Voice Activity Detection Based on Time-Delay Neural Networks

Ye Bai*†, Jiangyan Yi*, Jianhua Tao*†‡, Zhengqi Wen* and Bin Liu*
* National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, China
† School of Artificial Intelligence, University of Chinese Academy of Sciences, China
‡ CAS Center for Excellence in Brain Science and Intelligence Technology,
Institute of Automation, Chinese Academy of Sciences, China
E-mail: baiye2016@ia.ac.cn, {jiangyan.yi, jhtao, zqwen, liubin}@nlpr.ia.ac.cn

***Abstract*—Voice activity detection (VAD) is an important preprocessing part of many speech applications. Context information is important for VAD. Time-delay neural networks (TDNNs) capture long context information with a few parameters. This paper investigates a TDNN based VAD framework. A simple chunk based decision method is proposed to smooth raw posteriors and decide border points of utterances. To evaluate decision performance, a metric intersection-over-union (IoU) is introduced from image object detection. The experiment results are evaluated on Wall Street Journal (WSJ0) corpus. Frame classification performance is measured by area under the curve (AUC) and equal error rate (EER). Compared with long short-term memory baseline, the TDNN based system achieves a 41.26% EER relative reduction on average in matched noise condition, and relative improvement of average AUC is 3.82%. Proposed decision method achieves an 18.74% IoU relative improvement on average compared with moving average method on average.**

I. Introduction

Voice activity detection (VAD), also sometimes known as endpoint detection, is an important preprocessing step in many speech applications, such as automatic speech recognition (ASR), keywords spotting, speaker verification, and speech emotion recognition. An accurate VAD module can detect speech segments in continuous audio stream to improve performance of speech applications. Many conventional VAD approaches are based on manual rules. In this case, acoustic features are used to determine utterance endpoints in terms of thresholds based on heuristic rules [1]. Frame energy and zero-crossing rate (ZCR) are measured to detect endpoints of isolated utterances [2], [3]. Least-squares periodicity estimator based VAD is also implemented [4] to detect endpoints. Long-term spectral divergence (LTSD) is adopted to discriminate speech and non-speech segments [5]. These methods are sensitive to thresholds so that the rules need to be configured for different environments. Pattern matching based methods are also introduced to VAD. Itakura LPC distance is used to measure distance between test signals and reference patterns to determine whether the test sample is speech [6]. In another category, statistical model based methods are used to model speech and non-speech signals. Gaussian, generalized Gaussian, Gamma, or Laplacian distributions are adopted to model speech and non-speech segments, and soft decision method is used for inference [7], [8], [9]. Hidden Markov model (HMM) is also used to model speech and non-speech sequences [10]. These methods assume distributions of features, so that the representation ability is constrained. Statistical learning based discriminative models are also used to classify speech and non-speech segments, such as support vector machines (SVMs) [11], AdaBoost [12]. However, generalization performance of these methods is limited.

Instead of raw features, deep learning techniques capture complex structures from data by multiple non-linear transformations, and outperform other systems in many speech tasks [13], [14], [15]. Deep neural networks (DNNs) [16], [17], [18], [19], [20], recurrent neural networks (RNNs) [21], [22], and convolutional neural networks (CNNs) [23], [24], [25] based approaches improved robustness of VAD systems in multiple low signal-to-noise ratio (SNR) conditions. Comparative study of these three types of neural networks was analyzed, and long short-term memory (LSTM) outperformed DNNs and CNNs [26].

Context information is important for improving VAD performance [19]. A common approach to capture context information is stacking contiguous frames as input to a neural network. However, it costs more computation to capture long context information because of big initial affine layers. RNNs can capture all context information in the past by recurrent structure. However, it's difficult to perform parallelization and its computation cost is higher than feedforward neural networks. Time-delay neural networks (TDNNs) can capture longer context information through time [27]. Its higher layers capture time invariance from longer temporal context information. In addition, its computation can be reduced by subsampling [28]. This architecture has been applied to ASR successfully.

In this paper, we present a TDNN based VAD system. A chunk based border point decision method is also proposed. To evaluate decision performance, a metric intersection-over-union (IoU) is introduced from image object detection [29] and speech/non-speech segmentation performance is evaluated. This metric indicates accuracy of border point decision explicitly. To the best of our knowledge, it is the first work which uses IoU to evaluate VAD performance. Frame classification

is also measured by area under the curve (AUC) and equal error rate (EER). The experiments are evaluated on Wall Street Journal (WSJ) corpus. Our TDNN based VAD system achieves a 41.26% EER relative reduction on average in matched noise condition, and relative improvement of average AUC is 3.82%, compared with LSTM baseline. And compared with moving average method, our border point decision method achieves 18.74% IoU relative improvement on average.

The rest of this paper is organized as follows. Section II describes proposed VAD method. The experimental setup, the segmentation metric, and results are described in Section III. Finally Section IV summarizes the paper.

II. Proposed VAD System

The proposed VAD framework is illustrated in Figure 1. The feature extraction part extracts Mel-frequency cepstral coefficients (MFCCs) every 10ms with 25ms of frame length. Part (2) is a time-delay neural network to estimate speech/non-speech probabilities for every frame. Part (3) decides border points based on chunk. Part (4) detects endpoints of utterances and segments input signals.

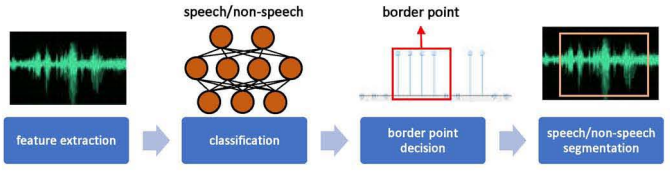

Fig. 1. System framework: (1) feature extraction (2) speech/non-speech probabilities estimation (3) border points decision (4) speech/non-speech segmentation.

A. TDNN Architecture

A TDNN architecture is used for VAD to classify each frame as speech or non-speech. It is inspired by [28]. In TDNN architecture, narrow temporal context are captured at low-level layer, and longer temporal context are captured at high level hidden layers. A TDNN can be seen as a non-linear transformation sliding along feature sequences, and time-invariant feature transform is learned during training. Otherwise, compared with standard unidirectional RNNs, the TDNN architecture can capture information from the "future", and it's effective for VAD prediction. Because TDNNs are a kind of feedforward neural networks, it can be parallelized to improve computational efficiency.

The network used in the system is showed in Figure 2. The overall architecture is like a pyramid. We use notation $\{n\}$ denotes offset value of the frame. Specifically, in Figure 2, $\{-2,-1,0,1,2\}$ means current frame (offset is zero) and current frame minus one, current frame minus two, current frame plus one, and current frame plus two are input to the network together. To reduce computation, sub-sampling technique is also used in the architecture. At hidden layers, several noncontiguous high-level representations are propagated to next layer. For instance, at second layer, the configuration is $\{-2,0,2\}$, and the configuration of last layer is $\{-3,0,3\}$.

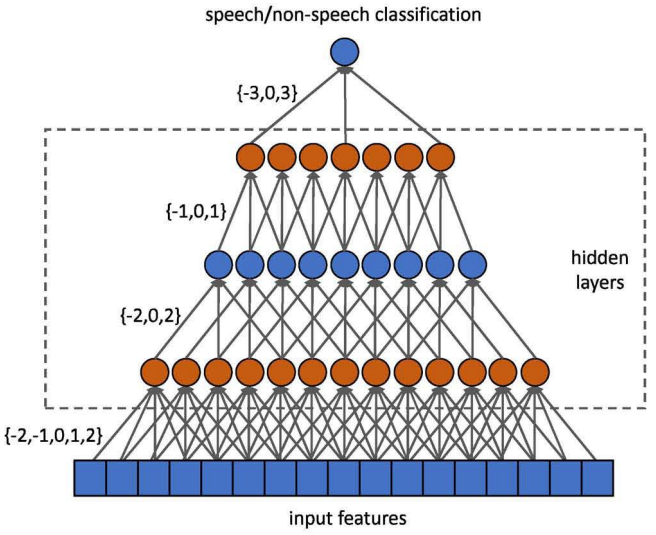

Fig. 2. Proposed neural network architecture.

Using this technique, long context information is remained and computation cost is reduced.

We use rectifier linear units (ReLUs) as activation functions in the proposed neural networks [30]. ReLUs provide sparse representation capacity for neural networks. This capacity can be considered as sparsity-inducing regularization, which can improve generalization of the networks. Compared with other activation functions, e.g., sigmoid or tanh, computation of ReLUs is very cheap: exponential function is not needed in these units, only a comparison operation is used.

B. Chunk Based Decision

Because noise exists in raw posteriors from the neural networks, we propose a chunk based decision method to smooth the raw results and decide border points.

Consider a chunk of N contiguous frames. The problem is whether the endpoint e_i is in this chunk. Assuming probability of being non-speech of j-th frame is p_j, then probability of the endpoint in this chunk can be calculated as follows:

$$P_{ep} = 1 - \prod_{j=1}^{N}(1 - p_j), \tag{1}$$

where P_{ep} denotes the probability of an endpoint existing in the chunk. In this case, we assume that the probability of each frame is independent, and estimated by proposed TDNN. Note that $N \to \infty$, $P_{ep} \to 1$ means that an endpoint exists in the whole utterance, and when $N = 1$, $P_{ep} = p_1$, P_{ep} is determined by the probability of the current frame. This method is a simple smoothing strategy, which use context information in a fixed window, to improve segmentation performance.

We use this chunk based endpoint decision method to detect endpoints of utterances. When P_{ep} is larger than some threshold, e.g., 0.95, a border point is considered in that chunk at high confidence level. Then a simple two-state finite state machine (FSM) is used to segment the signal. When the system

is in "speech" state, contiguous frames between two adjacent border points will be recorded as a speech segment. And when the system is in "non-speech" state, contiguous frames between two adjacent border points will be recorded as a non-speech segment. Transitions between the states are made in terms of border point decision.

III. Experiments

A. Datasets

Proposed VAD system is evaluated on Wall Street Journal (WSJ0) corpus. The sample rate of the data is 16kHz. The training set consists of 8343 utterances. To improve robustness of the models, we randomly add different noises to raw data. The noise corpora are MUSAN [31], HuNoises [1], and self-collected 250 music. Clean utterances and noisy utterances are combined to train the system, so total number of training data is 16686. The labels of clean utterances are generated by forced alignment using a GMM-HMM ASR system, and labels of noise added utterances are generated in terms of corresponding clean utterances. We use WSJ0 test set, which consists of 330 utterances, to evaluate the system performance. Similar noise data is randomly selected to added to be test set. To test the system in more realistic scenarios, unseen noises, i.e., noises not used in training stage, are also used to synthesize test data. Noisex92[2] [32] corpus and 50 self-collected music are added to WSJ0 test set with different SNRs.

B. Segmentation Evaluation Metric

To evaluate segmentation performance, we introduce intersection-over-union (IoU) metric, which is often used to evaluate accuracy in object detection task [29]. For a predicted speech segment, and a corresponding ground-truth speech segment (e.g., a voice segment labeled by forced alignment), IoU is defined as follows:

$$IoU = \frac{T_{pred} \cap T_{groundtruth}}{T_{pred} \cup T_{groundtruth}}, \tag{2}$$

where T_{pred} denotes predicted segment, i.e., endpoints of the predicted segment, and $T_{groundtruth}$ denotes corresponding ground-truth segment.

This metric evaluates overlap between the predicted segment and the ground-truth segment. When the predicted segment overlaps the ground-truth segment entirely, the IoU value is 1. The corresponding ground-truth segment of the predicted segment is selected by:

$$\hat{i} = \arg\max_{i} IoU(i), \tag{3}$$

where i is index of ground-truth segment, and $\hat{i}$ is selected ground-truth segment corresponding to predicted segment. Only when IoU value is larger than a threshold, that ground-truth is considered recalled.

$$\hat{IoU} = \begin{cases} IoU(\hat{i}) & IoU(\hat{i}) > threshold \\ 0 & IoU(\hat{i}) \leq threshold \end{cases}, \tag{4}$$

Finally, mean $\hat{IoU}$, which is denoted as ratio of total $\hat{IoU}$ value to the number of predictions, is used to evaluate speech segmentation performance of the VAD system. Otherwise, we also evaluate recall value of ground-truth segments, which is denoted as ratio of recalled predictions to total number of ground-truth, to evaluate the system. Mean $\hat{IoU}$ is used to measure precision of systems, and recall value is used to measure recall capacity.

C. Results

Proposed system used 13-dimensional MFCCs as feature. To capture absolute energy information of each frame, normalization technique, e.g., cepstral mean and variance normalization (CMVN), was not used. To compare the effectiveness of the TDNN system, we also implemented LSTM and DNN based VAD systems as baselines. All experiments were implemented using Kaldi speech recognition toolkit [33].

The baseline DNN architecture consists of three hidden layers and a binary classification softmax output layer. Each layer has 200 nodes. Three frames in the past and three future frames are stacked with current frame to input to the network. All non-linear activation units are ReLUs. The total number of parameters is 144602. The model outputs speech/non-speech label for each frame. The baseline unidirectional LSTM architecture has two layers. Each layer consists of 256 LSTM cells. The total number of parameters is 398338.

The architecture of proposed TDNN consists of four hidden layers. The structure of the networks is illustrated in Figure 2. Sub-sampling technique was used in second hidden layer and last layer. Each layer has 120 ReLUs. The total number of the network is 138122.

Because of imbalance of distribution of speech data and non-speech data, we did not use accuracy as evaluation metric directly. Area-under-ROC-curve (AUC) and equal equal error rate (EER) were used to evaluate Table I shows experimental result in matched noise condition. The proposed TDNN outperforms DNN and LSTM baseline systems. The TDNN shows robustness in low SNR conditions, especially 5dB, 0dB, -5dB.

TABLE I
FRAME LEVEL CLASSIFICATION PERFORMANCE IN MATCHED NOISE CONDITIONS.

		15dB	10dB	5dB	0dB	-5dB	average
AUC	DNN	0.9828	0.9732	0.9573	0.9315	0.8905	0.9471
	LSTM	0.9874	0.9782	0.9605	0.9287	0.8778	0.9465
	TDNN	**0.9958**	**0.9935**	**0.9888**	**0.9788**	**0.9564**	**0.9827**
EER	DNN	0.0623	0.0835	0.1112	0.1488	0.1990	0.1210
	LSTM	0.0498	0.0692	0.1019	0.1494	0.2102	0.1001
	TDNN	**0.0258**	**0.0338**	**0.0481**	**0.0724**	**0.1138**	**0.0588**

[1] http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html
[2] http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

TABLE II
FRAME LEVEL CLASSIFICATION PERFORMANCE IN UNSEEN NOISE CONDITIONS.

Noise	System	Metric	20dB	15dB	10dB	5dB	0dB	-5dB	Average
babble	DNN	AUC	0.9759	0.9525	0.9170	0.8632	0.7869	0.6973	0.8655
		EER	0.0724	0.1045	0.1477	0.2078	0.2809	0.3546	0.1947
	LSTM	AUC	**0.9900**	**0.9784**	**0.9507**	0.8868	0.7796	0.6609	0.8744
		EER	**0.0392**	**0.0684**	**0.1161**	0.195	0.2937	0.3868	0.1832
	TDNN	AUC	0.9810	0.9622	0.9335	**0.8912**	**0.8263**	**0.7393**	**0.8889**
		EER	0.0651	0.0931	0.1266	**0.1695**	**0.2310**	**0.3031**	**0.1647**
factory	DNN	AUC	0.9879	0.9759	0.9543	0.9163	0.8473	0.7379	0.9033
		EER	0.0406	0.0648	0.1013	0.1548	0.233	0.3305	0.1542
	LSTM	AUC	0.9931	0.9879	0.9768	0.948	0.8753	0.7441	0.9209
		EER	0.0261	0.0383	0.0643	0.1157	0.2062	0.3244	0.1292
	TDNN	AUC	**0.9955**	**0.9916**	**0.9845**	**0.9688**	**0.9270**	**0.8284**	**0.9493**
		EER	**0.0206**	**0.0300**	**0.0466**	**0.0777**	**0.1424**	**0.2559**	**0.0955**
volvo	DNN	AUC	0.9959	0.995	0.9937	0.9916	0.9886	0.9838	0.9914
		EER	0.0244	0.0271	0.0295	0.0332	0.0386	0.047	0.0333
	LSTM	AUC	0.9951	0.9933	0.9911	0.9885	0.9852	0.9794	0.9888
		EER	0.0252	0.0294	0.0341	0.0388	0.0439	0.0557	0.0379
	TDNN	AUC	**0.9982**	**0.9981**	**0.9977**	**0.9971**	**0.9961**	**0.9947**	**0.9970**
		EER	**0.0148**	**0.0159**	**0.0172**	**0.0188**	**0.0209**	**0.0242**	**0.0186**
music	DNN	AUC	0.9887	0.9804	0.9658	0.9399	0.8958	0.8287	0.9332
		EER	0.0472	0.0656	0.0929	0.1329	0.1874	0.2554	0.1302
	LSTM	AUC	0.9895	0.9808	0.9639	0.9289	0.8643	0.7747	0.9170
		EER	0.0433	0.0648	0.0977	0.1486	0.2188	0.2978	0.1452
	TDNN	AUC	**0.9960**	**0.9932**	**0.9879**	**0.9783**	**0.9584**	**0.9170**	**0.9718**
		EER	**0.0238**	**0.0325**	**0.0467**	**0.0675**	**0.1029**	**0.1593**	**0.0721**

From Table I, it can be seen that proposed TDNN outperforms the baseline DNN and LSTM based systems. Compared with LSTM, average EER is reduced by 41.26% relatively, and relative improvement of average AUC was 3.82%.

Table II shows experimental results of the three networks in different unseen noise conditions. The volvo noise recorded in a running car is a kind of typical low-frequency noise. The factory noise consists of intermittent knocking noise and roars of machines. These two kinds of noise were used to test influence on system performance of low-frequency noise and non-stationary noise. Babble is a kind of noisy human speech sound. We used it to test speech-like noise influence on systems. Music, which covers various frequency range, was used to test influence of harmonic noise. All these four noisy environments are common in real-life applications.

The three networks perform robustly in these four conditions. Even in -5dB, influence of the volvo noise, i.e., continuous low-frequency noise from motors in car, is limited. Influence of the music noise and factory noise is distinct to these three models, especially in very low SNRs, i.e., 0dB and -5dB. Babble noise is consists of human voice. And it is difficult for the models to distinguish foreground speech and background noise voice in low SNR conditions. So performance in babble noise environment is worst. In this comparison, proposed TDNN also outperforms baseline systems on average. Note that the TDNN performs better than LSTM, it is due to the fact, that the TDNN captures left context and right context simultaneously, but LSTM is unidirectional. In addition, the parameters of TDNN is fewer than DNN and LSTM baselines.

To evaluate proposed chunk based border point decision method, we did segmentation experiments. The network used in these experiments was proposed TDNN. In this experiment, we concatenated all 330 test set audio to one file, which is 40.3 minutes. Noise from Noisex92 set is added to the file randomly in different SNRs. The ground-truth labels of segments were generated for each audio file using force alignment, and then were concatenated. As comparison, direct decision on

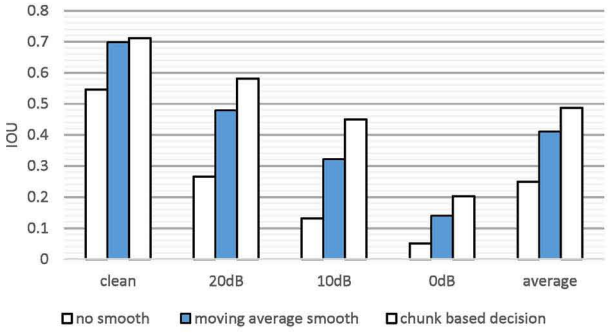

Fig. 3. IoUs vs. SNRs

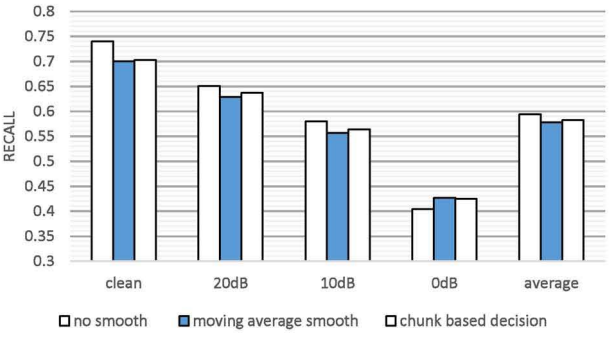

Fig. 4. Recalls vs. SNRs

raw prediction results and commonly used moving average method [26] to smooth the raw prediction results was used as baselines. We did experiments to find proper hyper-parameters of the systems. We set threshold of direct decision as 0.55. The baseline moving average smoothing method used fixed window of five and threshold as 0.45. The chunk size of proposed chunk based decision was nine, and the threshold was 0.95.

The IoU results are illustrated in Figure 3. The proposed decision method performs more robustly in noise environments than baseline moving average smoothing method. In 0dB noise condition, IoU value of proposed chunk based decision improves by 45.26% than moving average method relatively. Proposed method achieves 18.74% IoU relative improvement over to moving average method on average. Even though moving average method smooths sudden change in raw posteriors, it doesn't fuse context information for decision. However, proposed method reduces influence of a single frame, and fuses a chunk of frames for decision. The recall values in different conditions are showed in Figure 4. Because smoothing posteriors decreases sensitivity of decision, recall values are hurt a little. Relative to moving average method, the proposed method attains higher recall values in high SNR conditions.

IV. Conclusions and Future Work

This paper proposes a TDNN based VAD framework. A chunk based decision method is used to segment speech/non-speech signals. IoU metric is introduced from image object detection to measure performance of segmentation. Performance in matched and unseen noise conditions are evaluated separately. In future work, we will improve the system performance in low SNR condition with speech enhancement techniques. Otherwise, more efficient decision method and more complex FSMs will be investigated to improve segmentation performance.

V. Acknowledgements

This work is supported by the National Key Research Development Plan of China (No.2017YFC0820602), the National Natural Science Foundation of China (NSFC) (No.61425017, No.61831022, No.61773379, No.61603390), the Strategic Priority Research Program of Chinese Academy of Sciences (No.XDC02050100), and Inria-CAS Joint Research Project (No.173211KYSB20170061).

References

[1] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," *IEEE Transactions on speech and audio processing*, vol. 8, no. 4, pp. 478–482, 2000.

[2] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.

[3] L. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 4, pp. 777–785, 1981.

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

[5] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3, pp. 271–287, 2004.

[6] L. Rabiner and M. Sambur, "Voiced-unvoiced-silence detection using the itakura lpc distance measure," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'77.*, vol. 2. IEEE, 1977, pp. 323–326.

[7] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1. IEEE, 1998, pp. 365–368.

[8] T. Petsatodis, C. Boukis, F. Talantzis, Z. Tan, and R. Prasad, "Convex combination of multiple statistical models with application to vad," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2314–2327, 2011.

[9] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204–207, 2003.

[10] J. Wilpon and L. Rabiner, "Application of hidden markov models to automatic speech endpoint detection," *Computer Speech & Language*, vol. 2, no. 3-4, pp. 321–341, 1987.

[11] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, "Applying support vector machines to voice activity detection," in *Signal Processing, 2002 6th International Conference on*, vol. 2. IEEE, 2002, pp. 1124–1127.

[12] O.-W. Kwon and T.-W. Lee, "Optimizing speech/non-speech classifier design using adaboost," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1. IEEE, 2003, pp. I–I.

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

[14] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[15] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 241–245.

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

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

[18] B. Liu, J. Tao, F. Mo, Y. Li, Z. Wen, and S. Liu, "Efficient voice activity detection algorithm based on sub-band temporal envelope and sub-band long-term signal variability," in *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*. IEEE, 2014, pp. 531–535.

[19] X.-L. Zhang and D. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[20] Q. Wang, J. Du, X. Bao, Z.-R. Wang, L.-R. Dai, and C.-H. Lee, "A universal vad based on jointly trained deep neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[21] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7378–7382.

[22] M. Shannon, G. Simko, S.-y. Chang, and C. Parada, "Improved end-of-query detection for streaming speech recognition," *Proc. Interspeech 2017*, pp. 1909–1913, 2017.

[23] G. Saon, S. Thomas, H. Soltau, S. Ganapathy, and B. Kingsbury, "The ibm speech activity detection system for the darpa rats program." in *INTERSPEECH*, 2013, pp. 3497–3501.

[24] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference On*. IEEE, 2014, pp. 2519–2523.

[25] A. Sehgal and N. Kehtarnavaz, "A convolutional neural network smartphone app for real-time voice activity detection," *IEEE Access*, vol. 6, pp. 9017–9026, 2018.
[26] S. Tong, H. Gu, and K. Yu, "A comparative study of robustness of deep learning approaches for vad," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5695–5699.
[27] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE transactions on acoustics, speech, and signal processing*, vol. 37, no. 3, pp. 328–339, 1989.
[28] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
[29] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.
[30] X. Glorot, A. Bordes, Y. Bengio, X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.
[31] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *Computer Science*, 2015.
[32] 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.
[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.