



Boosted Deep Neural Networks and Multi-resolution Cochleagram Features for Voice Activity Detection

Xiao-Lei Zhang¹, DeLiang Wang²

¹TNList, Department of Electronic Engineering, Tsinghua University, Beijing, China

²Department of Computer Science & Engineering and Center for Cognitive & Brain Sciences, The Ohio State University, Columbus, OH, USA

huoshan6@126.com, dwang@cse.ohio-state.edu

Abstract

Voice activity detection (VAD) is an important frontend of many speech processing systems. In this paper, we describe a new VAD algorithm based on boosted deep neural networks (bDNNs). The proposed algorithm first generates multiple base predictions for a single frame from only one DNN and then aggregates the base predictions for a better prediction of the frame. Moreover, we employ a new acoustic feature, multi-resolution cochleagram (MRCG), that concatenates the cochleagram features at multiple spectrotemporal resolutions and shows superior speech separation results over many acoustic features. Experimental results show that bDNN-based VAD with the MRCG feature outperforms state-of-the-art VADs by a considerable margin.

Index Terms: Boosting, cochleagram, deep neural network, MRCG, voice activity detection

1. Introduction

Voice activity detection (VAD) is an important preprocessor of many speech systems, such as speech communication and speech recognition [1]. Perhaps the most challenging problem of VAD is to make it perform in low signal-to-noise ratio (SNR) environments. Early research focused on acoustic features, including energy in the time domain, pitch detection, zero-crossing rate, and many spectral energy based features [2]. Later on, effort shifted to statistical signal processing. These techniques first make assumptions on the distributions of speech and background noise (usually in the spectral domain) respectively, and then design statistical algorithms to dynamically estimate the model parameters, making them flexible in dealing with nonstationary noises. Typical models include the Gaussian distribution [3], Laplace distribution, Gamma distribution, or their combinations [4]. But statistical model based methods have limitations. First, model assumptions may not fully capture data distributions since the models usually have too few parameters. Second, with relatively few parameters, they may not be flexible enough in fusing multiple acoustic features. Third, they estimate parameters from limited observations, which may not fully utilize rich information embodied in speech corpora.

Recently, supervised learning methods are becoming more popular, as they have the potential to overcome the limitations of statistical model based methods. Typical models for VAD include support vector machine [5], conditional random field [6], sparse coding [7], spectral clustering [8], Gaussian models

[9–12], Gaussian mixture model [8], recursive neural network [13], and deep neural network (DNN) [14, 15].

In this paper, we investigate supervised learning for VAD at low SNRs. The main contributions of this paper are summarized as follows: (i) We propose a new deep model for VAD, named boosted deep neural network (bDNN). (ii) We employ a new acoustic feature for VAD, named multi-resolution cochleagram (MRCG) [16]. (iii) The boosting idea in bDNN and the multi-resolution scheme in MRCG, we believe, can be applied to other speech processing tasks, such as speech separation and speech recognition. Empirical results on the AURORA4 corpus [17] show that the bDNN-based VAD with the MRCG feature outperforms 5 comparison methods by a considerable margin, including the supervised DNN-based VAD [14].

2. Boosted DNN

In this section, we present the bDNN algorithm for the VAD problem. bDNN was motivated by ensemble learning, an important branch of machine learning [18]. Ensemble learning learns a strong classifier by grouping the predictions of multiple weak classifiers. The key idea behind bDNN is to generate multiple different base predictions for a single frame, so that when the base predictions are aggregated, the final prediction is boosted to be better than any of the base predictions. It contains two phases—training and test.

Training Phase. Suppose we have a manually-labeled training speech corpus that consists of V utterances, denoted as $\mathcal{X} \times \mathcal{Y} = \{(\mathbf{x}_k, y_k)\}_{k=1}^{K_v}\}_{v=1}^V$, where K_v is number of frames of the v th utterance, $\mathbf{x}_k \in \mathbb{R}^d$ is the k th frame of the v th utterance, and $y_k \in \{-1, 1\}$ is the label of \mathbf{x}_k . If \mathbf{x}_k is a noisy speech frame, then $y_k = 1$; if \mathbf{x}_k is a noise-only frame, then $y_k = -1$. Without loss of generality, we further represent the corpus by $\mathcal{X} \times \mathcal{Y} = \{(\mathbf{x}_m, y_m)\}_{m=1}^M$ where $M = \sum_{t=1}^T K_t$, which means we concatenate all utterances to a long one.

We aim to train a DNN model for VAD, which consists of two steps. The first step expands each speech frame $\mathbf{x}'_m = [\mathbf{x}_{m-W}^T, \mathbf{x}_{m-W+1}^T, \dots, \mathbf{x}_m^T, \dots, \mathbf{x}_{m+W-1}^T, \mathbf{x}_{m+W}^T]^T$ and $\mathbf{y}'_m = [y_{m-W}, y_{m-W+1}, \dots, y_m, \dots, y_{m+W-1}, y_{m+W}]^T$, where W is a user defined half-window size. The second step uses the new training corpus $\{(\mathbf{x}'_m, \mathbf{y}'_m)\}_{m=1}^M$ to train a DNN model that has $(2W + 1)d$ input units and $2W + 1$ output units.

Test Phase. Suppose we have an unlabeled test speech corpus $\{\mathbf{x}_n\}_{n=1}^N$ and a trained DNN model. We aim to predict the label of frame \mathbf{x}_n , which consists of three

steps. The first step reformulates \mathbf{x}_n to a large observation \mathbf{x}'_n as same as in the training phase, so as to get a new test corpus $\{\mathbf{x}'_n\}_{n=1}^N$. The second step gets the $(2W + 1)$ -dimensional prediction of \mathbf{x}'_n from DNN, denoted as $\hat{\mathbf{y}}'_n = [y_{n-W}^{(-W)}, y_{n-W+1}^{(-W+1)}, \dots, y_n^{(0)}, \dots, y_{n+W-1}^{(W-1)}, y_{n+W}^{(W)}]^T$. The third step aggregates the results, which is to predict the soft decision of \mathbf{x}_n , denoted as \hat{y}_n :

$$\hat{y}_n = \frac{y_n^{(-W)} + \dots + y_n^{(-1)} + y_n^{(0)} + y_n^{(1)} + \dots + y_n^{(W)}}{2W + 1} \quad (1)$$

Finally, we make a hard decision on \hat{y}_n by

$$\bar{y}_n = \begin{cases} 1 & \text{if } \hat{y} \geq \eta \\ -1 & \text{otherwise} \end{cases} \quad (2)$$

where $\eta \in [-1, 1]$ is the decision threshold tuned on the development set according to some predefined performance measurement.

When the training corpus and the size of the half-window W are both large, one can pick a subset of the channels within the window instead of all channels, based on our observation that the window size has a larger impact on the performance than the total number of channels within the window. In this paper, we pick the channels indexed by $\{-W, -W + u, -W + 2u, \dots, -1 - u, -1, 0, 1, 1 + u, \dots, W - 2u, W - u, W\}$, where u is a user defined integer parameter.

For the DNN model, different from [14], we use the rectified linear unit for hidden layers, sigmoid function for the output layer, and a dropout strategy to specify the DNN model [19]. These regularization strategies aim to overcome the overfitting problem of DNN. In addition, we employ the adaptive stochastic gradient descent [20] and a momentum term [21] to train the DNN. These training schemes accelerate traditional gradient descent training and facilitate large-scale parallel computing. Note that no pretraining is used in our DNN training.

3. MRCG Feature

In this section, we introduce the MRCG feature which was first proposed in [16]. This feature has shown its advantage over many acoustic features in a speech separation problem. The key idea of MRCG is to incorporate the local information and global information (a.k.a, contextual information) together through multi-resolution extraction.

As illustrated in Fig. 1a, MRCG is a concatenation of 4 cochleagram features with different window sizes and different frame lengths. The first and fourth cochleagram features are generated from two 64-channel gammatone filterbanks with frame lengths set to 20 ms and 200 ms respectively. The second and third cochleagram features are calculated by smoothing each time-frequency unit of the first cochleagram feature with two square windows that are centered on the unit and have the sizes of 11×11 and 23×23 . Because the windows on the first and last few channels (or frames) of the two cochleagram features may overflow, we cut off the overflowed parts of the windows. Note that the multi-resolution strategy is a common technique but not limited to the cochleagram feature [22, 23].

After calculating the 256-dimensional MRCG feature, we further calculate its Deltas and double Deltas, and then combine all three into a 768-dimensional feature (Fig. 1b). A Delta feature is calculated by

$$\Delta x_n = \frac{(x_{n+1} - x_{n-1}) + 2(x_{n+2} - x_{n-2})}{10} \quad (3)$$

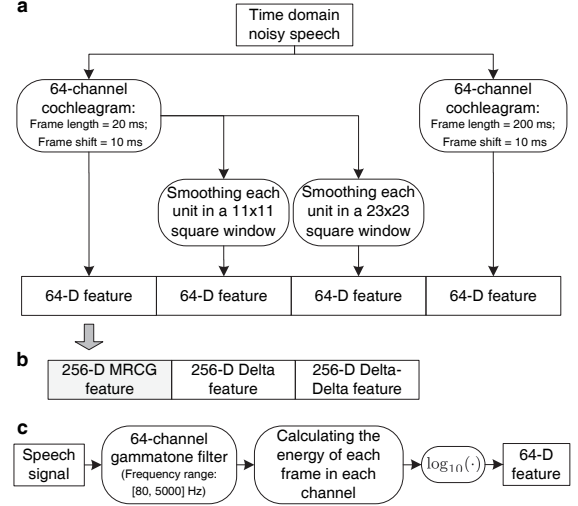


Figure 1: The MRCG feature. (a) Diagram of the process of extracting a 256-dimensional MRCG feature. “ $(2W + 1) \times (2W + 1)$ square window” means that the value of a given time-frequency unit is replaced by the average value of its neighboring units that fall into the window centered at the given unit and extending in the axes of time and frequency. (b) Expanding MRCG to a 768-dimensional feature that consists of the original MRCG feature, its Delta feature and Delta-Delta feature. (c) Calculation of the 64-dimensional cochleagram features in detail.

where x_k is the k th unit of MRCG in a given channel. The double-Delta feature is also calculated by applying equation (3) to the Delta feature. This calculation method is the same as that from MFCC to its Delta and double-Delta features.

The calculation of the 64-dimensional cochleagram feature in Fig. 1a is detailed in Fig. 1c. We first filter input noisy speech by the 64-channel gammatone filterbank, then calculate the energy of each time-frequency unit by $\sum_{k=1}^K s_{c,k}^2$ given the frame length K , and finally rescale the energy by $\log_{10}(\cdot)$, where $s_{c,k}$ represents the k -th sample of a given frame in the c -th channel [24].

4. Experiments

4.1. Experimental Settings

We used the clean speech corpus of AURORA4 [17]. The clean speech corpus consists of 7,138 training utterances and 330 test utterances. The sampling rate is 16 kHz. We randomly selected 300 and 30 utterances from the training utterances as our training set and development set respectively, and used all 330 test utterances for test. We chose three noises from the NOISEX-92 noise corpus—“babble”, “factory”, and “volvo”—to mix with the clean speech corpus at three SNR levels: $-5, 0, \text{ and } 5$ dB. As a result, we constructed 9 noisy speech corpora for evaluation. Note that for each noisy corpora, the additive noises for training, development, and test were cut from different intervals of a given noise. The manual labels of each noisy speech corpus were the results of Sohn’s VAD [3] applied to the corresponding clean speech corpus.

The area-under-ROC-curve (AUC) was used as the evaluation metric. Because over 70% frames are speech, we did not use the detection accuracy as the evaluation metric, so as to pre-

Table 1: AUC (%) comparison between the comparison VADs and proposed bDNN-based VAD. The number in bold indicates the best results.

Noise	SNR	Sohn	Ramirez05	Ying	SVM	Zhang13	bDNN
Babble	-5 dB	70.69	75.90	64.63	81.05	82.84	89.05
	0 dB	77.67	83.05	70.72	86.06	88.33	91.70
	5 dB	84.53	87.85	78.70	90.49	91.61	93.60
Factory	-5 dB	58.17	58.37	62.56	78.63	81.81	87.42
	0 dB	64.56	67.21	68.79	86.05	88.39	91.67
	5 dB	72.92	76.82	75.83	89.10	91.72	93.37
Volvo	-5 dB	84.43	89.63	92.51	93.91	94.58	94.71
	0 dB	88.25	90.44	93.42	93.43	94.80	95.04
	5 dB	90.89	90.99	94.13	94.12	95.02	95.19

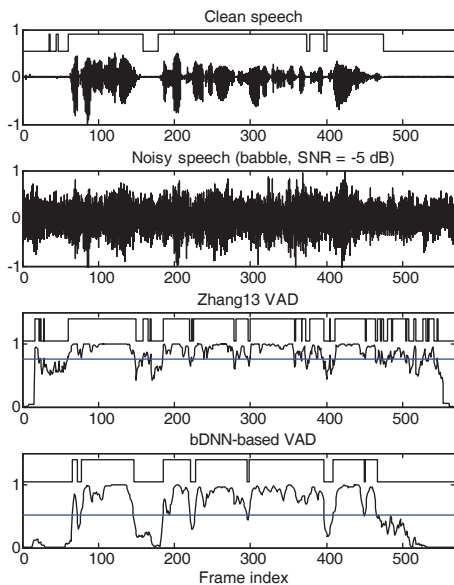


Figure 2: Illustration of the proposed and comparison methods in the babble noise environment with SNR = -5 dB. The soft outputs have been normalized so as to be shown clearly in the range [0, 1]. The straight lines are the optimal decision thresholds (on the entire test corpus) in terms of HIT-FA, and the notched lines show the hard decisions on the soft outputs.

vent reporting misleading results caused by class imbalance.

We compared the bDNN-based VAD with the following 5 VADs—Sohn VAD [3], Ramirez05 VAD [25], Ying VAD [10], Zhang13 VAD [14], and SVM-based VAD that uses the same acoustic feature as in [14].

The parameter setting of the boostDNN-based VAD was as follows. The recent advanced DNN model [20, 21] was used. The numbers of hidden units were set to 800 and 200 for the first and second hidden layer respectively. The number of epochs was set to 130. The batch size was set to 512, the scaling factor for the adaptive stochastic gradient descent [20] was set to 0.0015, and the learning rate decreased linearly from 0.08 to 0.001. The momentum [21] of the first 5 epochs was set to 0.5, and the momentum of other epochs was adjusted to 0.9. The dropout rate of the hidden units was set to 0.2. The half-window size W was set to 19, and the parameter u of the window was set to 9, i.e. only 7 channels within the window were selected.

Table 2: AUC (%) analysis on the advantages of the bDNN model and MRCG feature. “COMB” represents a serial combination of 11 acoustic features in [14]. The source code of all DNN models in this table is different from the DNN model in [14] (i.e., the DNN model of Zhang13 VAD in Table 1).

Noise	SNR	DNN+ COMB	DNN+ MRCG	bDNN +COMB	bDNN +MRCG
Babble	-5 dB	82.76	85.44	87.36	89.05
	0 dB	88.78	89.97	91.35	91.70
	5 dB	92.07	92.87	93.36	93.60
Factory	-5 dB	81.77	83.77	85.68	87.42
	0 dB	88.97	90.32	90.20	91.67
	5 dB	92.16	92.66	92.83	93.37

4.2. Results

Table 1 lists the AUC results of all 6 VAD methods. Figure 2 illustrates the soft outputs of our proposed and Zhang13 VADs for the babble noise at -5 dB SNR. From the table and figure, we observe that (i) the proposed method overall outperforms all 5 others methods when the background is very noisy; (ii) the proposed method clearly ranks the best for the two more difficult noises of babble and factory; for the volvo noise, its performance is nearly identical to that of Zhang13 VAD.

To separate the contributions of bDNN and MRCG to this significant improvement for babble and factory noises, we ran 4 experiments using either DNN or bDNN as the model with either the combination (COMB) of 11 acoustic features in Zhang13 VAD [14] or MRCG as the input feature, where the model “DNN” used the same DNN source code as that of bDNN but set $W = 0$. Table 2 lists the AUC comparison between these 4 combinations. From the table, we observe that (i) MRCG performs better than COMB, and bDNN better than DNN; (ii) both MRCG and bDNN contribute to the overall performance improvement.

To investigate how the window size of bDNN affects the performance, We evaluated the bDNN-based VAD with different windows whose parameters (W, u) were selected from $\{(3, 1), (5, 2), (9, 4), (13, 6), (19, 9)\}$ in babble and factory noises at -5 dB SNR. The results in Fig. 3 show that the ROC curve is improved steadily when the window size is gradually enlarged. Note that although different windows were used, only 7 channels within each window were selected, that is, the bDNNs maintained the same computational complexity.

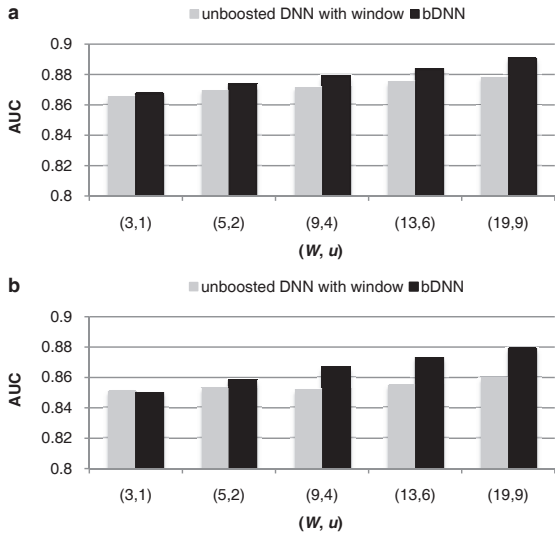


Figure 3: AUC analysis of the advantage of the boosted algorithm in bDNN-based VAD over the unboosted counterpart that uses the same input \mathbf{x}'_n as bDNN but uses the original output y_n as the training target instead of \mathbf{y}'_n . (a) Comparison in the babble noise environment with SNR = -5 dB. (b) Comparison in the factory noise environment with SNR = -5 dB. Note that (W, u) are two parameters of the window of bDNN.

To investigate how the boosted method is better than the unboosted one, we compared bDNN with a DNN model that used the same input as bDNN (i.e., \mathbf{x}'_n) but aimed to predict the label of only the central frame of the input (i.e., y_n) in two difficult environments. Results show that (i) bDNN significantly outperforms the unboosted DNN, and its superiority becomes more and more apparent when the window is gradually enlarged; (ii) the unboosted DNN can also benefit from the contextual information when comparing Fig. 3 with the corresponding results of the “DNN+MRCG” method in Table 2, but this performance gain is limited, particularly when W is large. Note that the boosted method had the same computational complexity with the unboosted one.

To show how the multi-resolution method affects the performance, we ran bDNN with MRCG and its 4 components respectively. Figure 4 gives the ROC curve comparison between the MRCG feature and its four components in the two difficult noise environments with parameters (W, u) set to $(0, 0)$ and $(19, 9)$, where $W = 0$ means that bDNN reduces to DNN. From the figure, we observe that (i) MRCG is at least as good as the best one of its 4 components in all cases, which demonstrates the effectiveness of the multi-resolution technique; (ii) CG2 yields a better ROC curve than the other 3 components; (iii) the gaps between the ROC curves are reduced when W is enlarged.

5. Concluding Remarks

In this paper, we have proposed a supervised VAD method, named bDNN-based VAD, which employs a newly introduced acoustic feature—MRCG. Specifically, bDNN first produces multiple base predictions for a single frame by boosting the contextual information (encoded in neighboring frames) and then aggregates the base predictions for a stronger one. MRCG consists of cochleagram features at multiple spectrotemporal res-

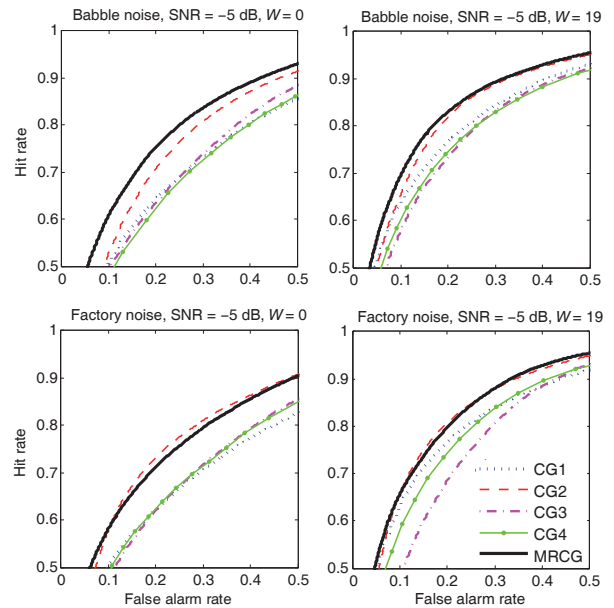


Figure 4: ROC curve analysis on the advantage of the MRCG feature over its CG components. CG1 is short for the original cochleagram feature with a frame length of 20 ms (Fig. 1). CG2 is short for the feature of the CG1 smoothed by a 11×11 sliding window. CG3 is short for the feature of the CG1 smoothed by a 23×23 sliding window. CG4 is short for the original cochleagram feature with a frame length of 200 ms. The variable W represents the half-window size of the window of bDNN.

olutions. Experimental results have shown that the proposed method outperforms the state-of-the-art VADs by a considerable margin at low SNRs. Our further analysis shows that the contextual information encoded by MRCG and bDNN both contribute to the improvement. Moreover, the window size of bDNN affects the performance significantly, and the boosted algorithm is significantly better than the unboosted version in which a DNN receives the input from a large window. Our investigation demonstrates that MRCG, originally proposed for speech separation, is effective for VAD as well. We believe that the boosting and multi-resolution ideas are not limited to DNN and cochleagram.

6. Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable advices.

This work was performed while the first author was a visiting scholar at The Ohio State University. We thank Yuxuan Wang for providing his DNN code and help in the usage of the code, Jitong Chen for providing the MRCG code, and Arun Narayanan for helping with the AURORA4 corpus. We also thank the Ohio Supercomputing Center for providing computing resources. The research was supported in part by an AFOSR grant (FA9550-12-1-0130).

7. References

- [1] D. Yu and L. Deng, “Deep-structured hidden conditional random fields for phonetic recognition,” in *Proc. Interspeech*, 2010, pp. 2986–2989.

- [2] A. Tsiartas, T. Chaspari, N. Katsamanis, P. Ghosh, M. Li, M. Van Segbroeck, A. Potamianos, and S. S. Narayanan, "Multi-band long-term signal variability features for robust voice activity detection," in *Proc. Interspeech*, 2013, pp. 718–722.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] A. Saito, Y. Nankaku, A. Lee, and K. Tokuda, "Voice activity detection based on conditional random fields using multiple features," in *Proc. Interspeech*, 2010, pp. 2086–2089.
- [7] 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.
- [8] S. Mousazadeh and I. Cohen, "Voice activity detection in presence of transient noise using spectral clustering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 6, pp. 1261–1271, 2013.
- [9] T. Yu and J. H. L. Hansen, "Discriminative training for multiple observation likelihood ratio based voice activity detection," *IEEE Signal Process. Lett.*, vol. 17, no. 11, pp. 897–900, 2010.
- [10] D. Ying, Y. Yan, J. Dang, and F. Soong, "Voice activity detection based on an unsupervised learning framework," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2624–2644, 2011.
- [11] Y. Suh and H. Kim, "Multiple acoustic model-based discriminative likelihood ratio weighting for voice activity detection," *IEEE Signal Process. Lett.*, vol. 19, no. 8, pp. 507–510, 2012.
- [12] S. O. Sadjadi and J. H. Hansen, "Robust front-end processing for speaker identification over extremely degraded communication channels," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7214–7218.
- [13] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7378–7382.
- [14] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 697–710, 2013.
- [15] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on youtube using deep neural networks," in *Proc. Interspeech*, 2013, pp. 728–731.
- [16] J. Chen, Y. Wang, and D. L. Wang, "A feature study for classification-based speech separation at very low signal-to-noise ratio," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, in press.
- [17] D. Pearce and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," Inst. for Signal & Inform. Process., Mississippi State Univ., Tech. Rep., 2002.
- [18] T. G. Dietterich, "Ensemble methods in machine learning," *Multiple Classifier Sys.*, pp. 1–15, 2000.
- [19] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 8609–8613.
- [20] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. W. Senior, P. A. Tucker *et al.*, "Large scale distributed deep networks," in *Adv. Neural Inform. Process. Sys.*, 2012, pp. 1232–1240.
- [21] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. Int. Conf. Machine Learn.*, 2013, pp. 1–8.
- [22] G. Hu and D. L. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 396–405, 2007.
- [23] S. K. Nemala, K. Patil, and M. Elhilali, "A multistream feature framework based on bandpass modulation filtering for robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 416–426, 2013.
- [24] D. L. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley-IEEE Press, 2006.
- [25] J. Ramírez, J. C. Segura, C. Benítez, L. García, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 689–692, 2005.