A Statistical Model-Based Voice Activity Detection Using Multiple DNNs and Noise Awareness

Inyoung Hwang, Jaeseong Sim, Sang-Hyeon Kim, Kwang-Sub Song, Joon-Hyuk Chang

Department of Electronic and Computer Engineering
Hanyang University, Seoul, Korea.
{i.yhwang88, sjsjaesung, wkdustksala, sentel103, jchang}@hanyang.ac.kr

Abstract

In this paper, we propose the ensemble of deep neural networks (DNNs) by using acoustic environment classification for statistical model-based voice activity detection (VAD). Since conventional decision functions for statistical model-based VAD are based on shallow model and it cannot take advantage of the diversity of the space distribution of features, we present to use the multiple DNNs separately trained on different noise condition as decision function for the statistical model-based VAD. And, environmental noise classification is also performed based on the separate DNN since acoustic environment classification makes it possible to achieve high detection performance at various type of noise environment by using different algorithm according to current noise condition. In the training stage, a number of DNNs are independently trained according to different type of noise environments, and separate DNN is organized to detect one of the environmental conditions. In an online stage, the environmental knowledge on each frame is contributed to allow us to combine the speech presence probabilities, which are derived from the ensemble of the trained DNNs for the individual environment. Our approach for VAD was evaluated in terms of objective measures and showed significant improvement compared to the conventional algorithm.

Index Terms: voice activity detection, statistical model, acoustic environment classification, deep neural network, ensemble

1. Introduction

Voice activity detection (VAD) which classifies the period of speech and non-speech from an input speech signal is an essential part of speech signal processing. Among the various VAD methods, we focus on a statistical model-based approach, which was originated from the work of Ephraim and Malah [1] for speech enhancement, due to its high detection accuracy as well as low computational complexity. Sohn et al. [2] devised the VAD based on a Gaussian statistical model by employing the decision rule based on the geometric mean of the likelihood ratio (LR). The novelty of the statistical model-based VAD was extensively reviewed, so that further improved methods based on the LR test according to two hypotheses for speech presence and absence have been presented in many studies [3]-[5]. However, it has been found that the LR corresponding to speech absence and presence cannot be separated by conventional decision functions based on shallow model such as support vector machine (SVM) due to considerable class overlap of LR in the feature space [5]. Recently, the deep belief network (DBN) has been proposed, by Hinton and Salakhutdinov [6], as a powerful hierarchical generative model not only for feature representation but also for classification by taking nonlinear multiple-layered deep architecture. It is noted that the DBN is known to avoid the poor local-minima and over-fitting which are occurred during a training stage by the greedy layer-wise unsupervised learning process called pre-training. The superiority of the DBN has been reviewed and successfully applied to various pattern recognition applications as a state-of-the-art technique [7]-[9]. Especially, the key idea behind the method of Zhang and Yu [9] is to extract new features by transferring the acoustic features through deep nonlinear hidden layers since the deep model can combine multiple features in a nonlinear way to discover the regularity among the features. However, the deep neural network (DNN)-based VAD is far from fully investigated yet under an area of the statistical model-based VAD, which is a main topic of interest in this study.

Before presenting our work, it is worthwhile to mention the performance of the VAD by incorporating the acoustic environment classification technique since it is useful to build and use a different DNN scheme for various acoustic environments [10]-[12]. However, these methods are restricted in tracking the subtle changes in acoustics, which cause nonlinear change in the feature space since the acoustic features are extracted through the fixed filters such as a Mel-filter bank or linear prediction filter. In addition, they cannot represent the diversity of the feature yet in this acoustic environmental classification task since conventional algorithms are based on shallow methods. We note that the DNN can be successfully used in representing raw speech data to encapsulate the underlying information associated with various acoustic scenes.

In this paper, we develop the statistical model-based VAD by employing the DNN with a multiple layered deep architecture as a novel decision rule in classifying the input signal into speech or noise. The first step is to establish the baseline of the DNN by which the improved speech presence probability (SPP) is obtained based on the novel features of the statistical model-based VAD, namely the LR, the a priori signal-to-noise ratio (SNR), and the a posteriori SNR. As the key point that contributes to the success of DNN-based VAD, distinct DNNs according to different noise types are established via the separate training. Then different SPPs which correspond to a number of total noises are derived. As for the environmental awareness, an independent DNN module is constructed by a separate training process to offer the probabilities of occurrence for each noise type. The probability of occurrence for each noise environment is calculated to combine the SPPs, derived from the multiple DNNs and thus the final decision for VAD is obtained by comparing the combined SPP with a given threshold. The proposed VAD was evaluated in terms of an objective measure and found to have better results than the conventional SVM-based algorithm.
2. Review of Statistical Model-Based Voice Activity Detection

We briefly review the statistical model-based VAD. It is assumed that a noise signal \( d(t) \) is added to a speech signal \( x(t) \) in a time domain, with their sum being denoted as the noisy speech signal \( y(t) \). After taking a short-time Fourier transform (STFT) of the noisy speech \( y(t) \), we then have the following in the time-frequency domain as

\[
Y(k, n) = X(k, n) + D(k, n) \tag{1}
\]

where \( Y(k, n), X(k, n), \) and \( D(k, n) \) denote the STFT coefficients of the noisy speech signal, clean speech signal, and added noise signal, respectively, where \( k \) and \( n \) denote the frequency-bin index \( (k = 0, 1, \ldots, L - 1) \) and frame index \( (n = 0, 1, \ldots) \), respectively. Given two hypotheses, \( H_0 \) and \( H_1 \) which respectively indicate speech absence and presence, it is assumed that

\[
H_0 : Y(k, n) = D(k, n) \tag{2}
\]

\[
H_1 : Y(k, n) = X(k, n) + D(k, n) \tag{3}
\]

With the complex Gaussian probability distribution assumption, the probability density functions conditioned on the two hypotheses of \( H_0 \) and \( H_1 \) are given by

\[
p(Y(k, n)|H_0) = \frac{1}{\pi \lambda_d(k, n)} \exp \left\{ -\frac{|Y(k, n)|^2}{\lambda_d(k, n)} \right\} \tag{4}
\]

\[
p(Y(k, n)|H_1) = \frac{1}{\pi (\lambda_s(k, n) + \lambda_d(k, n))} \cdot \exp \left\{ -\frac{|Y(k, n)|^2}{\lambda_s(k, n) + \lambda_d(k, n)} \right\} \tag{5}
\]

where \( \lambda_s(k, n) \) and \( \lambda_d(k, n) \) denote the variances of the clean speech and of the noise signal for the individual frequency bin, respectively. The LR of the \( f \)th frequency-bin is derived as

\[
\Lambda(Y(k, n)) = \frac{p(Y(k, n)|H_1)}{p(Y(k, n)|H_0)} = \frac{1 + \xi(k, n)}{1 + \xi(k, n)} \cdot \exp \left\{ \frac{\gamma(k, n) \xi(k, n)}{1 + \xi(k, n)} \right\} \tag{6}
\]

where \( \xi(k, n) = \lambda_s(k, n)/\lambda_d(k, n) \) and \( \gamma(k, n) = |Y(k, n)|^2/\lambda_d(k, n) \) which are called the \textit{a priori} SNR and the \textit{a posteriori} SNR, respectively. The \textit{a posteriori} SNR \( \gamma(k, n) \) is estimated by \( \lambda_s(k, n) \), which is updated during the periods of speech absence and the \textit{a priori} SNR \( \xi(k, n) \) is estimated based on the well-known decision-directed approach [2] as follows:

\[
\xi(k, n) \equiv \alpha \frac{|\hat{X}(k, n) - 1|^2}{\lambda_d(k, n - 1)} + (1 - \alpha)P[\gamma(k, n) - 1] \tag{7}
\]

where \( \hat{X}(k, n - 1) \), the speech spectral amplitude estimate obtained in the previous frame, is obtained by a minimum mean square error estimator [1] and \( P[z] = \max(z, 0) \). Also, \( \alpha \) is a weight that is usually determined in the range \((0.95, 0.99)\) [1].

The final decision in the statistical model-based VAD has been accomplished by taking the geometric mean of the LR computed from the individual frequency-bins and it is obtained by

\[
\log \Lambda(n) = \frac{1}{L} \sum_{k=0}^{L-1} \log \Lambda(k, n) \gtrless \eta \tag{8}
\]

where \( L \) and \( \eta \) denote the total number of frequency-bins and a given threshold for speech detection, respectively. As in (8), an input frame is classified as the period of speech if the geometric mean of the LR is greater than the given threshold \( \eta \) and the period of noise otherwise.

3. Proposed Voice Activity Detection using Deep Neural Network

Since the shallow machine learning techniques such as SVM and geometric mean as in (8) are not desirable for the statistical model-based VAD as we mentioned earlier, the DNN, a powerful hierarchical generative model, is herewith applied to the decision function for the statistical model-based VAD. To fully take a consideration of nonlinear distribution of input features, the key parameters of the statistical model-based VAD such as the \textit{a priori} SNR, the \textit{a posteriori} SNR, and the LR are employed into the input layer of the DNN and go through the multiple hidden layers in the training and test steps. As we reviewed in Section 2, the microphone input signal is transformed to the STFT coefficients and then the \textit{a priori} SNR, the \textit{a posteriori} SNR, and the LR are estimated as in (6) and (7). The input feature vector \( Z \) is composed of the aforementioned three parameters and their delta and delta-delta components. Then, these are mapped into the target values, which are given in the training stage, through multiple nonlinear hidden layers as follows:

\[
f(Z(n)|\Theta) = g(\tilde{g}(g(Z(n)w^{(0)} + b^{(0)})w^{(1)} + b^{(1)})w^{(2)} + b^{(2)})w^{(3)} + b^{(3)} \tag{9}
\]

where \( Z(n) \) is the feature vector at the \( n \)th frame, \( \Theta = \{w^{(0)}, b^{(0)}, w^{(1)}, b^{(1)}, w^{(2)}, b^{(2)}, w^{(3)}, b^{(3)}\} \), and \( g(\cdot) \) denotes the activation function. And, \( w^{(j)} \) and \( b^{(j)} \) denote the weight term and bias term between the \( j \)th layer and the \((j+1)\)th layer. Note that all activation functions adopt the logistic function as given by

\[
g(z) = \frac{1}{1 + e^{-z}} \tag{10}
\]

One of the important issues in training the DNN is how to pretrain the weights and biases of the restricted Boltzmann machine (RBM). The solution adopted in this paper is to stack the multiple RBMs, which can be trained layer-by-layer in an unsupervised greedy fashion [6, 13]. After the weights and biases of each RBM are initialized by the layer-wise unsupervised learning procedure, they are fine-tuned by the gradient-based back-propagation algorithm to minimize the multi-class cross-entropy error function. The number of output units for the VAD is two, speech presence and absence and thus target output values of the speech frames are \([0 1]^T\) and those of the non-speech frames are \([1 0]^T\). The output units of each DNN for speech detection can be represented simply by the output of the DNN for VAD such that

\[
y_{out}(n) = f_0(Z(n)|\Theta) - f_1(Z(n)|\Theta) \tag{11}
\]

where \( f_0(Z(n)|\Theta) \) is the output value of the \textit{mth} output node and can be computed by

\[
f_m(Z(n)|\Theta) = g(\tilde{g}(g(Z(n)w^{(0)} + b^{(0)})w^{(1)} + b^{(1)})w^{(2)} + b^{(2)})w^{(3)} + b^{(3)} \tag{12}
\]

where \( \tilde{w}^{(3)} = \left[w_{b(1)m}, w_{b(1)m}, w_{b(1)m}, \ldots, w_{N_{\text{out}}-1,m}\right]^T \) which is the \textit{mth} column vector of \( \tilde{w}^{(3)} \), where \( N_{\text{out}} \) and \( T \) denote the number of nodes on the third hidden layer and the vector transposition, respectively. And, \( b_{m(3)}^{(3)} \) denotes the \textit{mth} element of \( b^{(3)} \). The SPP can be derived from the output of the DNN by adopting a parametric way to employ the sigmoid function as given by

\[
p(H(n) = H_1|y_{out}(n)) = \frac{1}{1 + \exp(A \cdot y_{out}(n) + B)} \tag{13}
\]
Here, the principal parameters such as the slope parameter $A$ and the bias parameter $B$ can be obtained by discriminative training in a way to minimize the negative log likelihood of the data, which is the cross-entropy error function \[14\]. The model-trust algorithm based on the Levenberg-Marquardt algorithm \[15\] is performed to obtain the optimal parameters $A$ and $B$, which minimize the cross-entropy error function.

First, the ensemble of DNNs used for VAD can be built through the layer wise pre-training process by using all type of the noisy speech dataset without a label, namely, the unsupervised learning. Then, each DNN is optimized, separately, through the fine-tuning process by using each type of the noisy training data with the label and by starting with the initialized weights and biases. As a result, the optimal parameters for estimating the SPP from the output of the DNNs are obtained by the model-trust algorithm \[15\]. In the online classification stage, each DNN is optimized to the parameter for the $i$th type of noisy training data where $N_{DNN}$ denotes the number of DNNs on the number of the acoustic environments.

For the realtime implementation, each type of noise should be classified during the period of non-speech. This process is performed by applying the DNN in a similar reason described earlier. Indeed, the DNN for acoustic environment classification is implemented by incorporating the log-power spectrum as the feature vector. For this, the acoustic signal data manually labeled as noise frames was used, so optimal weights and biases set $\Theta_{AEC}$ were found at the training stage. Generally, each frame is classified from the value of output nodes of the DNN into the classes by taking the soft-max and winner-take-all methods. Instead of the hard decision, we adopt the parametric way to employ the sigmoid function, which was also used to predict the probabilities for each type of noise as follows:

\[
\tilde{y}_i(n) = f_i(\log(Y(n)))|\Theta_{AEC} - \sum_{j=1}^{i-1} f_j(\log(Y(n)))|\Theta_{AEC} - \sum_{j=i+1}^{N_{DNN}} f_j(\log(Y(n)))|\Theta_{AEC}
\]

where $\tilde{y}_i(n)$ is the final output value of the DNN for the $i$th type of noise condition at the $n$th frame and $Y(n) = [Y(0, n), Y(1, n), \ldots, Y(L - 1, n)]$, which is a vector of power spectra at the $n$th frame. The probability of the $i$th type of noise given the observation $Y(n)$ is derived by

\[
p(s(n) = s_i|Y(n)) = \frac{1}{1 + \exp(A_{AEC}^{(i)} \cdot \tilde{y}_i(n) + B_{AEC}^{(i)})}
\]

where $A_{AEC}^{(i)}$ and $B_{AEC}^{(i)}$ denote the slope parameter and the bias parameter for the $i$th type of noises, respectively, and can be optimized to the $i$th type of noise environment based on the model-trust algorithm \[15\]. In the online classification stage, SPPs can be computed from each DNN by transferring the statistical feature vector into each well-trained DNN and then they are derived from the simplified output of each DNN by the parametric sigmoid function with specific optimal parameters $A_{VAD}^{(i)}$ and $B_{VAD}^{(i)}$. A final SPP can be obtained by the linear weighted combination way with the weights derived from the result of the acoustic environment classification. Note that the weights must satisfy the following conditions

\[
\sum_{i=1}^{N_{noise}} w_i(n) = 1
\]

\[
w_i(n) \geq 0.
\]

The weights satisfying the above conditions can be derived from the probabilities for each type of noise as follows:

\[
w_i(n) = \frac{\exp(p(s(n) = s_i|\log(Y(n))))}{\sum_{j=1}^{N_{noise}} \exp(p(s(n) = s_j|\log(Y(n))))}.
\]

The final SPP is estimated by fusing the probability of each DNN with the weights for each noise as follows:

\[
p(H(n) = H_1|Z(n)) = \sum_{i=1}^{N_{noise}} w_i(n) \cdot p_i(H(n) = H_1|y_{out}(n))
\]

where $p_i(H(n) = H_1|y_{out}(n))$ is calculated by using the parameters $\Theta_i$, $A_{VAD}^{(i)}$, and $B_{VAD}^{(i)}$ used to estimate the SPP at $i$th DNN for VAD and $A_{AEC}^{(i)}$ and $B_{AEC}^{(i)}$ used to calculate the probability of the $i$th type of noise. In order to install the weights adaptively, weights have to be updated during the period of noise. Here, the threshold for the criterion of speech absence was set to 0.35 for updating the weights. This scheme is revised by the hang-over scheme method \[2\] when the final SPP of the previous frame is less than the given threshold as given by

\[
w_i(n) = a \cdot w_i(n-1) + (1-a) \cdot \frac{\exp(p(s(n) = s_i|\log(Y(n))))}{\sum_{j=1}^{N_{noise}} \exp(p(s(n) = s_j|\log(Y(n))))}
\]

where $a$ is a smoothing factor. Note that the acoustic environment classification operates on the first few frames since those can be considered as the non-speech frames, which can be an acceptable rule \[5\]. Finally, each frame is initially classified as a the period of speech if the SPP as in \(19\) is greater than a given threshold $\eta$ for speech detection or the period of noise otherwise.

### 4. Experiments and results

This section describes the performance evaluation of the proposed VAD approaches based on the statistical model and ensemble of DNNs employing acoustic environment classification. For the objective comparison, the proposed approach was compared with the conventional SVM-based VAD \[5\] and the method based on single DNN without the acoustic environment classification. Each VAD algorithm was assessed by investigating the probability of error $P_F$, the probability of miss $P_M$, and the probability of false-alarm $P_Y$ which are defined as the ratio of incorrect decision to the overall frames, as the ratio of miss speech decision to the hand-marked speech frames, and as the ratio of false speech decisions to the hand marked non-speech frames, respectively. In this work, the input signal was sampled at 8 kHz, and the analysis window size was 10 ms, which implies that each frame consists of 80 samples.

To train the DNNs, we made reference decisions on clean speech corpus of 304 s long by manually labeling the active and inactive regions of the speech signal for every 10 ms frame. The proportions of voiced speech, unvoiced speech, and silence frames of the training materials were 45.29%, 13.37%, and 41.34%, respectively. In order to make a input noisy speech...
data, we added the babble, car, and street noises, which stem from the Aurora 2 database, to clean speech signals while keeping the SNR at -5, 0, 5, and 10 dB. As a result, about 1.2 hours of labeled noisy training data was prepared. We also built the other unlabeled noisy speech corpus of about 116.6 hours by adding the aforementioned noise signals to the Aurora 2 and TIMIT databases while keeping the same noise conditions. In training stage, about 117.8 hours of labeled and unlabeled noisy training data were used in the pre-training process and 1.2 hours of labeled noisy training data was used only in the fine-tuning process.

To test the proposed approach, we used different speech data of 256 s in duration, and manually labeled as the active and inactive regions of the speech signal for every 10 ms frame. The proportions of voiced speech, unvoiced speech, and silence frames of the materials for testing were 43.86%, 13.40%, and 42.74%, respectively. To simulate noisy conditions, aforementioned noises were added to clean speech signals at -5, 0, 5, and 10 dB SNR. The office noise from NOISEX-92 [16] was also added to clean speech while keeping the same noise conditions to evaluate the proposed VAD approach under the unseen noise conditions.

The parameter settings used to train DNNs for VAD and acoustic environmental classification are explained in the following. For training each DNN for the VAD, the number of hidden layers was set to three and the numbers of units on each hidden layers were set to 256, 128, and 64, respectively. To train the DNN for the acoustic environmental classification, the number of hidden layers was also set to three and the numbers of units on hidden layers were set to same as 512. The minibatch sizes for the pre-training and fine-tuning were set to 100 and 1000, respectively. In the unsupervised pre-training process, the maximum epoch was set to 80 and the learning rate was set to 0.001. For the fine-tuning process, the maximum epoch was set to 150 and the learning rate was set to 0.1 for the first 10 epochs and then the learning rate was decreased by 10% after 10 epochs at which we did not consider the early stopping scheme. The parameters used to train the SVM compared with the proposed approach were set rigorously according to the author’s settings that the radial basis function is used as the kernel function and the kernel parameter was set to 1.0 [5].

Table 1: Performance comparison of $P_L$, $P_M$, $P_F$ among conventional algorithms and proposed approach

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise SNR(dB)</td>
<td>$P_L$</td>
<td>$P_M$</td>
<td>$P_F$</td>
</tr>
<tr>
<td>babble</td>
<td>-5</td>
<td>15.53</td>
<td>20.74</td>
</tr>
<tr>
<td>0</td>
<td>25.38</td>
<td>20.94</td>
<td>23.21</td>
</tr>
<tr>
<td>5</td>
<td>21.16</td>
<td>24.82</td>
<td>17.82</td>
</tr>
<tr>
<td>10</td>
<td>16.45</td>
<td>18.75</td>
<td>14.73</td>
</tr>
<tr>
<td>car</td>
<td>-5</td>
<td>21.62</td>
<td>23.37</td>
</tr>
<tr>
<td>0</td>
<td>26.35</td>
<td>24.91</td>
<td>10.52</td>
</tr>
<tr>
<td>10</td>
<td>10.35</td>
<td>13.75</td>
<td>7.82</td>
</tr>
<tr>
<td>street</td>
<td>-5</td>
<td>17.00</td>
<td>15.65</td>
</tr>
<tr>
<td>0</td>
<td>22.97</td>
<td>30.00</td>
<td>17.06</td>
</tr>
<tr>
<td>5</td>
<td>17.45</td>
<td>23.73</td>
<td>14.47</td>
</tr>
<tr>
<td>office</td>
<td>-5</td>
<td>27.00</td>
<td>28.37</td>
</tr>
<tr>
<td>0</td>
<td>43.71</td>
<td>52.67</td>
<td>13.55</td>
</tr>
<tr>
<td>5</td>
<td>16.15</td>
<td>28.22</td>
<td>8.93</td>
</tr>
<tr>
<td>10</td>
<td>15.00</td>
<td>26.78</td>
<td>6.24</td>
</tr>
</tbody>
</table>

% and the average accuracy of 99.33 %.

5. Conclusions

In this paper, we have presented the novel VAD technique provided by the ensemble of DNNs with acoustic environment classification. The first contribution of this work is the use on the statistical model parameters such as the a priori SNR, the a posteriori SNR, and the LR in the feature extraction part and then they map to the target output value in a nonlinear way for transferring them into multiple DNNs, which were trained for each kind of noise conditions. In order to implement the frame-by-frame pooling of the ensemble of DNN, the SPPs are first derived from the output of DNNs by using parametric sigmoid function with optimized parameters. Then, final SPP was computed by fusing the estimate of each DNN in a linearly weighted combination way by the use of weights derived from the result of DNN-based acoustic environment classification. The proposed method was evaluated in terms of an objective measure and was found to have significant improvement compared to the previous method under the various noise conditions in terms of types and SNR levels.

6. Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. 2014R1A2A1A10049735). This research was also supported by NRF (14035195).

7. References


