbol{H}_s$ and noise bases $\boldsymbol{H}_n$, their activations ($\boldsymbol{U}_s$ and $\boldsymbol{U}_n$) correspond to speech and noise bases, respectively. To separate noisy speech into speech and noise, initial values of speech bases are picked up from clean speech [17, 18]. Here, we use either speech activation, $\boldsymbol{U}_s$, or speech and noise activation, $\boldsymbol{U} = [\boldsymbol{U}_s; \boldsymbol{U}_n]$. NMF activation represents speech characteristics of an utterance well when the $\boldsymbol{H}$ bases are appropriately selected for the target speech. Actually, [19] proposed using speech activations combined with conditional random fields in order to detect speech. This paper defines $\boldsymbol{U}_s$ and $\boldsymbol{U}$ for an auxiliary feature as

$$\boldsymbol{y} = \mathcal{F}([\boldsymbol{X}'; \boldsymbol{U}]) \text{ or } \boldsymbol{y} = \mathcal{F}([\boldsymbol{X}'; \boldsymbol{U}_s]). \tag{9}$$

4.2. Acoustic score from ASR acoustic models

Precedent studies [12, 13] showed that a feedback of ASR results, such as one-hot vector of phonemes, to SE is effective. Here, we

propose to use acoustic scores of each phoneme computed by ASR acoustic models as auxiliary features of VAD. Acoustic features $\boldsymbol{X}''$ are input into acoustic models and the output is the acoustic scores of each phoneme, $\boldsymbol{s}$, via transformation $\mathcal{G}$, which can be Gaussian mixture models (GMMs) or DNN acoustic models. At that time, acoustic models with a small number of parameters are used because the computational load of VAD should be much smaller than that of ASR. Acoustic scores $\boldsymbol{s}$ are obtained as

$$\boldsymbol{s} = \mathcal{G}(\boldsymbol{X}''), \tag{10}$$

and output is obtained as

$$\boldsymbol{y} = \mathcal{F}\left(\left[\boldsymbol{X}''; \boldsymbol{s}\right]\right). \tag{11}$$

5. EXPERIMENTS

5.1. Experimental setups

We validated the effectiveness of our proposed method on the CENSREC-2 [16] dataset, which was recorded in real in-car environments[1]. There were three levels of driving speed (idling (i_a), low-speed (city) driving (c_a), and high-speed (highway) driving (e_a)). At each driving speed, the number of speakers were 58 in the training set and 15 in the evaluation set[2]. In CENSREC-2, each speech file includes one continuous utterance. For VAD experiments, we concatenated these utterances for each speaker in one file. The duration of each file was about one minute. At each driving speed, there were four types of in-car environment: regular driving, with the air conditioner switched on, with car audio on, and with window open. These different conditions were all mixed at almost the same portion, so that, in total, there were twelve environments. The results were summarized for each driving speed. Utterances were composed of eleven continuous digits (1–9, 0 (oh), and Z (zero)). CENSREC-2 did not include time labels for the speech or non-speech, thus, time labels were obtained from the ASR results of speech that were recorded by close-talking microphones. ASR acoustic models were trained in the matched condition ("Condition 3") by using attached scripts. Labeling of speech and non-speech with 10ms shifts was done using the time alignments of the ASR results with these models.

Table 1 shows the setup of the experiments. The acoustic features of the DNNs both for VAD and acoustic score calculation were

Table 1. Setup for VAD system.

Sampling frequency	16 kHz
Window length / shift	25 ms / 10 ms
Features	0–22th FBANK
Splice	9 frames
# NMF bases	50
# DNN output nodes	2
# DNN nodes per layer	1,000 nodes
DNN layer size	3 layers

[1]Dataset "CENSREC-1C" [20] was prepared for the VAD experiments but we constructed an original dataset from "CENSREC-2" because the sampling frequency of "CENSREC-1C" was 8kHz, which is unrealistic today. The utterances of "CENSREC-2" were the same as those in "CENSREC-1C".

[2]Training and evaluation sets were newly constructed by dividing the training set of CENSREC-2 because it was difficult to make labels for the original evaluation set of CENSREC-2, which did not include speech recorded by close-talking microphones.

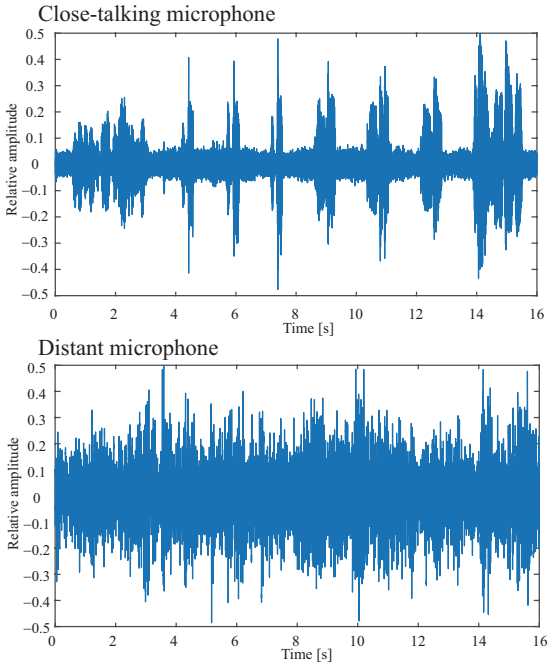

Fig. 2. Waveforms recorded by close-talking and distant microphones in the case of highway driving.

0–22 dimensional filterbank (FBANK) features with context expansion before and after four frames. One DNN was trained using all of the training utterances. The acoustic score calculation of GMM and ASR used 0–12 dimensional mel-frequency cepstral coefficients (MFCCs) with delta features. Two NMF activation patters were used; speech activation $\boldsymbol{U}_s$ and speech and noise activation $\boldsymbol{U}$. The threshold η of Sohn's method, which achieved the highest VAD performance on average, was common across all experimental conditions.

Fig. 2 compares the speech recorded by close-talking and distant microphones in the case of highway driving (e_a). In the case of close talking, voice activation can be seen, but in the case of distant talking, we cannot discriminate speech from background noise in terms of waveforms.

5.2. Baseline

Table 2 shows the average frame-level VAD accuracy for the baseline method (Sohn's method), Sohn's method with noise reduction by using minimum mean-square error short-time spectral amplitude estimator (MMSE-STSA) [21], DNN, and DNN with an MMSE-STSA. The DNNs' results were much better than those of Sohn's method. Although noise reduction significantly improved the performance using Sohn's method, it did not improve the performance when using the DNNs. As well as the DNN-based ASR, speech distortions were more harmful than noise regarding the DNN-based method [22].

5.3. NMF Activation

Table 2 also shows the results of our proposed method with NMF activations as auxiliary features. There were two types of NMF activation: activation of speech bases $\boldsymbol{U}_s$ (speech activation) and activation of all bases $\boldsymbol{U}$ including speech and noise (speech & noise activation). Both auxiliary features improved VAD accuracy from

Table 2. Average frame-level VAD accuracy [%]. Performance of DNNs was compared with that of conventional Sohn's method. DNNs used FBANK features with NMF activations.

	e_a	c_a	i_a
Sohn baseline	52.08	63.34	63.23
Sohn baseline (w MMSE-STSA)	62.25	60.81	65.06
DNN baseline (*1)	77.76	86.03	91.62
DNN baseline (w MMSE-STSA)	77.78	85.38	90.96
*1 + speech activation	79.37	**87.92**	**92.70**
*1 + speech & noise activation	**79.38**	87.79	92.64

Table 3. Average frame-level VAD accuracy [%]. DNNs used FBANK features with clean speech acoustic model (GMM/DNN) outputs.

	e_a	c_a	i_a
DNN baseline (*1)	77.76	86.03	91.62
*1 + speech GMM	80.23	88.33	92.94
*1 + speech DNN	**81.70**	**90.14**	**94.28**

the DNN baseline, but noise activation did not improve the accuracy further from that with speech-only activation. These comparisons show the effectiveness of activations corresponding to the speech bases.

5.4. Acoustic scores

Table 3 shows the results using acoustic scores output from acoustic models (GMM/DNN) as auxiliary features. The results were better than those using NMF activation in Section 5.3. In addition, the auxiliary DNN acoustic model was more effective than the GMM acoustic model.

Fig. 3 compares the log likelihood ratio by using Sohn's method, $\log \Lambda$, with the output speech posterior probability of the DNN baseline and out proposed DNN-based VAD with acoustic scores (calculated from the DNN models) for the utterance in Fig. 2. Both methods can detect at least a part of all utterances, but because the likelihood ratio had a larger variance, VAD performance of Sohn's method heavily depended on threshold η. On the other hand, the performance of the DNNs was better and it was easy to determine the optimal thresholds because the output posterior probabilities of the DNNs were stable. Our proposed method was more robust for non-speech than DNN baseline.

5.5. Necessity of smoothing

Table 4 shows the results with smoothing of frame-level VAD results across some contiguous frames. Smoothing significantly improved the performance of the DNN-based method. Sohn's method already used a smoothing-like method, namely HMM hangover, but the DNNs did not use such an explicit method, except implicit context expansion that used input features across some contiguous frames. Smoothing was effective for VAD because the speech was continuous.

Table 4. Average frame-level VAD accuracy [%] with smoothing.

	e_a	c_a	i_a
Sohn baseline	52.14	63.66	63.46
DNN baseline (*1)	80.91	89.02	94.03
*1 + speech activation	81.90	89.93	94.02
*1 + speech & noise activation	82.13	90.25	94.39
*1 + speech GMM	82.25	88.33	92.94
*1 + speech DNN	**82.68**	**91.19**	**94.91**

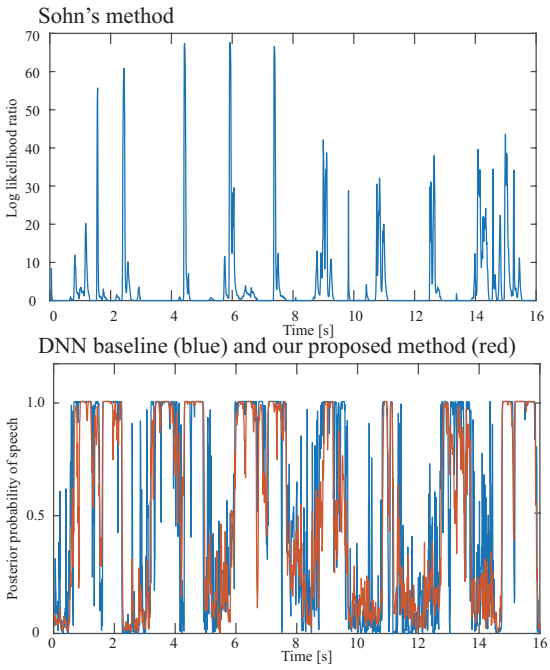

Fig. 3. Log likelihood ratio of Sohn's method, $\log \Lambda$, and speech posterior probability of DNN baseline and our proposed method.

Table 5. Word accuracy [%] of ASR for detected speech.

	e_a	c_a	i_a
Sohn baseline	18.37	40.79	44.70
DNN baseline (*1)	69.33	78.30	87.25
*1 + speech activation	72.70	78.87	87.37
*1 + speech & noise activation	73.00	79.15	87.06
*1 + speech GMM	**73.75**	**80.57**	88.35
*1 + speech DNN	72.70	80.28	**89.03**

5.6. Evaluation of ASR task

Table 5 shows the word accuracy for ASR experiments that were performed by using acoustic models trained with the baseline scripts attached to CENSREC-2. In this case, DNN-based VAD also outperformed Sohn's method. The proposed auxiliary features also improved ASR performance. It is important to detect speech with high accuracy because a lack of VAD directly leads to a drop in ASR performance.

6. CONCLUSION

This paper proposed auxiliary features output from auxiliary speech models in order to improve the performance of DNN-based VAD by adapting models to the target environments. The results of VAD experiments in noisy environments show that DNN-based VAD outperformed one of the most effective conventional methods. Experiments also show that auxiliary features that use NMF activations and acoustic scores of ASR models improved VAD performance. In addition, the proposed methods also improved the performance of ASR. Future work includes consideration of the time structure of acoustic features by using hidden-layer outputs of recurrent networks as an auxiliary feature.

7. REFERENCES

[1] L.R. Rabiner and M.R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *The Bell System Technical Journal*, vol. 54, pp. 297–315, 1975.

[2] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, 1999.

[3] Y. Tachioka, T. Narita, T. Hanazawa, and J. Ishii, "Voice activity detection based on density ratio estimation and system combination," in *Proceedings of APSIPA*, 2013, pp. 1–4.

[4] M. Fujimoto and K. Ishizuka, "Noise robust voice activity detection based on switching Kalman filter," *IEICE Transactions on Information and Systems*, vol. E91-D, pp. 467–477, 2008.

[5] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 28, pp. 82–97, 2012.

[6] 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. 1–14, 2013.

[7] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proceedings of ICASSP*, 2013, pp. 7378–7382.

[8] M. Delcroix, K. Kinoshita, T. Hori, and T. Nakatani, "Context adaptive deep neural networks for fast acoustic model adaptation," in *Proceedings of ICASSP*, 2015, pp. 4535–4539.

[9] D. Tran, M. Delcroix, A. Ogawa, C. Hümmer, and T. Nakatani, "Feedback connection for deep neural network-based acoustic modeling," in *Proceedings of ICASSP*, 2017.

[10] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[11] N. Hojo, Y. Ijima, and H. Mizuno, "An investigation of DNN-based speech synthesis using speaker codes," in *Proceedings of INTERSPEECH*, 2016, pp. 2278–2282.

[12] F. Sohrab and H. Erdogan, "Recognize and separate approach for speech denoising using nonnegative matrix factorization," in *Proceedings of EUSIPCO*, 2015.

[13] K. Kinoshita, M. Delcroix, A. Ogawa, and T. Nakatani, "Text-informed speech enhancement with deep neural networks," in *Proceedings of INTERSPEECH*, 2015, pp. 1760–1764.

[14] D.D. Lee and S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[15] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[16] K. Takeda, H. Fujimura, K. Itou, N. Kawaguchi, S. Matsubara, and F. Itakura, "Construction and evaluation a large in-car speech corpus," *IEICE Transactions on Information and Systems*, vol. E88-D, no. 3, pp. 553–561, 2005.

[17] B. Raj, R. Singh, and T. Virtanen, "Phoneme-dependent NMF for speech enhancement in monaural mixtures," in *Proceedings of INTERSPEECH*, 2011, pp. 1217–1220.

[18] Y. Tachioka, T. Narita, I. Miura, T. Uramoto, N. Monta, S. Uenohara, K. Furuya, S. Watanabe, and J. Le Roux, "Coupled initialization of multi-channel non-negative matrix factorization based on spatial and spectral information," in *Proceedings of INTERSPEECH*, 2017, pp. 2461–2465.

[19] P. Teng and Y. Jia, "Voice activity detection via noise reducing using non-negative sparse coding," *IEEE Signal Processing Letters*, vol. 20, no. 5, pp. 475–478, 2013.

[20] N Kitaoka, T Yamada, S Tsuge, C Miyajima, K Yamamoto, T Nishiura, Y Denda, M Fujimoto, T Takiguchi, S Tamura, S Matsuda, T Ogawa, S Kuroiwa, K Takeda, and S Nakamura, "CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments," *Acoustical Science & Technology*, vol. 30, pp. 363–371, 2009.

[21] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109–1121, 1984.

[22] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W.J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proceedings of ASRU*. 2015, pp. 436–443.