A COMPARATIVE STUDY OF ROBUSTNESS OF DEEP LEARNING APPROACHES FOR VAD

Sibo Tong, Hao Gu, Kai Yu

Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
SpeechLab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
{supertongsibo, guhao1994, kai.yu}@sjtu.edu.cn

ABSTRACT

Voice activity detection (VAD) is an important step for real-world automatic speech recognition (ASR) systems. Deep learning approaches, such as DNN, RNN or CNN, have been widely used in model-based VAD. Although they have achieved success in practice, they are developed on different VAD tasks separately. Whilst VAD performance under noisy conditions, especially with unseen noise or very low SNR, are of great interest, there has no robustness comparison of different deep learning approaches so far. In this paper, to learn the robustness property, VAD models based on DNN, LSTM and CNN are thoroughly compared at both frame and segment level under various noisy conditions on Aurora 4, a commonly used speech corpus with rich noises. To improve the robustness of deep learning based VAD models, a new noise-aware training (NAT) approach is also proposed. Experiments show that LSTM-based VAD is most robust but the performance degrades dramatically in the conditions with unseen noise or diverse SNR. By incorporating NAT, significant performance gains can be obtained in these conditions.

Index Terms— VAD, Deep learning, Robustness

1. INTRODUCTION

Voice activity detection (VAD) is a technique used in speech processing in which the presence or absence of human speech is detected. It is broadly applied to various speech applications such as automatic speech recognition (ASR), speech synthesis, speech coding and speech enhancement. It can directly influence the performance of these applications.

A number of techniques have been proposed for VAD, including both unsupervised systems mostly based on energy[1], zero crossing rate[2], the periodicity measure[3], higher-order statistics in LPC residual domain[4] and supervised systems, including support vector machines[5], Gaussian mixture models (GMM)[6], deep neural networks (DNN)[7, 8]. In a clean signal, or one that has high signal-to-noise ratio (SNR), the VAD problem can be solved directly using methods mentioned above. However, when the signal is corrupted by noise, it is difficult to distinguish between speech and non-speech. Researchers traditionally paid much attention to exploring new complicated acoustic features that are more discriminative, and used some specific approaches to handle various noisy conditions[9, 10, 11].

This work was supported by the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, the China NSFC project No. 61222208 and JiangSu NSF project No. 201302060012.

More recently, deep learning approaches attract great research interest due to its success in speech recognition. DNN-based VAD system can fuse the advantages of multiple features much better than traditional VADs. Recurrent neural networks (RNN)[12] and long short-term memory (LSTM) recurrent neural networks[13] have also been adopted for the reason that such kind of models model long range dependencies between the inputs and improve the robustness in real-life applications. Besides, convolutional neural network (CNN) is also applied to VAD[14, 15], since CNN can generate stronger feature vectors that are more invariant to input distortion and position and is easier to train due to parameter sharing[16].

However, each of these deep learning approaches is usually proposed aiming at some specific noise conditions and was experimented on different data sets. Although all of these deep learning methods have achieved desirable performance, the lack of thorough comparisons and analysis between them makes people still unaware of the advantages and disadvantages among these deep learning approaches on VAD task, especially under unseen noisy environment and low SNR conditions. In this paper, we investigate and analyse the noise robustness of VAD systems based on DNN, LSTM and CNN at both frame and segment level under various noisy conditions on Aurora 4, a commonly used speech corpus with rich noises. In addition, the clean speech in Aurora 4 is also used to manually mix with multiple unseen noise types at lower SNR. To improve the robustness of deep learning based VAD models, noise-aware training (NAT) is also proposed. NAT can be considered as model-space noise-adaptive training, using information about the environmental distortion during network training. Through a series of experiments on the Aurora 4 task, we show that the LSTM-based VAD system has remarkable noise robustness. By using NAT proposed in this paper, performance is further improved.

The remainder of this paper is organized as follow: in section 2, we first briefly discuss the traditional DNN-based, LSTM-based and CNN-based VAD algorithms. In section 3, the noise-aware training approach is proposed. Experimental results and analysis are provided in section 4. Finally section 5 concludes the whole paper.

2. DEEP LEARNING APPROACHES FOR VAD

2.1. DNN-based VAD system

As has been demonstrated in [7, 8], DNN-based VAD not only outperforms numerous other model-based VAD algorithms, but has a low detection complexity. This section introduces the VAD system based on a frame-based DNN classifier. DNN is used to classify an acoustic observation $\mathbf{x}$ into one of a set of classes. In VAD problem, a two-class classifier is used, which consists of a speech class and a silence class. The input vector of the DNN, which is constructed

for every frame of the input signal, is simply an extended context window of the input observation.

$$\mathbf{o_t} = [\boldsymbol{x}_{t-r}, ..., \mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}, ..., \mathbf{x}_{t+r}] \quad (1)$$

where $\mathbf{x}_t$ is the feature vector of the t^{th} frame, r denotes the length of context extension. Frame-based classification is performed by a comparison among posterior probabilities of the two classes for each frame. DNN is optimized using the cross-entropy criterion by stochastic gradient descent algorithm.

2.2. LSTM-based VAD system

Paper [13] presented a VAD approach based on LSTM RNN, which takes advantage of its ability to model long range dependencies between the inputs. The LSTM contains special units called *memory blocks*. Each memory block contains an *input gate*, an *output gate* and a *forget gate*. A memory block can be regarded as a complex and smart network unit capable of memorizing information for a long duration of time.

An LSTM network computes a mapping from an input sequence $\mathbf{x} = [\mathbf{x}_1, ..., \mathbf{x}_T]$ to an output sequence $\mathbf{y} = [\mathbf{y}_1, ..., \mathbf{y}_T]$ by calculating the network unit activations iteratively from $t = 1$ to T. More details about this architecture and training can be found in [17, 18]

When it comes to VAD problems, the input vector of the LSTM, similar to DNN, is an extended context window of the input observation. Posterior probabilities of silence class and speech class are computed for each input vector respectively. Cross-entropy is used as the optimization criterion and is minimized using the truncated back propagation through time (BPTT) learning algorithm.

2.3. CNN-based VAD system

A convolutional neural network (CNN) is composed of several *convolution layers* and fully-connected layers. Input maps are fed into convolution layers, which are composed of *convolution sub-parts* and optional *pooling sub-parts* [16]. Each convolution sub-part performs the operation of convolution with many 2-D filters. For each hidden map in convolution layer, all the inputs share one *filter*, which reduces the complexity of the entire network greatly. The pooling layers are quite simple. It is only a sampling using operation like maximum or average to reduce the dimension of data. And such structure has been proved great performance over DNN in many fields [19, 20].

A basic CNN-based VAD system has been proposed in [14, 15]. Two 2-dimension maps, representing filter-bank and first order derivative features of one frame, are fed into CNN. Also, features of each frame have been extended across time like equation (1), so that time and frequency are the two dimensions of input maps. This ensures the capacity of CNN to obtain the topology information and train time sequences properly. The objective function is cross-entropy and is minimized using back propagation[21], which is performed by stochastic gradient descent.

The above three deep learning approaches have been successfully applied on VAD task. However, they were proposed with some specific noisy environment and the used corpora were different. DNN as proposed in [8] was tested on Aurora 2 data set; Paper [13] reported the remarkable performance of LSTM based on Buckeye and TIMIT corpus and was also applied to Hollywood movie audio tracks; In paper [14, 15] IBM proposed and tested CNN-based system mainly on RATS data. Until now, no comparative study between these deep learning approaches on the same corpus has been done to investigate the robustness under noisy conditions.

3. NOISE-AWARE TRAINING FOR NEURAL NETWORKS

Researchers have attempted to add some noise information to the input of DNN in ASR systems and achieved some improvements[22]. Inspired by this, NAT-based VAD is proposed here. The noise information of each utterance is not specifically utilized in the basic structured neural network framework described above. To actualize this noise awareness, the network is fed with the noisy speech features augmented with extra estimated information about current environment conditions.

In this work, not only noise but also noisy speech information is taken into consideration. Thus, the network's input vector $\mathbf{o}_t$ is modified to an extended context window appended with noise code and noisy speech code:

$$\mathbf{o}_t = [\mathbf{x}_{t-r}, ..., \mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}, ..., \mathbf{x}_{t+r}, \mathbf{n}, \mathbf{s}] \quad (2)$$

where $\mathbf{n}$ and $\mathbf{s}$ are the appended noise code and speech code respectively and are fixed for the entire utterance as environment is assumed to remain the same for each utterance. To simplified the problem in this work, the average of filterbank features of silence frames is used as $\mathbf{n}$ and the average over speech frames is used as $\mathbf{s}$.

$$\mathbf{n} = \frac{\sum_{t \in T_{sil}} \mathbf{x}_t}{|T_{sil}|}, \mathbf{s} = \frac{\sum_{t \in T_{spch}} \mathbf{x}_t}{|T_{spch}|} \quad (3)$$

where T_{sil} and T_{spch} denote silence frames set and speech frames set of an utterance, which can directly obtained from the VAD label during training. When testing, the non-NAT VAD system is used as an assistance to get a preliminary classification. Based on those frames which have higher posterior probabilities, $\mathbf{n}$ and $\mathbf{s}$ can be calculated respectively.

4. EXPERIMENTAL COMPARISONS ON ROBUSTNESS

In this section, experiments were conducted on Aurora 4[23]. There are three sets of training data, each comprising 7138 utterances from 83 speakers. Multi-noise training set was used in this paper. One half of Aurora 4 is recorded by the primary Sennheiser microphone and the rest is recorded by different secondary microphones. Both halves include a combination of clean speech and speech corrupted by one of six different noises (street traffic, train station, car, babble, restaurant, airport) at 10-20 dB SNR. The evaluation was performed using Wall Street (WSJ0)[24] test set. This test set is recorded in the same two channels as training set and corrupted by the same six noises at 5-15 dB SNR, creating a total of 14 test sets. Notice that the types of noise are common across training and test set but SNRs are not. These 14 test sets can be grouped into 4 subsets: clean, noisy, clean with channel distortion, and noisy with channel distortion, which are indicated as A, B, C and D respectively.

In addition, a more realistic scenario is set up with many unseen noise types compared to the Aurora 4, which is more similar to the real-world application. The new noisy speech data was synthesized based on clean and clean with channel distortion test sets in Aurora 4, and 100 noise types[1] were randomly selected and added to the clean speech manually at different random SNRs from 5 to 15 dB. The two test sets are indicated as SEN and 2ND respectively.

[1] Source: http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/

To further compare the performances of each deep learning approach when the signal is severely corrupted by noise, we chose the same kinds of noise as previous test sets, including seen noise types and unseen types, to mix with the same clean speech corpus at two SNR levels: -5, 0dB, thus generating 4 new test sets. SEEN is used to denote those test sets in which the noise condition is seen in the training data and UNSEEN represents those absent noise conditions. The integration of these 4 sets is indicated as LOWSNR.

The basic feature used for all neural networks was 24-dimensional log mel filterbank and first order derivative features. In order to directly compare the performances of DNN, LSTM and CNN, the amount of parameters of each neural network was fixed to almost the same. In all following experiments, the input layers were formed from a context window of 11 frames creating an input layer of 528 visible units. The hidden layer structure of DNN was 437[units] $\times$4[layers] while it was 256[LSTM blocks] $\times$ 1[layer] for LSTM. For CNN, the entire network layer is 2[24$\times$11 input maps], 64[7$\times$2-filter and 2$\times$2-pooling maps], and 3 fully-connected layers of 64, 545, 256 units, respectively. A final layer of 2 units included speech and silence class. By such configuration, the numbers of parameters of DNN, LSTM and CNN are 804517, 804096, and 804288, respectively. That means the parameters of DNN is nearly equal to CNN's and LSTM's.

4.1. Frame-level Evaluation

Notice that no post-processing was adopted here to directly compare the modelling abilities of DNN, LSTN and CNN. The area-under-ROC-curve (AUC)[25] and equal error rate (EER) were 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 prevent reporting misleading results caused by class imbalance. The results are listed in Table 1. It can be found that LSTM model has the best performance on frame-based classification in all test sets and CNN performs worse than other two models, although the gaps between these models are very small.

Table 1. *AUC and EER (%) comparison between deep learning approaches.*

Metric	System	A	B	C	D	AVG
AUC	DNN	99.86	98.87	99.64	97.46	98.45
	LSTM	**99.89**	**99.49**	**99.72**	**98.72**	**99.19**
	CNN	99.83	98.36	99.68	97.06	98.10
EER	DNN	1.29	4.47	2.23	7.62	5.56
	LSTM	**1.10**	**2.88**	**1.94**	**5.01**	**3.79**
	CNN	1.50	5.71	1.97	8.35	6.47

Table 2. *AUC and EER (%) comparison under unseen noise conditions.*

Metric	System	SEN	2ND	AVG
AUC	DNN	85.21	88.90	86.80
	LSTM	**93.63**	**94.22**	**93.64**
	CNN	92.82	91.17	91.78
EER	DNN	16.11	14.67	15.51
	LSTM	**10.55**	**10.82**	**11.09**
	CNN	11.62	14.39	13.32

Since all the three deep learning models perform excellent under seen noisy conditions, we want to investigate their generalization abilities under unseen environment and noisier conditions. Experiments were conducted on SEN, 2ND, SEEN and UNSEEN. The results are illustrated in Table 2 and Table 3. Compared with LSTM and CNN, DNN shows a poor generalization ability and performs terrible under low SNR. Since CNN can generate stronger feature vectors that are more invariant to input distortion and position, it obtains better performance under unseen noisy conditions. When SNR is low, CNN performs better than DNN under seen noise but worse under unseen noise types. This is likely because of the mismatches in the feature transforms, especially those in the feature extracting layers of CNN. LSTM consistently has the best performance due to its strong ability to model long range dependencies between inputs.

Table 3. *AUC and EER (%) comparison under noisier conditions.*

Metric	Test set	SEEN		UNSEEN		AVG
	SNR	0db	-5db	0db	-5db	
AUC	DNN	55.15	43.54	74.53	64.71	59.48
	LSTM	**81.90**	**70.45**	**81.25**	**71.91**	**76.37**
	CNN	69.49	59.55	69.28	63.37	65.42
EER	DNN	36.00	41.15	24.05	29.45	32.61
	LSTM	**21.37**	**28.05**	**20.73**	**26.15**	**24.08**
	CNN	29.44	34.86	27.68	31.26	30.81

4.2. Segment-level Evaluation

AUC and EER are only indications of frame classification ability. In this section, we want to further investigate the specific advantages of each deep learning model from other segment-level aspects which are of importance for VAD task. To do so, another evaluation metric $\mathcal{J}_{VAD}$[26] is introduced here. $\mathcal{J}_{VAD}$ evaluates four different aspects, namely start boundary accuracy (SBA), end boundary accuracy (EBA), border precision (BP) and frame accuracy (ACC). ACC is the basic percentage of correctly recognized frames. For the start boundary s_r of utterance r, calculate a start boundary score $\mathcal{J}_S^r$ based on the interval $[s_r - L, s_r + L]$ if there exists a speech start boundary in VAD output matching it (allowing a plus or minus error margin L) following

$$\mathcal{J}_S^r = \frac{\sum_{i\in[s_r-L,s_r+L]} f(i-s_r)\delta(c^{(i)}, c_{\mathtt{ref}}^{(i)})}{\sum_{i\in[s_r-L,s_r+L]} f_s(i-s_r)} \tag{4}$$

where $c^{(i)}$ is the VAD output of the i^{th} frame and $c_{\mathtt{ref}}^{(i)}$ is the corresponding label. $\delta(\cdot)$ is the Kronecker-δ function. $f(\cdot)$ is a weighting function to give a heavier weight to the frames in speech period since they usually lead to more serious speech recognition error if misclassified. Therefore SBA is defined as the average of $\mathcal{J}_S^r$ over all utterances. It is similar for EBA. Thus, SBA and EBA are indications of boundary-level accuracy, based on which border precision can be defined as

$$\mathcal{J}_B = \frac{R}{2M}(\mathcal{J}_S + \mathcal{J}_E) \tag{5}$$

where M indicates the number of speech segments in VAD output and R denotes the number of speech segments in VAD label. Low border precision indicates that an algorithm returned substantially more incorrect borders than correct ones, which means the VAD output is more fragile. Therefore, $\mathcal{J}_B$ is a measure of the integrity of the VAD output segments. The harmonic mean of above four subcriteria is defined as $\mathcal{J}_{VAD}$. Analysis is conducted from these 4 aspects. The details are shown in Fig. 1.

ACC shows the consistent results as AUC and EER. If focusing on test A, B, C and D, we can find DNN performs slightly better than LSTM and CNN around boundaries. However BP of DNN is much worse than LSTM and CNN. That means there are a lot of false transitions between speech and non-speech in DNN output. Therefore many false alarms from non-speech periods and deletion errors from

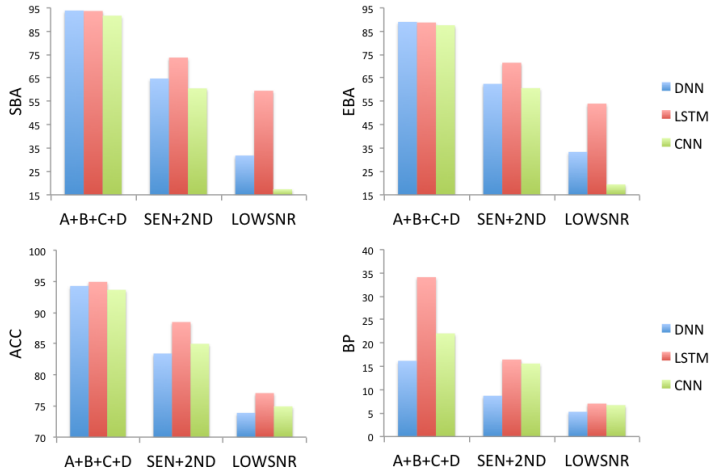

Fig. 1. *Segment-level evaluation of DNN, LSTM and CNN based VAD.*

speech segments may occur. The reason why DNN has the worst BP in all test sets can be attributed to the fact that LSTM and CNN have the better ability to capture time-series dependence between the inputs, which help reduce false transitions between speech and non-speech. When it comes to unseen noisy scenarios and low SNR conditions, LSTM consistently performs the best on both boundary accuracy and BP. DNN performs slightly better than CNN around boundaries while BP of CNN is much better than DNN. When SNR degrades to very low levels, although CNN retains its relative higher BP, the boundaries accuracy degrades severely as well. In fact, CNN tended to recognize all frames as speech when SNR degrades lower than 0db. Too much extension of speech boundaries made boundaries accuracy extremely low. In sum, under the same amount of parameters, DNN performs better around boundaries on matched test sets, but contains too much false transitions. Although CNN has a better generalization ability under unseen noisy conditions, it is not a stable system when signal is severely corrupted by noise. On the other hand, LSTM is the best frame-based classifier, which has stronger modelling and generalization ability.

4.3. Comparison After Post-processing

Since the nature of VAD is different from normal binary classification problems, it is important to apply post-processing to smooth the raw output and reduce false transitions due to weak speech tails or abrupt noise presence. Therefore, performances of different VAD systems were compared after applying post-processing. Specifically, two distinct post-processing stages are applied. In the first stage, preliminary decisions made by the system are smoothed using a running window in order to reduce short term variations. In the second stage, we merge the small segments with very short durations.

Table 4. *Performance of VAD after post-processing.*

Metric	Model	A+B+C+D	SEN+2ND	LOWSNR
EER	DNN	3.52	11.19	29.31
	LSTM	**3.15**	**9.24**	**23.64**
	CNN	5.05	9.76	28.36
$\mathcal{J}_{VAD}$	DNN	**84.66**	48.68	20.80
	LSTM	83.80	**65.39**	**41.47**
	CNN	82.28	55.23	14.43

Considering that post-processing based on flawed preliminary classifications when choosing some extreme thresholds may lead to more errors, AUC might not be reliable. Therefore, EER and $\mathcal{J}_{VAD}$ are used as the evaluation metrics here. From the table, it shows that LSTM obtains the best performance on most test sets after post-processing. As is discussed before, the weakness of DNN lies in the frequent false transitions. Since post-processing is specially used to address this error and DNN detects boundaries more accurately, DNN gains more improvement than LSTM and CNN and obtains the best $\mathcal{J}_{VAD}$ on matched test sets. As for the unseen scenarios, CNN consistently outperforms DNN, which means CNN has advantage in dealing with unseen noise distortion. When SNR degrades to low level, CNN fails to detect correct boundary, thus making $\mathcal{J}_{VAD}$ lower than DNN, although it has better EER.

4.4. Effect of Noise-aware Training

To evaluate the proposed technique designed to increase the noise robustness of these systems, noise-aware training was adopted on all systems. CNN, due to its special input format, needs a different noise-aware training method. The noise code and speech code should be appended on the first fully-connected layer of CNN, rather than input maps. The estimation of noise and noisy speech was computed simply by averaging the static input features over silence frames and speech frames separately and fixed for the entire utterance. The dimensions of the NAT features are 24. We compared DNN-based, LSTM-based and CNN-based NAT system and the results are listed in Table 5.

Table 5. *Performance of NAT-based VAD systems.*

Metric	Model	A+B+C+D	SEN+2ND	LOWSNR
EER	DNN-NAT	3.14	8.58	28.74
	LSTM-NAT	**2.82**	**6.72**	**16.86**
	CNN-NAT	3.30	7.14	21.22
$\mathcal{J}_{VAD}$	DNN-NAT	**85.82**	64.10	35.90
	LSTM-NAT	85.52	**72.04**	**51.56**
	CNN-NAT	84.74	70.21	38.11

Compared with Table 4, all the three systems achieve significant improvement on all test sets after adopting noise-aware training method. DNN with noise-aware training achieves a level of performance equivalent to or better than LSTM. For CNN, noise-aware training dramatically remedies the disadvantages of CNN when SNR degrades to lower than 0db, and CNN achieves better performance on all test sets. LSTM-NAT consistently obtains the best overall performance. Therefore, the proposed NAT is an effective method to further improve VAD performance especially under noisy conditions, although it introduces another detection pass.

5. CONCLUSION

In this paper, VAD systems based on different deep learning approaches, DNN, LSTM and CNN are thoroughly compared from the robustness aspect. Through a series of experiments on Aurora4, it is demonstrated that LSTM is more robust than CNN and DNN under various circumstances. Although all deep learning approaches performed well under noise-matched conditions, very large performance degradations were observed in conditions with unseen noise or very low SNR for all approaches. To address this issue, noise aware training (NAT) is proposed in this paper. Experiments showed that with NAT, significant performance gains can be achieved for all deep learning approaches under unseen noise or very low SNR conditions. We observed the same conclusion on real-world noisy data but the results are not listed due to the limited space in this paper.

6. REFERENCES

[1] Kyoung Ho Woo, Tae Young Yang, Kun Jung Park, and Chungyong Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.

[2] Jean Claude Junqua, Ben Reaves, and Brian Mak, "A study of endpoint detection algorithms in adverse conditions: incidence on a DTW and HMM recognizer," in *Second European Conference on Speech Communication and Technology*, 1991.

[3] R Tucker, "Voice activity detection using a periodicity measure," *IEE Proceedings I (Communications, Speech and Vision)*, vol. 139, no. 4, pp. 377–380, 1992.

[4] Elias Nemer, Rafik Goubran, and Samy Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 217–231, 2001.

[5] Nima Mesgarani, Malcolm Slaney, and Shihab A Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, 2006.

[6] Tim Ng, Bing Zhang, Long Nguyen, Spyros Matsoukas, Xinhui Zhou, Nima Mesgarani, Karel Veselỳ, and Pavel Matejka, "Developing a speech activity detection system for the DARPA RATS program," in *Proc. Interspeech*, 2012.

[7] Neville Ryant, Mark Liberman, and Jiahong Yuan, "Speech activity detection on YouTube using deep neural networks," in *Proc. Interspeech*, 2013.

[8] Xiao-Lei Zhang and Ji Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.

[9] Jinsoo Park, Wooil Kim, David K Han, and Hanseok Ko, "Voice activity detection in noisy environments based on double-combined fourier transform and line fitting," *The Scientific World Journal*, vol. 2014, 2014.

[10] Masakiyo Fujimoto, Kentaro Ishizuka, and Tomohiro Nakatani, "A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme," in *Proc. ICASSP*, 2008.

[11] Peng Teng and Yunde Jia, "Voice activity detection via noise reducing using non-negative sparse coding," *Signal Processing Letters, IEEE*, vol. 20, no. 5, pp. 475–478, 2013.

[12] Thad Hughes and Keir Mierle, "Recurrent neural networks for voice activity detection," in *Proc. ICASSP*, 2013.

[13] Florian Eyben, Felix Weninger, Stefano Squartini, and Björn Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to hollywood movies," in *Proc. ICASSP*, 2013.

[14] Samuel Thomas, Sriram Ganapathy, George Saon, and Hagen Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," in *Proc. ICASSP*, 2014.

[15] George Saon, Samuel Thomas, Hagen Soltau, Sriram Ganapathy, and Brian Kingsbury, "The IBM speech activity detection system for the DARPA RATS program," in *Proc. Interspeech*, 2013.

[16] Yann LeCun and Yoshua Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, 1995.

[17] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] Hasim Sak, Andrew Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. Interspeech*, 2014.

[19] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back, "Face recognition: A convolutional neural-network approach," *Neural Networks, IEEE Transactions on*, vol. 8, no. 1, pp. 98–113, 1997.

[20] Yann LeCun, Fu Jie Huang, and Leon Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*. IEEE, 2004, vol. 2, pp. II–97.

[21] Ossama Abdel-Hamid, Li Deng, and Dong Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition.," in *Proc. Interspeech*, 2013.

[22] Michael L Seltzer, Dong Yu, and Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. ICASSP*, 2013.

[23] N Parihar and J Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," *Inst. for Signal and Information Process, Mississippi State University, Tech. Rep*, vol. 40, pp. 94, 2002.

[24] Douglas B Paul and Janet M Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[25] James P Egan, *Signal detection theory and ROC-analysis*, Academic Press, 1975.

[26] Sibo Tong, Nanxin Chen, Yanmin Qian, and Kai Yu, "Evaluating VAD for automatic speech recognition," in *Proc. ICSP*, 2014.