DNN-Based Voice Activity Detection with Local Feature Shift Technique

Tae Gyoon Kang, Kang Hyun Lee, Woo Hyun Kang, Soo Hyun Bae, and Nam Soo Kim
Department of Electrical and Computer Engineering and INMC, Seoul National University, Seoul, Korea
E-mail: {tgkang, khlee, whkang, shbae}@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract—Recently, the deep neural networks (DNNs) are successfully adopted into the voice activity detection (VAD) area. However, the performance of the DNN-based VAD is still unsatisfactory in noise environments where the feature subspace of the training database and the test environments are not matched with each other. In this paper, we propose a local feature shift technique which normalizes the feature subspaces over various noise environments. The proposed technique considers the local minimum vectors of the log-Mel filterbank features as noise power estimates and produces feature shift vectors from them. The experimental results in stationary and non-stationary noise environments show that the DNN with the proposed technique outperforms the conventional DNN-based VAD algorithms.

I. INTRODUCTION

Voice activity detection (VAD) has been widely studied as a pre-processing algorithm for many speech communication and processing systems. Early studies on the VAD area tracked the speech and noise statistics in order to estimate the voice activity status under the assumption of stationary background noise [1], [2]. Though the algorithms with this approach can estimate the voice active intervals from noisy utterances with affordable computational cost, they have difficulties in the non-stationary noise environments where the stationary background noise assumption does not hold.

To estimate the voice active intervals without the stationary noise assumption, recent studies adopt the machine learning approach to the VAD area [3], [4], [5] in which the VAD problem is considered as a two-class classification task. By learning the mapping between the noisy speech features and the corresponding voice active labels from a large amount of training database, they can estimate voice activity status in non-stationary noise environments. Among the algorithms with this approach, the DNN-based algorithm showed promising results compared to other statistical model-based and machine learning-based algorithms [6].

One of the critical issues of the DNN-based VAD algorithm is that its performance decreases in the unseen noise environments. In these environments, the feature subspaces of noisy speech features do not match with those of the training database due to the bias from different noise characteristics. This performance degradation of the VAD could reduce the reliability of the speech signal processing and communication systems in some environments.

A similar problem has been addressed in the speaker adaptive automatic speech recognition (ASR) area where the feature subspaces vary over different speakers. Many studies investigate the speaker adaptive ASR task by adapting the model parameter subspace or normalizing the feature subspaces. Xue et al. propose a speaker code vector which modifies the model parameter subspace for each speaker [7]. In contrast, Miao et al. use i-vectors [8] to obtain speaker-specific linear feature shifts with the adaptation neural network [9]. In this paper, we adopt a similar approach with [9] and propose the local feature shift (LFS) technique for robust DNN-based VAD.

In contrast to speaker information which is represented over the utterances, the noise characteristics for non-stationary noise environments need to be estimated locally from the set of adjacent noisy speech frames. In this paper, we consider the local minimum vectors of the log-Mel filterbank feature as the noise power and shift the input features with them. The feature learning approach is also combined to the proposed algorithm in order to tune the parameters of LFS and DNN. The experimental results show that the performance of the proposed algorithm outperformed the conventional DNN-based VAD in both the NOISEX-92 [10] and ITU-T recommendation P.501 [11] noise environments.

II. A REVIEW ON DNN-BASED VAD ALGORITHM

The DNN consists of an input layer, several hidden layers, and an output layer which are fully connected to their adjacent layers. For the input layer, the feature vector is extracted from each frame of the noisy speech and concatenated over several adjacent frames for input context expansion. In this paper, we use the log-Mel filterbank coefficients which are widely used in many speech processing systems.

Let us denote an $f$-th log-Mel filterbank coefficient at $t$-th frame as $x(t,f)$. Then, the input feature of the DNN $x_t$ is defined as follows:

\[
x_t = [x(t,1), x(t,2), \ldots, x(t,F)], \tag{1}
\]

\[
x_t = [x(t-T, T), x(t-T+1), \ldots, x(t+T)] \tag{2}
\]

where $F$ denotes the number of Mel-scale filterbanks and $T$ denotes the input context expansion parameter. In this paper, $F$ and $T$ are fixed to 26 and 5, respectively. Before $x_t$ is fed
to the DNN, they are normalized to have zero mean and unit variance.

For each hidden layer of the DNN, the rectified linear activation function is used as the activation function [12]. We use a single node with the logistic sigmoid function for the output layer.

The parameters of the DNN are randomly initialized and no pre-training is performed. The network is trained by the stochastic gradient descent algorithm to minimize the cross entropy loss function $C_{CE}$ which is given by

$$C_{CE} = -y \ln(\hat{y}) - (1 - y) \ln(1 - \hat{y})$$  \hspace{1cm} (3)

where $\hat{y}$ and $y$ denote the estimated and reference voice active probabilities, respectively.

In the test stage, $\hat{y}$ is estimated from $\bar{x}_t$ using the feed-forward algorithm. We apply a simple decision rule to $\hat{y}$ as follows:

$$H = \begin{cases} H_{speech}, & \text{if } \hat{y} > \eta, \\ H_{noise}, & \text{otherwise} \end{cases}$$  \hspace{1cm} (4)

where $H_{speech}$ and $H_{noise}$ denote the voice active and noise-only hypothesis, respectively. $\eta$ is a threshold parameter which is fixed to 0.5 in this paper.

### III. LOCAL FEATURE SHIFT TECHNIQUE

In this section, the proposed LFS technique is described in detail. Similar to the statistical-model based algorithms in the speech enhancement area [13], [14], we describe the noise power with the local minimum statistics from several adjacent frames. In the proposed algorithm, the log-Mel filterbank coefficient vectors are shifted by the local minimum vectors. Then, the shifted features and local minimum vectors are concatenated before they are fed to DNN to compensate the distortion that might be introduced from normalization.

Fig. 1 shows the scheme of the proposed LFS technique with the DNN-based VAD algorithm. In the proposed technique, an element-wise minimum function is applied to each set of the log-Mel filterbank vectors and the local minimum vector $\bar{x}_t$ is defined as follows:

$$\bar{x}(t, f) = \min(x(t - T, f), x(t + T, f))$$  \hspace{1cm} (5)

where $\min(a_1, a_2, \cdots, a_N)$ is the minimum function over $N$ elements $(a_1, a_2, \cdots, a_N)$. Then, the shifted feature vector $\tilde{x}_t$ is obtained from $x_t$ and $\bar{x}_t$ as follows:

$$\tilde{x}_t = \tilde{x}_t - \bar{x}_t W, \hspace{1cm} (7)$$

where $W$ denotes an $F \times F$-dimensional identity matrix.

After $x_t$ and $\tilde{x}_t$ are obtained, we concatenate them and fed to the DNN. The DNN in the proposed algorithm is trained by the same stochastic gradient descent algorithm with the conventional DNN-based VAD.

Similar to the cases of filterbank and delta operation in the ASR area, the hand-crafted transform matrix $W$ does not guaranteed to be optimal value and can be jointly trained for improved performance [15]. After the DNN is fine-tuned with the fixed $W$ initialized by (8), we fix the parameters of the DNN and optimize $W$ with gradients from $C_{CE}$ considering the local feature shift technique as a hidden layer.

Compared to the algorithm without the joint training scheme, the optimal value for $W$ can be determined from the training data while the over-fitting problem could limit the performance of VAD in some environments. The performance of the proposed algorithms with or without joint training are compared in the experiments to verify if the joint training scheme is effective.

### IV. RELATION TO PRIOR WORK

The robustness of the input feature structure heavily affects the performance of the VAD algorithm. Many studies focused on improving the performance of the input features in two ways.

First, various feature structures are extracted and concatenated for each frame. By incorporating various feature structures, the DNN can classify the noisy speech features with richer information than that with the single feature structure. Zhang et al. show the performance increment of the DNN-based VAD with respect to a number of features [6] in which the performance of the DNN-based VAD gradually increases when information from various feature structures is utilized by the DNN simultaneously. Drugman et al. also merge various filter-based and source-based features and evaluate the performance of them [16].
Second, the novel feature structures which emphasize discriminant characteristics of speech are proposed. Zhang et al. apply the multi-resolution cochleagram feature to improve the DNN-based VAD algorithm [17]. Yoo et al. propose the VAD algorithm which utilizes the unique distribution of formant frequency in human voice [18] despite they do not incorporated the proposed feature to the DNN framework.

Rather than concatenating various features or developing more discriminative features, we focus on the compensating feature subspace mismatch over noise environments with the log-Mel filterbank feature which is popular for ASR or speaker recognition systems [16], [19], [20]. The proposed algorithm does not increase the complexity of the speech processing systems significantly since the proposed VAD algorithm utilizes the feature that is already extracted for other applications. Also, compared to the previous studies which mainly depend on hand-crafted features, the proposed algorithm adopts the feature learning approach and modifies the parameters of the LFS techniques in the training stage.

V. EXPERIMENTS

In order to evaluate the performance of the proposed technique, we conducted a set of VAD experiments in various mismatched noise environments. In the experiments, 1,500 utterances of clean speech data were randomly taken from the TIMIT training database to build the DNN training set. The white, factory, babble, and machinegun noises from NOISEX-92 database were used for the training set. For each clean speech utterance and noise type, the noisy speech utterances were artificially generated with SNRs from -5 to 5 dB with 5 dB step. Each waveform was sampled at 16 kHz, and a 512-point Hamming window with 50% overlap was applied.

We conducted experiments in the NOISEX-92 database and the ITU-T recommendation P.501 database [11]. In both test sets, 200 utterances of clean speech data were randomly taken from the TIMIT test set. The buccaneer, hfchannel, and volvo noises were used for NOISEX-92 test set and the cafeteria, incar, and street noises were used for ITU-T recommendation P.501 test set. For each clean speech utterance and noise type, the noisy speech utterances were artificially generated with SNRs from -5 to 5 dB with 5 dB step. Note that the NOISEX-92 test set represented stationary noise environments while the ITU-T recommendation P.501 test set contained non-stationary background noises.

In the experiments, we trained DNNs with various input features and training methods as follows:

- **Mel**: the log-Mel filterbank feature \( \tilde{x}_t \).
- **Mel+Min**: the log-Mel filterbank feature concatenated with the local minimum vector \( \tilde{x}_t, s_t \).
- **Mel**: the shifted log-Mel filterbank feature \( \tilde{z}_t \).
- **Mel+Min**: the shifted log-Mel filterbank feature concatenated with the local minimum vector \( \tilde{z}_t, s_t \).
- **Mel+Min(joint)**: the shifted log-Mel filterbank feature concatenated with the local minimum vector \( \tilde{z}_t, s_t \) and jointly trained \( W \).

The DNNs were implemented using the Theano neural network toolkit [21]. The DNNs were constructed by stacking 2 hidden layers with 128 nodes each. In the training stage, we used the learning rate of 0.05 in the first 10 epochs and decreased by 10% after each subsequent epoch. The momentum was 0.5 for the first 5 epochs and increased to 0.9 afterward. The dropout rates of the input layer and all hidden layers were set to 0.1 and 0.2, respectively. The mini-batch size was fixed to 128. In Mel+Min(joint), \( W \) was fine-tuned with the learning rate of 0.00001 for 10 epochs.

We used the frame accuracies of DNNs as an evaluation metric. Table 1 shows the frame accuracies of various VADs in the NOISEX-92 test set. In this table, the Mel+Min feature showed slightly worse performance than the Mel feature. The results show that the local minimum vector was not helpful when it was simply concatenated with the log-Mel filterbank feature. In contrast, the Mel+Min feature outperformed Mel in buccaneer and hfchannel noises while it showed worse performance than Mel in volvo noise. Concatenating the local minimum vector in Mel+Min also improved the performance of the DNN. Note that the Mel+Min outperformed or showed similar performance with Mel for all types of SNR values and noise environments. Finally, the joint training of \( W \) slightly increased the performance of DNN while the effect of the joint training was inconsistent over noise types.

Tables 2 shows the frame accuracies of various DNNs in the ITU-T recommendation P.501 test set. In this test set, both Mel+Min and Mel did not show significant performance improvement over Mel. In this test set, the performance gain from the input feature shift was not significant since the simple local minimum vector had difficulty to represent the highly non-stationary noise features.

However, by alleviating the speech distortion with feature concatenation, Mel+Min outperformed the Mel in all types of SNR values and noise environments. The results show that the proposed technique is helpful in both the stationary and non-stationary noise environments. Similar to the results in the NOISEX-92 database, the jointly training \( W \) increased the performance of the DNN in the cafeteria noise while Mel+Min showed slightly better performance in inca and street noises.

From the results, we observed that (i) feature shift with the local minimum vectors improves the performance of the DNN in stationary noise environments. (ii) Concatenating the local minimum vectors with the shifted features increases the performance of VAD in both the stationary and non-stationary noise environments. (iii) The joint training scheme could be helpful to the proposed LFS technique while its contribution is not consistent over noise environments.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed the LFS technique to ameliorate the performance degradation caused by feature subspace mismatch between the training and test environments. From the experimental results, it has been found that the proposed technique outperforms the conventional DNN-based
VAD in both the stationary and non-stationary noise environments. The future work will focus on the end-to-end DNN-based VAD algorithm whose parameters are initialized with the knowledge of hand-crafted speech features and optimized in the training stage.

ACKNOWLEDGEMENT

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (NRF-2015R1A2A1A15054343), and by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2016-H8501-16-1016) supervised by the IITP (Institute for Information & Communications Technology Promotion).

REFERENCES