A Universal VAD Based on Jointly Trained Deep Neural Networks

Qing Wang[1], *Jun Du*[1], *Xiao Bao*[1], *Zi-Rui Wang*[1], *Li-Rong Dai*[1], *Chin-Hui Lee*[2]

[1]University of Science and Technology of China, P. R. China
[2]Georgia Institute of Technology, USA

{xiaosong,baox,cs211}@mail.ustc.edu.cn, {jundu,lrdai}@ustc.edu.cn, chl@ece.gatech.edu

Abstract

In this paper, we propose a joint training approach to voice activity detection (VAD) to address the issue of performance degradation due to unseen noise conditions. Two key techniques are integrated into this deep neural network (DNN) based VAD framework. First, a regression DNN is trained to map the noisy to clean speech features similar to DNN-based speech enhancement. Second, the VAD part to discriminate speech against noise backgrounds is also a DNN trained with a large amount of diversified noisy data synthesized by a wide range of additive noise types. By stacking the classification DNN on top of the enhancement DNN, this integrated DNN can be jointly trained to perform VAD. The feature mapping DNN serves as a noise normalization module aiming at explicitly generating the "clean" features which are easier to be correctly recognized by the following classification DNN. Our experiment results demonstrate the proposed noise-universal DNN-based VAD algorithm achieves a good generalization capacity to unseen noises, and the jointly trained DNNs consistently and significantly outperform the conventional classification-based DNN for all the noise types and signal-to-noise levels tested.

Index Terms: voice activity detection, deep neural network, feature mapping, joint training

1. Introduction

Voice activity detection (VAD) is a very fundamental preprocessing module for many speech applications, such as speech coding, speech recognition, speaker recognition, and spoken language identification. In the mobile internet era, most speech-activated devices use a push-to-talk function as a manual VAD mechanism to record speech, implying that high-performance VAD is still an unsolved problem in real-world scenarios, especially in non-stationary or low signal-to-noise ratio (SNR) environments. Recent VAD research could be traced back to the late 1950s [1]. For the past several decades, many approaches were investigated and they could be categorized into three broad classes. The first class focused on the study of different acoustic features or metrics, e.g., linear prediction coding (LPC) parameters [2], zero-crossing rate (ZCR) [3], periodicity measure [4], cepstral features [5], formant shape [6], the higher-order statistics of the LPC residual [7], the long-term spectral divergence (LTSD) [8], and fusion of multiple features [9]. The second class was the statistical model based VAD algorithm originated from Ephraim & Malah's work for speech enhancement [10]. In [11], a Gaussian model was adopted for VAD with a decision-directed approach [12] to estimate the signal parameters. It achieved a better VAD performance over the conventional approaches. Later, the statistical model based approaches were improved by using soft decision schemes [13, 14], or other model assumptions, e.g., replacing the Gaussian by the Gamma and Laplacian distributions [15, 16]. The third class, often referred to as the so-called supervised learning approach, directly utilized classification models to discriminate speech against noise, instead of making model assumptions about the interaction between the speech and noise signals. Classifier designs, such as support vector machine (SVM) [17], conditional random field (CRF) [18], and non-negative sparse coding [19], have been investigated.

Recently, the deep learning techniques [20, 21] have been increasingly popular for many speech areas, e.g., speech recognition [22], speech enhancement [23, 24] and separation [25]. Several representative work for VAD were based on deep neural networks [26, 27, 28] and recurrent neural networks [29]. The deep learning approaches indeed could significantly improve the VAD performance compared with other classification models under the matched noise conditions. But the generalization capability problem to unseen noise conditions was not explicitly discussed and addressed in previous work. Inspired by the recent success to handle the unseen noises in speech enhancement [24], in this work first we propose a universal VAD based on deep neural network by using a large amount of diversified noisy data synthesized by a wide range of additive noises. But our preliminary experiments show that the classification DNN for VAD with only two-dimensional output can not handle the diversified noisy training data well and the performance of DNN is quickly saturated when using more than two hidden layers. Motivated by the recent work for noise robust speech recognition [30, 31, 32], we present a novel feature mapping front-end by using a regression DNN as a noise normalization module to estimate the clean speech features which make the VAD decision easier with the subsequent classification DNN. Furthermore, the feature mapping DNN can be jointly trained with the conventional classification DNN, namely the joint training of the front-end and back-end DNNs for VAD. Our experiments demonstrate the superiority of the jointly trained DNN for all unseen noise types and levels.

2. DNN Based VAD System Overview

The overall flowchart of VAD system is illustrated in Fig. 1. In the training stage, first the acoustic features of both clean speech and synthesized noisy speech training data are extracted. Multi-resolution cochleagram (MRCG) features are adopted, which are well verified for speech recognition [33] and VAD [28]. Then two DNNs, namely feature mapping DNN and the classification DNN, are trained. Please note that the stereo-data of clean speech and noisy speech MRCG features should be adopted to train the feature mapping DNN while only the noisy speech features are needed for the conventional classification DNN training. Finally a generic DNN can be generated by joint training of both feature mapping and classification DNN. In the

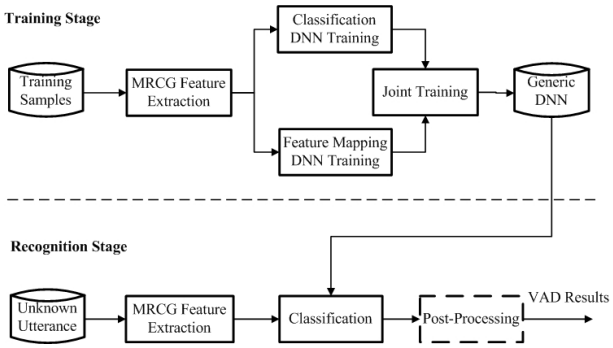

Figure 1: VAD system flowchart.

recognition stage, after the feature extraction of the unknown utterance, frame-level decision is first given by the generic DNN. To achieve better performance, a post-processing can be applied via a long-term smoothing of the multiple DNN outputs with a half-window size τ. The classification DNN with post-processing is quite similar to the boosted DNN proposed in [28]. The main difference is the acoustic context information, namely the neighboring frames is directly integrated into the output layer of boosted DNN. The details of both regression and classification DNNs are elaborated in Section 3.

3. Jointly Trained DNNs

3.1. Conventional DNN Training of VAD

The conventional DNN for VAD is designed as a classification DNN where the output refers to the probabilities of two classes. The input of DNN is the noisy MRCG features with neighboring frames. The training of this DNN consists of unsupervised pre-training and supervised fine-tuning. The pre-training treats each consecutive pair of layers as a restricted Boltzmann machine (RBM) while the parameters of RBM are trained layer by layer with the approximate contrastive divergence algorithm [21]. After pre-training for initializing the weights of the first several layers, a supervised fine-tuning of the parameters in the whole neural network is performed via a frame-level cross-entropy criterion. The main difference from other DNN approaches, e.g. [28], is the training data. In [28], only three noise types are used for training with a small amount of utterances and the noise types of the test set are the same as those of the training set. In this work, to designed a universal VAD robust to any noise environments, a large training set is formed by synthesizing the noisy speech data with a wide range of additive noises at different SNRs. And only the testing on unseen noises is conducted.

3.2. Feature Mapping

The viewpoint of the conventional classification DNN training is that the noise with other irrelevant variabilities might be implicitly normalized during the fine-tuning procedure. However, a single DNN even with deep architectures can not simultaneously perform irrelevant variabilities normalization and the content classification well, which is verified for noise robust speech recognition [31]. This observation should be well applied to our case due to the diversity of the training data consisting of many combinations with different noise types and levels. Furthermore our preliminary experiments show that the performance of classification DNN for VAD with only two-dimensional output is easily saturated when using more than two hidden layers. To address these problems, we propose a novel feature mapping DNN as an explicit noise normalization module. The DNN architecture for feature mapping is shown in Fig. 2. This DNN acts as a highly non-linear regression function to map the noisy speech features to clean speech features. As for the training data, the pairs of noisy and clean speech data should be used, unlike in classification DNN, only the noisy speech data is needed. The training procedure of this regression DNN is similar to classification DNN, namely the RBM pre-training plus a supervised fine-tuning. The main difference of regression DNN training from classification DNN training is the objective function. We aim at minimizing mean squared error (MMSE) between the DNN output and the reference clean speech features:

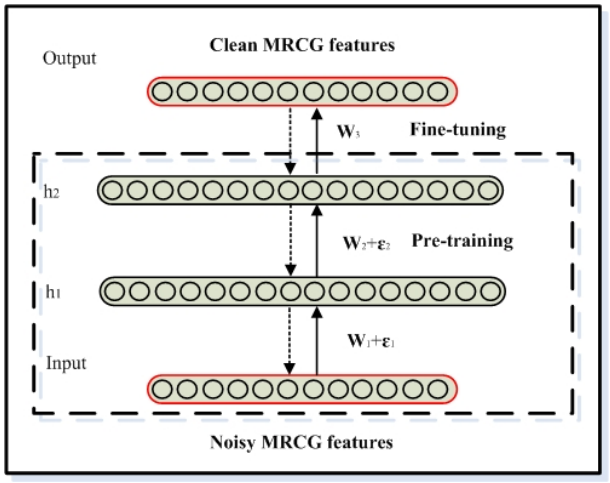

Figure 2: DNN for feature mapping.

$$E = \frac{1}{N}\sum_{n=1}^{N} \|\hat{\boldsymbol{x}}_{n-\tau}^{n+\tau}(\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W}, \boldsymbol{b}) - \boldsymbol{x}_{n-\tau}^{n+\tau}\|_2^2 + \kappa\|\boldsymbol{W}\|_2^2 \quad (1)$$

where $\hat{\boldsymbol{x}}_{n-\tau}^{n+\tau}$ and $\boldsymbol{x}_{n-\tau}^{n+\tau}$ are the $D(2\tau+1)$-dimensional vectors of estimated and reference clean MRCG features for the n^{th} frame, respectively. $\boldsymbol{y}_{n-\tau}^{n+\tau}$ is a $D(2\tau+1)$-dimensional vector of input noisy MRCG features with the neighbouring left and right τ frames as the acoustic context. $\boldsymbol{W}$ and $\boldsymbol{b}$ denote all the weight and bias parameters. κ is the regularization weighting coefficient to avoid over-fitting. The objective function is optimized using back-propagation with a stochastic gradient descent method in mini-batch mode of N sample frames. Please note that the acoustic context is also used in the DNN output which is similar to boosted DNN in [28]. All the input and output features are normalized with a global mean and variance of the noisy MRCG features of the training set.

3.3. Joint Training

The joint training procedure of two DNNs can be divided into two steps. The first step is to convert the classification DNN with the input of noisy MRCG features to the DNN with the input of estimated clean MRCG features, which is implemented via a simple fine-tuning of the original noisy DNN by only changing the input to the estimated clean MRCG features rather than the noisy MRCG features. After this step, it is very interesting to observe that the output of the feature mapping DNN

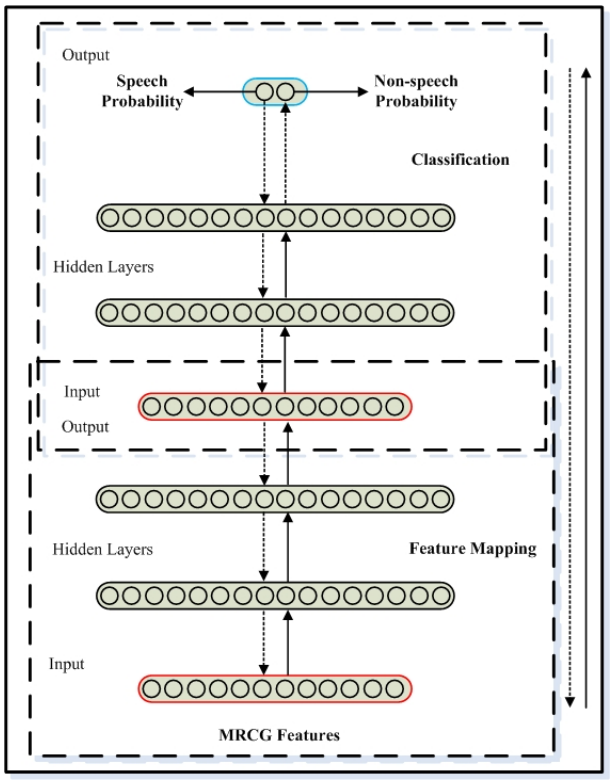

Figure 3: Jointly trained DNNs.

is exactly the same as the input of the newly updated classification DNN. So the second step is to concatenate two DNNs to a single generic DNN, which can be illustrated as in Fig. 3. We directly stack the classification layers on top of the feature mapping layers. The output layer of feature mapping and the input layer of classification is merged as one hidden layer in the generic DNN (or denoted as JT-DNN). It is noted that this is a hidden layer with a linear activation function while others are with sigmoid activation functions. Using the same object function as the classification DNN, all weight and bias parameters are then re-trained. After joint training, the generic DNN yields a better performance than two separated DNNs which can be explained as the feature mapping network is refined to enable a better classification performance rather than optimizing the original MMSE criterion.

4. Experiments and Result Analysis

4.1. Experimental Setup

Our experiments were conducted on the 16kHz clean utterances of the Aurora4 database [34]. As for the training data, all the 7138 utterances of the clean training set were used to synthesize the noisy speech with 115 noise types, including the public 100 noise types in [35], and 15 home-made noise types [1]. Each synthesized noisy utterance is obtained by corrupting each clean utterance with one of 115 noise types at one of six noise levels of SNR, i.e., 20dB, 15dB, 10dB, 5dB, 0dB, and -5dB. For the test data, 40 clean utterances randomly selected from the 330 test utterances of Aurora4 were corrupted by three unseen noise types from the NOISEX-92 corpus [36], namely Babble, Factory, and Machinegun, at three SNR levels: 5dB, 0dB, and -5dB. The frame-level reference labels of each noisy utterance were generated by forced alignment on the corresponding clean utterance using the acoustic model trained on clean speech data. 768-dimensional MRCG features were extracted according to [28]. The area-under-ROC-curve (AUC) [37] was adopted as the evaluation metric. For both the regression DNN and classification DNN, sigmoid activation function was used and the number of units in each hidden layer was set to 2048 by default. The mini-batch size N was set to 128. The regularization weighting coefficient κ in Eq.(1) was 0.8. The other tuning parameters of DNN were set according to [38]. The half-window size τ for post-processing was 19.

[1]The 115 noise types are N1-N17: Crowd noise; N18-N29: Machine noise; N30-N43: Alarm and siren; N44-N46: Traffic and car noise; N47-N55: Animal sound; N56-N69: Water sound; N70-N78: Wind; N79-N82: Bell; N83-N85: Cough; N86: Clap; N87: Snore; N88: Click; N88-N90: Laugh; N91-N92: Yawn; N93: Cry; N94: Shower; N95: Tooth brushing; N96-N97: Footsteps; N98: Door moving; N99-N100: Phone dialing; N101: AWGN; N102: Babble; N103-N105: Car; N106-N115: Musical instruments.

4.2. Comparison between DNN and JT-DNN

Table 1 gives a performance comparison of different DNN based VAD approaches for the three unseen noise environments with different SNRs averaged on the test set. For the conventional classification DNN (denoted as DNN), two configurations of 2 and 3 hidden layers were compared. It seemed that increasing the number of hidden layers for the conventional DNN could not guarantee to yield consistent performance gain for unseen noise types. Actually we could only observe the improvements for Factory noise. This might be due to the different characteristics of unseen noises or the weird architecture of DNN with only two-dimensional output which could not handle the diversified training data well.

Two configurations of 2+1 (2 hidden layers for feature mapping DNN and 1 hidden layer for classification DNN) and 2+2 were designed for experiments of jointly trained DNNs (denoted as JT-DNN). It was obvious that JT-DNN achieved consistent and significant improvements of AUC performance for all the unseen noises with different SNRs, especially at low SNRs, e.g., AUC improved from 82.26% (DNN with 2 hidden layers) to 89.76% (JT-DNN with 2+2 configuration) at -5dB., which demonstrated the importance of the feature mapping DNN. Even the worst results of JT-DNN were still much better than the best results of DNN. For JT-DNN we observed the mixed results between 2+1 and 2+2, e.g., the better 2+1 results for Babble noise while much better 2+2 results for machine gun noise.

Overall, JT-DNN improved the generalization capability of DNN, which could be explained as that adding one hidden layer directly in the conventional DNN can easily lead to over-fitting (DNN with 2 hidden layers and 3 hidden layers) while adding the feature mapping layers to the conventional classification DNN could yield significant performance gain (DNN with 2 hidden layers and JT-DNN with 2+2 configuration).

The DNN and JT-DNN with the post-processing were denoted as DNN-PP and JT-DNN-PP, respectively. After post-processing, all the AUC results were improved as long-term information was used. And almost all the above observations without post-processing could be applied to the post-processing versions. It was interesting that at low SNRs, e.g., -5dB, the improvements from DNN with 2 hidden layers to JT-DNN with 2+2 are more significant for the post-processing case, e.g., for the Babble noise case, 86.48% to 89.22% with no post-

Table 1: *Performance (AUC in %) comparison of different DNN based VAD approaches for the three unseen noise environments with different SNRs averaged on the test set.*

Noise Type	SNR	DNN		JT-DNN		DNN-PP		JT-DNN-PP	
		2	3	2+1	2+2	2	3	2+1	2+2
Babble	5dB	98.73	98.65	99.19	99.08	98.84	98.8	99.25	99.21
	0dB	95.44	95.11	96.99	96.54	96.04	96.13	97.45	97.48
	-5dB	86.48	85.75	90.19	89.22	88.98	88.94	92.26	92.29
Factory	5dB	98.28	98.47	99.09	99.01	99.12	99.18	99.5	99.42
	0dB	93.77	94.58	96.91	97.02	96.26	97.14	98.29	98.59
	-5dB	82.26	83.76	89.19	89.76	86.66	88.7	92.73	94.46
Machine Gun	5dB	95.95	95.82	96.5	97.99	96.79	96.57	96.92	98.02
	0dB	91.43	89.87	91.55	94.93	94.46	93.73	94.68	96.63
	-5dB	85.53	82.87	86.54	90.57	90.7	89.39	91.65	94.39

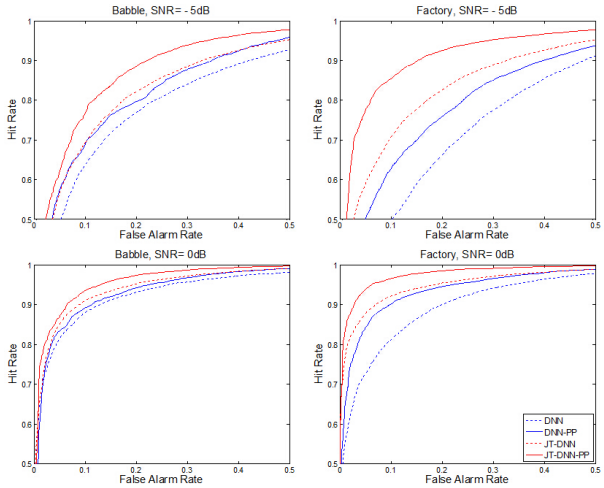

Figure 4: ROC curves for DNN with 2 hidden layers and JT-DNN with 2+2 configuration with/without post-processing for two unseen noise environments at two SNRs of 0dB and -5dB.

processing while 88.98% to 92.29% with post-processing. This indicated that with long-term information, JT-DNN could make more correct VAD decision than DNN at very low SNRs.

Fig. 4 shows the ROC curve analysis for DNN with 2 hidden layers and JT-DNN with 2+2 configuration with/without post-processing for two unseen noise environments at different SNRs of 0dB and -5dB. The similar observations on ROC curve could be made as the AUC metric. The gap of ROC curves between DNN and JT-DNN became larger at lower SNR for the same noise type, which demonstrated JT-DNN was more effective under low SNRs. Furthermore, the area between two solid line is larger than the area between two dotted line at -5dB, which implied that with post-processing JT-DNN could perform better than DNN at low SNRs.

4.3. Comparison among Different Sizes of JT-DNN

Table 2 lists a performance comparison of different number of units in the hidden layers of JT-DNN with 2+2 configuration for two unseen noise environments with different SNRs averaged on the test set. Three configurations, namely 2048, 1024, and 512 hidden nodes, were compared. First, the decreasing of hidden nodes could lead to the degradation of AUC performance for all the unseen noises at different SNRs. But even for JT-DNN with 1024 hidden nodes which had less parameters than the DNN with 2 hidden layers and 2048 hidden units for each layer in Table 1, its performances with/without post-processing were still consistently better for all noise types and levels, especially at low SNRs, e.g., AUC improved from 82.26% to 88.68% for Factory noise at -5dB. This implied that the feature mapping module of JT-DNN could make the generic DNN more compact and effective.

Table 2: *AUC (in %) comparison of different hidden layer sizes of JT-DNN with 2+2 configuration for two unseen noise environments with different SNRs averaged on the test set.*

Babble						
	JT-DNN			JT-DNN-PP		
SNR	512	1024	2048	512	1024	2048
5dB	98.71	98.92	99.08	98.94	99.07	99.21
0dB	95.31	96.08	96.54	96.78	97.04	97.48
-5dB	86.62	87.88	89.22	90.41	91.06	92.29
Factory						
	JT-DNN			JT-DNN-PP		
SNR	512	1024	2048	512	1024	2048
5dB	98.59	98.97	99.01	99.29	99.38	99.42
0dB	95.89	96.63	97.02	98.07	98.22	98.59
-5dB	87.69	88.68	89.76	93.15	93.12	94.46

5. Conclusions

In this paper, we have presented a universal DNN for VAD robust to any noise types. To address the diversity of the training data with a wide range of noise types and levels, a novel feature mapping DNN is built to estimate the clean acoustic features from the noisy acoustic features which can make the subsequent VAD decision easier. Joint training of feature mapping DNN and classification DNN can yield very promising VAD results compared with the conventional DNN training. As for the future work, we will focus on improving the practicability of our approach in both accuracy and efficiency.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants No. 61305002. We would like to thank Dr. Xiao-Lei Zhang for the help in providing the tool of MRCG feature extraction.

7. References

[1] K. Bullington and J. M. Fraser, "Engineering aspects of TASI," *Bell Syst. Tech. J.*, pp. 353-364, 1959.

[2] L. R. Rabiner and M. R.Sambur , "Voiced-unvoiced-silence detection using Itakura LPC distance measure," in *ICASSP*, 1977, pp. 323-326.

[3] J. C. Junqua, B. Reaves, and B. Mark, "A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize," in *EUROSPEECH*, 1991, pp. 1371-1374.

[4] R. Tucker, "Voice activity detection using a periodicity measure," *Proc. Inst. Electr. Eng.*, vol. 139, pp. 377-380. 1992.

[5] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral features," in *IEEE TELCON*, 1993, pp.321-324.

[6] J. D. Hoyt and H. Wechsler, "Detection of human speech in structured noise," in *ICASSP*, 1994, pp. 237-240.

[7] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. on Speech and Audio Process.*, vol. 9, no. 3, pp. 217-231, 2001.

[8] J. Ramírez, J. C. Segura, C. Benítez, Á. de la Torre, A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3, pp.271-287, 2004.

[9] S. G. Tanyer and H. Özer, "Voice activity detection in nonstationary noise," *IEEE Trans. on Speech and Audio Process.*, vol. 8, no. 4, pp. 478-482, 2000.

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109-1121, 1984.

[11] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *ICASSP*, pp. 365-368, 1998.

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

[13] N. S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Process. Lett.*, vol. 7, pp. 108-110, 2000.

[14] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model based voice activity detector," *IEEE Signal Process. Lett.*, vol. 8, pp. 276-278, Oct. 2001.

[15] S. Gazor and W. Zhang, "Speech probability distribution, " *IEEE Signal Process. Lett.*, vol. 10, pp. 204-207, 2003.

[16] T. Petsatodis, C. Boukis, F. Talantzis, Z. Tan, and R. Prasad, "Convex combination of multiple statistical models with application to vad," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2314-2327, 2011.

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

[18] A. Saito, Y. Nankaku, A. Lee, and K. Tokuda, "Voice activity detection based on conditional random fields using multiple features," in *INTERSPEECH*, 2010, pp. 2086-2089.

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

[20] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504-507, 2006.

[21] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527-1554, 2006.

[22] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, 2012.

[23] Y. X. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381-1390, 2013.

[24] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65-68, 2014.

[25] Y.-H. Tu, J. Du, Y. Xu, L.-R. Dai, and C.-H. Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," Submitted to Proc. ISCSLP, 2014.

[26] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. on Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 697-710, 2013.

[27] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on youtube using deep neural networks, in *INTERSPEECH*, 2013, pp. 728-731.

[28] X.-L. Zhang and D. L. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *INTERSPEECH*, 2014, pp. 1534-1538.

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

[30] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, and C.-H. Lee, "Robust speech recognition with speech enhanced deep neural networks," in *INTERSPEECH*, 2014, pp. 616-620.

[31] 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 *ICASSP*, 2015.

[32] A. Narayanan and D. L. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *ICASSP*, 2014,

[33] J. Chen, Y. Wang, and D. L. Wang, "A feature study for classification-based speech separation at very low signal-to-noise ratio," in *ICASSP*, 2014, pp. 7089-7093.

[34] N. Parihar and J. Picone, "DSR front end LVCSR evaluation," *Aurora Working Group*, 2002.

[35] G. Hu, "100 nonspeech environmental sounds," 2004. http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html.

[36] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247-251, 1993.

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

[38] G. Hinton. "A practical guide to training restricted Boltzmann machines," *Technical Report*, University of Toronto, 2010.