Context-Aware Neural Voice Activity Detection Using Auxiliary Networks for Phoneme Recognition, Speech Enhancement and Acoustic Scene Classification

Ryo Masumura, Kiyoaki Matsui, Yuma Koizumi, Takaaki Fukutomi, Takanobu Oba, Yushi Aono
NTT Media Intelligence Laboratories, NTT Corporation
masumura.ryou.ba@hco.ntt.co.jp

***Abstract*—This paper proposes a novel fully neural network based voice activity detection (VAD) method that estimates whether each speech segment is speech or non-speech even in very low signal-to-noise ratio (SNR) environments. Our innovation is to improve context-awareness of speech variability by introducing multiple auxiliary networks into the neural VAD framework. While previous studies reported that phonetic-aware auxiliary features extracted from a phoneme recognition network can improve VAD performance, none examined other effective auxiliary features for enhancing noise robustness. Thus, this paper present a neural VAD that uses auxiliary features extracted from not only the phoneme recognition network but also a speech enhancement network and an acoustic scene classification network. The last two networks are expected to improve context-awareness even in extremely low SNR environments since they can extract de-noised speech awareness and noisy environment awareness. In addition, we expect that combining these multiple auxiliary features yield synergistic improvements in VAD performance. Experiments verify the superiority of the proposed method in very low SNR environments.**

I. Introduction

Voice activity detection (VAD) which determines whether an acoustic segment is a speech or non-speech is essential for various speech applications such as automatic speech recognition (ASR), speaker recognition, and spoken language identification. One of the most important aspect for VAD is noise robustness in various environments because it is required to be used in any places and any environments.

Many VAD methods have been studied to achieve greater noise robustness. The basic method is to threshold time-domain energy [1]. While it performs well in clean environments, it is very weak against noise. Therefore, the main approach to VAD involves supervised learning. Likelihood-ratio-based methods that use generative models have been examined [2], [3]. In addition, discriminative models that uses support vector machine and conditional random fields have been introduced [4]. In recent studies, the fully neural network based VAD (neural VAD) has shown significant performance improvement [5]–[11]. One strength of neural VAD methods is their flexible capture of speech variability by using non-linear transformational functions such as, deep neural networks (DNNs) [5]–[7], recurrent neural networks (RNNs) [8], [9], and convolutional neural networks (CNNs) [10]. In particular, long short-term memory RNNs (LSTM-RNNs) can well handle the VAD problem since they can flexibly take long-range context information into consideration [11].

Although neural VAD schemes trained with multi-condition data sets offer significantly improved VAD performance, they falter when challenged with low signal-to-noise ratio (SNR) environments. In fact, it remains difficult to realize accurate VAD in very low SNR environments even if matched condition data sets can be prepared for training. In order to mitigate this problem, we focus on an approach that utilizes auxiliary features to enhance the context-awareness of speech variability. Some previous neural VAD methods used phonetic-aware auxiliary features extracted from phoneme recognition to improve VAD performance in noisy environments [12], [13]. However, no study has examined what other auxiliary features would be effective in achieving noise-robustness. In addition, the effect of combining multiple auxiliary features has not been addressed.

In this paper, we propose a context-aware neural VAD scheme that leverages multiple auxiliary features extracted from not only the phoneme recognition network, but also a speech enhancement network and an acoustic scene classification network. The last two are expected to improve context-awareness in very low SNR environments since they can extract de-noised speech awareness and noisy environment awareness. Inspired by the success of LSTM-RNN in VAD [11], we use just LSTM-RNNs to compose the auxiliary and main networks. Experiments show that each auxiliary network is effective for improving VAD performance in very low SNR environments. We also reveal that combining them provides a synergistic improvement in VAD performance.

II. Related Work

This section briefly describes fully neural network based methods for phoneme recognition, speech enhancement and acoustic scene classification. In addition, we present related work in which the networks are used for extracting auxiliary features in isolation.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

Phoneme recognition: In the ASR field, senone-based acoustic models are widely used for phoneme recognition. A senone represents a frame-level state within context-dependent phones. Many studies have introduced neural networks to senone-based acoustic models since they can attain significant performance superiority compared to Gaussian mixture model based methods [14], [15]. Senone-based acoustic models are currently being utilized for extracting phonetic-awareness features in speaker recognition [16], spoken language identification [17], and other speech based applications [18], [19].

Speech enhancement: There are two main approaches to neural network based speech enhancement. The first uses nonlinear mapping to convert noisy speech into clean speech [20]–[22]. The second estimates a soft mask that can eliminate noise interference [23]–[25]. This paper employs the first approach since it can extract de-noised speech awareness more directly. Previous studies combined neural speech enhancement with other speech applications for joint learning of front-end and back-end systems [26], [27]. However, no study has leveraged neural speech enhancement for auxiliary feature extraction.

Acoustic scene classification: Acoustic scene classification is now receiving a lot of attention. In contrast to frame-by-frame estimation such as used in phoneme recognition and speech enhancement, acoustic scene classification estimates a scene label from the entire input signal. Therefore, neural network based acoustic scene classification is often modeled by combining pooling or self-attention with RNNs or CNNs [28]–[31]. To the best of our knowledge, this paper is the first to utilize acoustic scene classification for auxiliary feature extraction in the speech field. The proposed method trains acoustic scene classification so as to discriminate the type of noisy environment (station, car, crowd, clean, etc.).

III. LSTM-RNN based Voice Activity Detection

This section describes LSTM-RNN based VAD as a baseline method [9], [11]. VAD demands the estimation of state sequence $\boldsymbol{S} = \{s_1, \cdots, s_T\}$ from input acoustic feature sequence $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$ in a frame-by-frame manner where the t-th state s_t represents either speech state or non-speech state. LSTM-RNN based VAD uses LSTM-RNN to estimate the conditional probability of $\boldsymbol{S}$ given $\boldsymbol{X}$, thus it can utilize long-range input information from start-of-utterance to the currently-being-processed frame. The conditional probability is defined as

$$P(\boldsymbol{S}|\boldsymbol{X}, \boldsymbol{\theta}) = \prod_{t=1}^{T} P(s_t|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta}$ is the model parameter. The t-th predictive probability is computed as

$$P(s_t|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \boldsymbol{\theta}) = \texttt{SOFTMAX}(\boldsymbol{z}_t; \boldsymbol{\theta}), \tag{2}$$

$$\begin{aligned} \boldsymbol{z}_t &= \texttt{LSTM}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t; \boldsymbol{\theta}) \\ &= \texttt{LSTM}(\boldsymbol{x}_t, \boldsymbol{z}_{t-1}; \boldsymbol{\theta}), \end{aligned} \tag{3}$$

where $\texttt{LSTM}()$ is a nonlinear transformational function based on unidirectional LSTM-RNNs and $\texttt{SOFTMAX}()$ is a linear transformational function with softmax activation.

The model parameter can be optimized by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} - \sum_{(\boldsymbol{X}, \boldsymbol{S}) \in \mathcal{D}^{\text{vad}}} \log P(\boldsymbol{S}|\boldsymbol{X}, \boldsymbol{\theta}), \tag{4}$$

where $\mathcal{D}^{\text{vad}}$ represents a training data set.

IV. Context-Aware Neural Voice Activity Detection using Auxiliary Networks

This section details context-aware neural VAD using auxiliary networks for phoneme recognition, speech enhancement and acoustic scene classification. The context-aware neural VAD is composed of three auxiliary networks and one main network. Auxiliary networks perform phoneme classification, speech enhancement and acoustic scene classification, and extract auxiliary features from the input acoustic features. The main network estimates the conditional probability of a state sequence from both input acoustic features and the auxiliary features. The context-aware neural VAD models conditional probability of state sequence $\boldsymbol{S} = \{s_1, \cdots, s_T\}$ given acoustic feature sequence $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$ as follows

$$\begin{aligned} P(\boldsymbol{S}|\boldsymbol{X}, \boldsymbol{\Lambda}, \boldsymbol{\theta}) &= \prod_{t=1}^{T} P(s_t|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \boldsymbol{\Lambda}, \boldsymbol{\theta}) \\ &= \prod_{t=1}^{T} P(s_t|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \boldsymbol{c}_t^{\text{pr}}, \boldsymbol{c}_t^{\text{se}}, \boldsymbol{c}_t^{\text{asc}}, \boldsymbol{\Lambda}, \boldsymbol{\theta}) \end{aligned} \tag{5}$$

where $\boldsymbol{\Lambda} = \{\boldsymbol{\lambda}^{\text{pr}}, \boldsymbol{\lambda}^{\text{se}}, \boldsymbol{\lambda}^{\text{asc}}\}$ is the model parameters for the auxiliary networks and $\boldsymbol{\theta}$ represents model parameter of the main network. $\boldsymbol{\lambda}^{\text{pr}}$, $\boldsymbol{\lambda}^{\text{se}}$, $\boldsymbol{\lambda}^{\text{asc}}$ represent model parameters of the auxiliary networks for a phoneme recognition, a speech enhancement and an acoustic scene classification. $\boldsymbol{c}_t^{\text{pr}}, \boldsymbol{c}_t^{\text{se}}, \boldsymbol{c}_t^{\text{asc}}$ are the t-th auxiliary features extracted from the auxiliary networks for phoneme recognition, speech enhancement and acoustic scene classification, respectively. The network structure of context-aware neural VAD composed by LSTM-RNNs is presented in Fig. 1. We detail each network structure and the optimization procedure in the following subsections.

A. Network strucutre

Auxiliary network for phoneme recognition: For phoneme recognition, we use neural networks to model the conditional probability of senone sequence $\boldsymbol{W} = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_T\}$ given input acoustic feature sequence $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$. The conditional probability is defined as

$$P(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{pr}}) = \prod_{t=1}^{T} P(w_t|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \boldsymbol{\lambda}^{\text{pr}}). \tag{6}$$

The t-th predictive probability is calculated as

$$P(w_t|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \boldsymbol{\theta}^{\text{pr}}) = \texttt{SOFTMAX}(\boldsymbol{c}_t^{\text{pr}}; \boldsymbol{\lambda}^{\text{pr}}), \tag{7}$$

$$\begin{aligned} \boldsymbol{c}_t^{\text{pr}} &= \texttt{LSTM}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t; \boldsymbol{\lambda}^{\text{pr}}) \\ &= \texttt{LSTM}(\boldsymbol{x}_t, \boldsymbol{c}_{t-1}^{\text{pr}}; \boldsymbol{\lambda}^{\text{pr}}), \end{aligned} \tag{8}$$

where $\boldsymbol{c}_t^{\text{pr}}$ represents a context vector that embeds information effective in determining the t-th phoneme state.

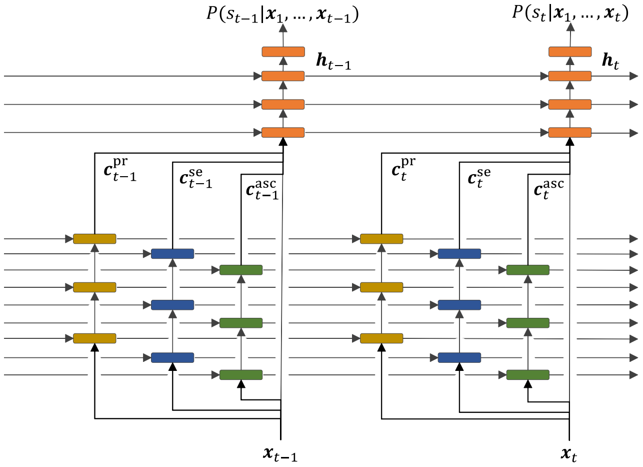

Fig. 1. Network structure of context-aware neural VAD.

Auxiliary network for speech enhancement: For speech enhancement, we use neural networks to model a regression problem that estimates de-noised acoustic features $\bar{\boldsymbol{X}} = \{\bar{\boldsymbol{x}}_1, \cdots, \bar{\boldsymbol{x}}_T\}$ from input acoustic features $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$. The t-th de-noised acoustic feature $\bar{\boldsymbol{x}}_t$ is estimated by

$$\bar{\boldsymbol{x}}_t = \texttt{LINEAR}(\boldsymbol{c}_t^{\text{se}}; \boldsymbol{\lambda}^{\text{se}}), \tag{9}$$

$$\begin{aligned} \boldsymbol{c}_t^{\text{se}} &= \texttt{LSTM}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t; \boldsymbol{\lambda}^{\text{se}}) \\ &= \texttt{LSTM}(\boldsymbol{x}_t, \boldsymbol{c}_{t-1}^{\text{se}}; \boldsymbol{\lambda}^{\text{se}}), \end{aligned} \tag{10}$$

where $\texttt{LINEAR}()$ represents the linear transformational function. $\boldsymbol{c}_t^{\text{se}}$ represents a context vector that embeds information effective in estimating the t-th de-noised acoustic feature.

Auxiliary network for acoustic scene classification: For acoustic scene classification, we use neural networks to model the conditional probability of scene label $L \in \mathcal{L}$ from an input acoustic features $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$ where $\mathcal{L}$ represents acoustic scene sets. The conditional probability is calculated by

$$P(L|\boldsymbol{X}) = \texttt{SOFTMAX}(\boldsymbol{U}; \boldsymbol{\lambda}^{\text{asc}}), \tag{11}$$

$$\boldsymbol{U} = \texttt{SelfAttention}(\boldsymbol{c}_1^{\text{asc}}, \cdots, \boldsymbol{c}_T^{\text{asc}}; \boldsymbol{\lambda}^{\text{asc}}), \tag{12}$$

$$\begin{aligned} \boldsymbol{c}_t^{\text{asc}} &= \texttt{LSTM}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t; \boldsymbol{\lambda}^{\text{asc}}) \\ &= \texttt{LSTM}(\boldsymbol{x}_t, \boldsymbol{c}_{t-1}^{\text{asc}}; \boldsymbol{\lambda}^{\text{asc}}), \end{aligned} \tag{13}$$

where $\texttt{SelfAttention}()$ is a function that summarizes frame-level vectors as one vector using the self-attention mechanism [30]. $\boldsymbol{c}_t^{\text{asc}}$ represents a context vector that embeds information effective in estimating the acoustic scene label.

Main network: The main network estimates frame-level predictive probability from both input acoustic features and auxiliary features extracted from the auxiliary networks. The t-th predictive probability is calculated as

$$P(s_t|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \boldsymbol{\Lambda}, \boldsymbol{\theta}) = \texttt{SOFTMAX}(\boldsymbol{h}_t; \boldsymbol{\theta}), \tag{14}$$

$$\begin{aligned} \boldsymbol{h}_t &= \texttt{LSTM}(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_t; \boldsymbol{\theta}) \\ &= \texttt{LSTM}(\boldsymbol{u}_t, \boldsymbol{h}_{t-1}; \boldsymbol{\theta}), \end{aligned} \tag{15}$$

$$\boldsymbol{u}_t = [\boldsymbol{x}_t^\top, \boldsymbol{c}_t^{\text{pr}\top}, \boldsymbol{c}_t^{\text{se}\top}, \boldsymbol{c}_t^{\text{asc}\top}]^\top, \tag{16}$$

where $\boldsymbol{u}_t$ is the t-th concatenated vector of both an input acoustic feature and the auxiliary features extracted from the auxiliary networks. While LSTM-RNN based VAD only handles input acoustic features, context-aware neural VAD handles the concatenated vector. This enables us to efficiently take multiple context-awareness features into consideration.

B. Optimization Procedure

The context-aware neural VAD is trained in two steps. In the first step, the auxiliary networks are initially trained using data sets prepared for individual auxiliary tasks. The model parameters $\boldsymbol{\lambda}^{\text{pr}}$, $\boldsymbol{\lambda}^{\text{se}}$, $\boldsymbol{\lambda}^{\text{asc}}$ are optimized by

$$\hat{\boldsymbol{\lambda}}^{\text{pr}} = \underset{\boldsymbol{\lambda}^{\text{pr}}}{\arg\min} - \sum_{(\boldsymbol{X},\boldsymbol{W}) \in \mathcal{D}^{\text{pr}}} \log P(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{\lambda}^{\text{pr}}), \tag{17}$$

$$\hat{\boldsymbol{\lambda}}^{\text{se}} = \underset{\boldsymbol{\lambda}^{\text{se}}}{\arg\min} \sum_{(\boldsymbol{X},\bar{\boldsymbol{X}}) \in \mathcal{D}^{\text{nr}}} \sum_{t=1}^{|\boldsymbol{X}|} |\bar{\boldsymbol{x}}_t - \boldsymbol{x}_t|^2, \tag{18}$$

$$\hat{\boldsymbol{\lambda}}^{\text{asc}} = \underset{\boldsymbol{\lambda}^{\text{asc}}}{\arg\min} - \sum_{(\boldsymbol{X},L) \in \mathcal{D}^{\text{asc}}} \log P(L|\boldsymbol{X}, \boldsymbol{\lambda}^{\text{asc}}), \tag{19}$$

where $\mathcal{D}^{\text{pr}}$, $\mathcal{D}^{\text{se}}$, $\mathcal{D}^{\text{asc}}$ are training data sets for phoneme recognition, speech enhancement and acoustic scene classification, respectively.

In the second step, the main network is trained using a data set for VAD while preserving the model parameters for the auxiliary networks. Thus, model parameter $\boldsymbol{\theta}$ is optimized by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\arg\min} - \sum_{(\boldsymbol{X},\boldsymbol{S}) \in \mathcal{D}^{\text{vad}}} \log P(\boldsymbol{S}|\boldsymbol{X}, \hat{\boldsymbol{\Lambda}}, \boldsymbol{\theta}), \tag{20}$$

where $\hat{\boldsymbol{\Lambda}}$ represents $\{\hat{\boldsymbol{\lambda}}^{\text{pr}}, \hat{\boldsymbol{\lambda}}^{\text{se}}, \hat{\boldsymbol{\lambda}}^{\text{asc}}\}$ and $\mathcal{D}^{\text{vad}}$ is the training data set for VAD.

V. Experiments

Our experiments used a large scale Japanese training data set. First, we prepared a home-made 1,500 hour clean speech data set with manual transcriptions; senone states and VAD states were automatically annotated using LSTM acoustic models with 3,072 senone states trained from the clean speech data set. Note that speech/non-speech states can be converted from the senone states. Next, we created a noisy speech data set by synthesizing 120 manually-constructed noise types (car, shopping mall, factory, etc.) to the clean speech data set; SNR levels were randomly varied between -10 dB to 30 dB. Both the clean and noisy speech data sets were used for learning VAD and phoneme recognition. A parallel data set of clean speech and noisy speech was employed for learning speech enhancement. The noisy speech data set, SNR under 0 dB, was leveraged for learning acoustic scene classification. In this case, noise type corresponds to acoustic scene. The sampling rate of all data sets was 16 kHz. Table 1 details each training data set.

For testing, we also prepared a clean data set that was not included in the training data set. The data set included 5,040 utterances and its VAD states were manually annotated. The

TABLE II
EXPERIMENTAL RESULTS IN TERMS OF AUC (%).

Noise	SNR	DNN-VAD	LSTM-VAD	PR-LSTM-VAD	SE-LSTM-VAD	ASC-LSTM-VAD	PR-SE-ASC-LSTM-VAD
Crowd	-10 dB	76.35	89.96	92.13	91.36	91.05	**93.14**
	-5 dB	86.83	95.56	96.75	96.65	96.21	**97.28**
	0 dB	94.45	98.25	98.82	98.81	98.35	**99.11**
	5 dB	97.70	99.30	99.52	99.51	99.32	**99.62**
	10 dB	98.85	99.67	99.77	99.74	99.66	**99.80**
Station	-10 dB	81.67	90.34	90.82	91.22	90.88	**92.18**
	-5 dB	89.01	95.47	95.85	96.00	95.78	**96.54**
	0 dB	94.42	98.01	98.32	98.54	98.24	**98.85**
	5 dB	97.25	99.03	99.24	99.30	99.05	**99.40**
	10 dB	98.53	99.49	99.62	99.63	99.44	**99.68**

TABLE I
TRAINING DATA SETS.

	Size (hours)
Clean speech data set	1,500
Data set for VAD: $\mathcal{D}^{\rm vad}$	10,500
Data set for phoneme recognition: $\mathcal{D}^{\rm pr}$	10,500
Data set for speech enhancement: $\mathcal{D}^{\rm se}$	9,000
Data set for acoustic scene classification: $\mathcal{D}^{\rm asc}$	3,000

evaluation data set was prepared by corrupting the utterances with two unseen noise types (crowd and station) at 6 noise levels; i.e., SNR values of -10 dB, -5 dB, 0 dB, 5 dB, 10 dB.

A. Setups

Our experiments evaluated 6 VAD methods. In each method, we used 38 dimensional mel-frequency cepstrum coefficients (12 MFCC, 12 ΔMFCC, 12ΔΔMFCC, Δpower and ΔΔpower) as acoustic features; they were extracted using 20 msec windows shifted by 10 msec. Additionally, 418 dimensional acoustic features were formed by stacking the current processed frame and its ±5 left-right context.

As baselines, we constructed "DNN-VAD" (5-layer sigmoid non-linear function with 256 units [6]), and "LSTM-VAD" (3-layer LSTM-RNN with 256 units [9]). For DNN-VAD, we introduced post-processing to smooth outputs since DNN cannot consider long-range information. Both baselines were trained using only the data set for VAD. As variants of the proposed method, we constructed "PR-LSTM-VAD", "SE-LSTM-VAD" and "ASC-LSTM-VAD"; they individually utilized an auxiliary network for phoneme recognition, speech enhancement, and acoustic scene classification. We also constructed "PR-SE-ASC-LSTM-VAD"; it utilizes the three auxiliary networks, simultaneously. The auxiliary networks and main network consisted of 3-layer LSTM-RNN with 256 units. To optimize the individual methods, we used Adam optimizer with default settings. For the mini-batch training, we partitioned each speech into 400 ms and the mini-batch size was set to 64. Note that a part of the training data set were used for early stopping.

B. Results

Experimental results in terms of the area under receive operating characteristic curve (AUC) [32] are shown in Table 2. We constructed five models for each setups and evaluated averaged results. First, LSTM-VAD outperformed DNN-VAD in all SNR environments even though LSTM-VAD did not use post-processing. This is due to its excellent ability to consider long-range context information. Next, PR-LSTM-VAD, SE-LSTM-VAD and ASC-LSTM-VAD yielded VAD performance improvements compared with LSTM-VAD in all conditions. In particular, VAD performance in low SNR environments was significantly improved by introducing the auxiliary networks. This indicates that phonetic awareness, de-noised speech awareness, and noisy environment awareness are effective in enhancing VAD performance. The experiments showed that phoneme awareness was the most effective in the crowd noise environment, while de-noised speech awareness was most effective in the station noise environment. These results suggest that the effectiveness of context-awareness depends on the noise environments. The highest AUC results were obtained by PR-SE-ASC-LSTM-VAD. This verifies that combining multiple context-awareness is effective.

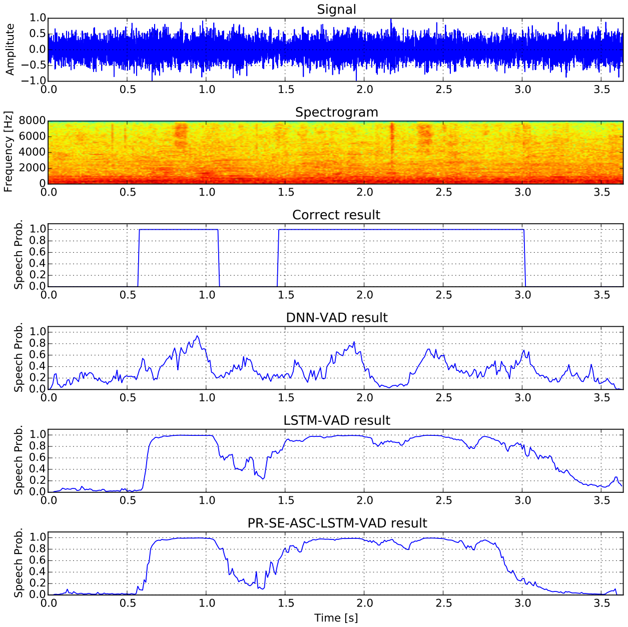

Fig. 2. VAD results for crowd noise at -10 dB SNR.

Figure 2 shows typical results for crowd noise environment with SNR of -10 dB. The result shows DNN-VAD could not estimate speech/non-speech segments at all. LSTM-VAD and PR-SE-ASC-LSTM-VAD correctly estimated speech segments. In particular, PR-SE-ASC-LSTM-VAD could more precisely identify speech/non-speech boundaries than LSTM-VAD.

VI. Conclusions

This paper proposed a context-aware neural VAD scheme that utilizes the power of auxiliary networks for phoneme recognition, speech enhancement and acoustic scene classification to improve VAD performance in very low SNR environments. The strength of the proposed method is to simultaneously leverage phonetic awareness, de-noised speech awareness, and noisy environment awareness (extracted by auxiliary networks) for detecting speech/non-speech segments. Experiments showed that each type of context-awareness was effective in improving VAD performance in low SNR environments. Furthermore, the highest VAD performance was attained by simultaneously utilizing all types of context-awareness.

References

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

[2] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[3] S. Gazor and W. Xhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, pp. 204–207, 2003.

[4] J. W. Shin, J. H. Chang, and N. S. Kim, "Voice activity detection based on statistical models and machine learning approaches," *Computer Speech & Language*, vol. 24, no. 3, pp. 515–530, 2010.

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

[6] N. Ryant, M. Liberman, and J. Yuan, "Speech actibity detection on youtube using deep neural networks," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 728–731, 2013.

[7] X.-L. Zhang and D. Wang, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 2, pp. 252–264, 2016.

[8] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 7378–7382, 2013.

[9] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to hollywood movies," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 483–487, 2013.

[10] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismathed acoustic conditions," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2519–2523, 2014.

[11] S. Tong, H. Gu, and K. Yu, "A comparative study of robustness of deep learning approraches for VAD," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5695–5699, 2016.

[12] L. Ferrer, M. Graciarena, and V. Mitra, "A phonetically aware system for speech activity detection," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5710–5714, 2016.

[13] Y. Tachioka, "DNN-based voice actibity detection using auxiliary speech models in noisy environments," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5529–5533, 2018.

[14] G. Hinton, L. Deng, D. Yu, G. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury, and T. Sainath, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, pp. 82–97, 2012.

[15] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dpendent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE transactions on Audio, Speech and Language Processing*, vol. 20, pp. 30–42, 2012.

[16] M. McLaren, L. Ferrer, and A. Lawson, "Exploring the role of phonetic bottlececk features for speaker and language rocognition," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5575–5579, 2016.

[17] R. Masumura, T. Asami, H. Masataki, and Y. Aono, "Parallel phonetically aware DNNs and LSTM-RNNs for frame-by-frame discriminative modeling of spoken language identification," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5260–5264, 2017.

[18] R. Masumura, T. Asami, H. Masataki, R. Ishii, and R. Higashinaka, "Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1661–1665, 2017.

[19] A. Ando, R. Asakawa, R. Masumura, H. Kamiyama, S. Kobashikawa, and Y. Aono, "Automatic question detection from acoustic and phonetic features using feature-wise pre-training," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1731–1735, 2018.

[20] A. L. Maas, Q. V. Le, T. M. O ' Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 22–25, 2012.

[21] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 436–440, 2013.

[22] Z. Chen, Y. Huang, J. Li, , and Y. Gong, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[23] A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 4, pp. 826–835, 2014.

[24] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 708–712, 2015.

[25] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "DNN-based source enhancement to increase objective sound quality assessment score," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 10, pp. 1780–1792, 2018.

[26] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4375–4379, 2015.

[27] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 4, pp. 796–806, 2016.

[28] Y. Wang, L. Neves, and F. Metze, "Audio-based multimedia event detection using deep recurrent neural networks," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2742–2746, 2016.

[29] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 559–563, 2015.

[30] Z. Ren, Q. Kong, K. Qian, M. D. Plumbley, and B. W. Schuller, "Attention-based convolutional neural networks for acoustic scene classification," *In Proc Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE Workshop)*, pp. 39–43, 2018.

[31] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 2, pp. 379–393, 2018.

[32] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receive operating characteristic (roc) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.