PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Gated convolutional neural network-based voice activity detection under high-level noise environments

Li LI, Kouei YAMAOKA, Yuki KOSHINO, Mitsuo MATSUMOTO, Shoji MAKINO

University of Tsukuba, Japan
lili@mmlab.cs.tsukuba.ac.jp

Abstract

This paper deals with voice activity detection (VAD) tasks under high-level noise environments where signal-to-noise ratios (SNRs) are lower than −5 dB. With the increasing needs for hands-free applications, it is unavoidable to face critically low SNR situations where the noise can be internal self-created ego noise or external noise occurring in the environment, e.g., rescue robots in a disaster or navigation in a high-speed moving car. To achieve accurate VAD results under such situations, this paper proposes a gated convolutional neural network-based approach that is able to capture long- and short-term dependencies in time series as cues for detection. Experimental evaluations using high-level ego noise of a hose-shaped rescue robot revealed that the proposed method was able to averagely achieve about 86% VAD accuracy in environments with SNR in the range of −30 dB to −5 dB.

Keywords: Voice activity detection (VAD), low SNR, gated convolutional neural networks, rescue robot, ego noise

1 INTRODUCTION

Voice activity detection (VAD) is referred to as the technique of identifying speech and non-speech regions in a recorded audio signal, which is an active research area in the field of speech processing since it plays an essential role in numerous speech applications [1, 2, 3]. In high-quality recording conditions where signal-to-noise ratios (SNR) are high, methods based on the energy perform well because the energy difference between speech and non-speech segments is noticeable. However, the performances of these methods decrease when SNR tends to be critically low. On the other hand, with the rapidly increasing needs for hands-free applications and robots, high VAD accuracies are required under various situations without regard to SNRs. This means that developing VAD systems handling critically low SNRs becomes indispensable since these situations frequently occur around us, such as navigation in a high-speed moving car or robot audition [4, 5].

To deal with such arduous tasks and motivated by the considerable success in many classification tasks achieved by deep learning-based methods, some attempts have recently been made to adopt deep neural networks (DNNs) to VAD tasks [6, 7, 8, 9, 10, 11]. In [7, 8], multiple layer perceptron (MLP) is employed to train a nonlinear speech/non-speech classifier. To capture the dependence in time series, these methods take segmental features that concatenate the feature vectors extracted from the current frame, preceding frames and succeeding frames as inputs of neural networks. However, capturing the temporal dependency in such naïve way will lead to a critical increase of the feature dimension, which significantly increases the difficulty of training a classifier with good generalization. Comparing with MLP, recurrent neural networks (RNN) and convolutional neural networks (CNN) are more efficient architectures to model time series data. Especially, long short-term memory (LSTM) networks [12, 13] (also bidirectional LSTM and gated recurrent units) and gated CNN [14] have shown their strong capability to capture long-term dependencies of time series in many recent studies, including VAD tasks [10, 11]. Although RNN, including LSTM, is proposed initially to model time series, enormous parameters involved in models lead to some well-known problems. Namely, it is cost-consuming to train an acceptable model, and RNN is prone to overfitting. Furthermore, it has been reported in [10] that LSTM suffers from state saturation problems for long-utterance in VAD tasks. In contrast, CNN has a relatively small number of parameters thanks to its parameter sharing scheme, and with the gated mechanism and the dilation process,

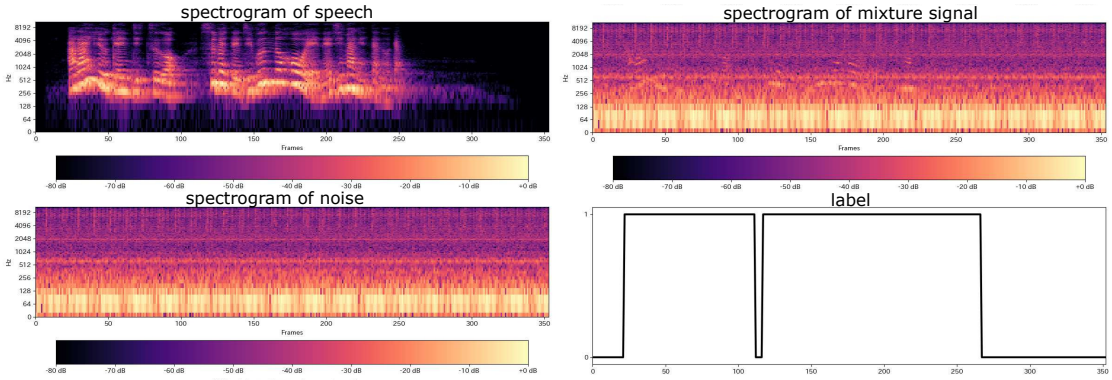

Figure 1. An example of spectrograms of speech, ego noise (left) and mixture signals (right upper) where SNR=-20 dB. Ground truth is shown in the right bottom.

CNN has achieved comparable performance with LSTM in modeling time series [10, 14], which attracts a lot of attention recently.

Similar to other classification tasks, deep learning-based methods have accomplished significant improvement in VAD tasks comparing to conventional methods, and many successful systems have been developed and reported. However, to our best knowledge, the effectiveness of these systems was only confirmed under noisy environments where SNRs were about [-5, 30] dB. There has been no investigation on performing VAD under high-level noisy situations where SNRs are lower than -5 dB. In this paper, we mainly provide two contributions. First, we propose a VAD system that uses gated CNN [14] to construct the VAD classifier together with a temporal smoothing as post-process. We focus on small network architectures to prevent high computational cost, which could be a severe problem when converting an existing method into a real-time system. This is the main difference between the proposed system and the one proposed in [10], which uses 36 layers and residual structures. Second, we investigate VAD methods in high-level noisy environments where SNRs are about [-30, -5] dB and show the limitation of the MLP-based method and the proposed method in such severe situations. Specifically, we evaluate the proposed method with speeches recorded by a hose-shaped rescue robot, which are critically interrupted by the self-created ego noise of the robot. Fig. 1 shows an example of spectrograms of the ego noise and the mixture signal whose SNR is -20 dB.

2 MLP-based VAD

Let $x(t)$ denotes a mixture signal (i.e, observation) consisting of a target speech $s(t)$ and a noise signal $n(t)$, and $x(\omega, f)$, $s(\omega, f)$ and $n(\omega, f)$ denote short-time Fourier transform (STFT) representations of the mixture, speech, and noise signal, respectively. Here $t = \{1, 2, \ldots, T\}$ denotes the discrete time index in time domain, and $\omega = \{1, 2, \ldots, \Omega\}$ and $f = \{1, 2, \ldots, F\}$ denote the frequency and frame indices in STFT domain, respectively. A VAD problem can be formulated as a binomial classification problem that classifies a segment of signal $\{x(t), \ldots, x(t+T)\}$ into speech and non-speech groups on the basis of a set of input features $\tilde{x}(\omega, f)$ extracted from the signal segment:

$$v(f) = \begin{cases} 1 & (\text{speech}), \\ 0 & (\text{non-speech}). \end{cases} \tag{1}$$

Here $v(f)$ denotes the ground truth at f-th frame. There are various acoustic features available for VAD, including magnitude/power spectrum, mel-frequency cepstral coefficients (MFCC). Considering that a neural network is able to serve as a further feature extractor to extract more task-effective features, in this paper, we use magnitude spectra of the observation $\tilde{x}(\omega, f) = |x(\omega, f)|$ as acoustic features.

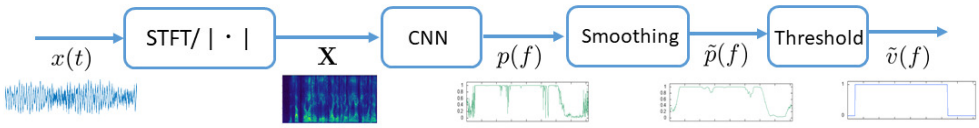

Figure 2. Flowchart of proposed method.

In an MLP-based VAD system, in order to capture the long-term dependencies in acoustic time series, inputs of a neural network in both of training and test stages are segmental features defined as $\mathrm{X(f)} = [\tilde{\mathrm{x}}(\mathrm{f}-\mathrm{M}+1)^{\mathrm{T}}, \tilde{\mathrm{x}}(\mathrm{f}-\mathrm{M}+2)^{\mathrm{T}}, \ldots, \tilde{\mathrm{x}}(\mathrm{f}-1)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{\Omega \times \mathrm{M}}$, which are concatenated feature vectors that combine the feature extracted from the current frame $\tilde{\mathrm{x}}(f) = [\tilde{x}(1,f), \tilde{x}(2,f), \ldots, \tilde{x}(\Omega,f)]^{\mathrm{T}}$ and those extracted from the previous $(M-1)$ frames. Since temporal dependency structures involve important information for analyzing time series such speech signals, it is expected that a more accurate classification system could be obtained with segmental features combining more feature vectors. However, it becomes challenging to train a well-generalized model when the dimension of input data increases, which results in a dramatic increase in the parameter number.

3 Proposed VAD system using gated CNN

RNN, in particular, LSTM, is a natural choice for modeling time series data since the recurrent connection architectures allow the networks to make a prediction with the entire input time series. However, the deeper the network architecture becomes, the more challenging its training becomes. Furthermore, it is difficult to employ parallel implementations for RNNs; thus, the training and prediction processing become computationally demanding. Motivated by the recent success achieved by CNN in language modeling and speech synthesis, and the merits of CNN that it is practically much easier to train and well suited to parallel implementation, in this paper, we propose a VAD classifier using CNN-based neural networks. The VAD decision is finally made by thresholding a smoothed predicted speech probability series, which is the output of the classifier. The flowchart of the proposed method is shown in Fig. 2.

3.1 CNN-based VAD Classifier

Considering the fact that spectrograms of audio signals have region dependency (i.e. they have different frequency structures in voiced and unvoiced segments), in particular, we employ gated CNN to model a VAD classifier $\mathbf{P} = \mathscr{F}(\mathbf{X})$, where $\mathscr{F}(\cdot)$ denotes a nonlinear function modeled with CNN, $\mathbf{X} = \{\tilde{x}(\omega,f)\}_{\omega,f}$ denotes magnitude spectrograms of mixture signals, and $\mathbf{P} = \{p(d,f)\}_{d,f} \in \mathbb{R}^{2\times F}$ denotes probability series of speech and non-speech, respectively. Note that $\mathbf{P}$ satisfies the constraint that $\sum_d p(d,f) = 1$. By using $\mathbf{H}_{l-1}$ to denote the output of the $(l-1)$-th layer, the output of the l-th layer $\mathbf{H}_l$ of a gated CNN is given as

$$\mathbf{H}_l = (\mathbf{H}_{l-1} * \mathbf{W}_l^{\mathrm{f}} + \mathbf{b}_l^{\mathrm{f}}) \otimes \sigma(\mathbf{H}_{l-1} * \mathbf{W}_l^{\mathrm{g}} + \mathbf{b}_l^{\mathrm{g}}), \tag{2}$$

where $\mathbf{W}_l^{\mathrm{f}}$, $\mathbf{W}_l^{\mathrm{g}}$, $\mathbf{b}_l^{\mathrm{f}}$ and $\mathbf{b}_g^{\mathrm{g}}$ are weight and bias parameters of the l-th layer, $\otimes$ denotes the element-wise multiplication and σ is the sigmoid function. Here, a gated linear unit (GLU) represented as the second term of (2) is used as a nonlinear activation function, which is the main difference between a gated CNN and a regular CNN layer. Similar to LSTM, GLU is a data-driven gate, which plays the role of controlling the information passed on in the hierarchy. Owing to this particular mechanism, it allows us to capture long-range context dependencies efficiently by deepening the layers without suffering from the vanishing gradient problem as well as apply different filters to different regions in a data-driven manner. Moreover, to capture the dependency of frequencies, we use 1-dimensional convolution, where the frequency dimension is regarded as the channel dimension, and an input spectrogram is convolved with a $(1,k_\tau)$ filter. Here k_τ is the filter width in the frame dimension. Dilated convolution [15] is used to efficiently obtain wider receptive field with fewer parameters by convolving a larger filter derived from the original filter with dilating zeros. Details of the network architecture we used in experiments are shown in Fig. 3.

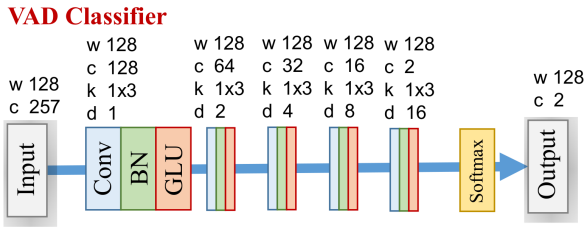

Figure 3. Network architecture of proposed VAD classifierThe inputs and outputs are 1-dimensional data. "w", "c", "k" and "d" denote the width, channel number, kernel size and dilation number, respectively. "Conv", "BN", and "GLU" denote 1-dimensional convolution, batch normalization and gated linear unit, respectively.

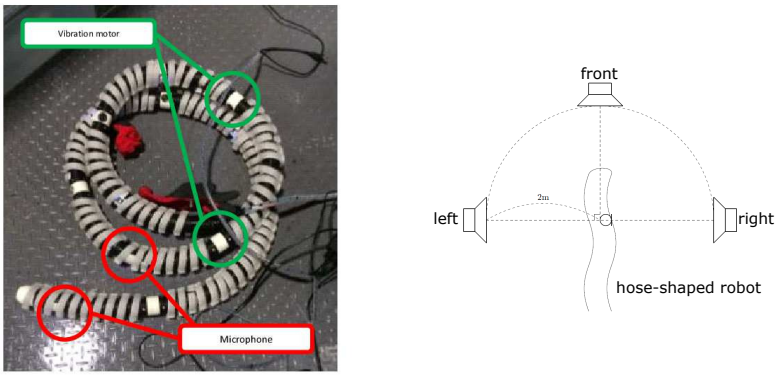

Figure 4. Hose-shaped rescue robot (left) and configuration of recording (right).

3.2 Post-processing: smoothing

After VAD classifier, we perform a temporal smoothing to the output probability series of speech to lessen the sudden changes of the decision since speech generally remains for a given period after a speech frame has been detected. The smoothing is applied with a rectangular window having $2L+1$ window length

$$\hat{p}(1,f) = \frac{1}{2L+1} \sum_{l=f-L}^{f+L} p(1,l). \tag{3}$$

Then the decision $\hat{v}(f)$ is made by thresholding the smoothed probability series $\hat{p}(1,f)$ with a manually set threshold θ. Frames with probability larger than the threshold are recognized as voiced frames otherwise unvoiced frames.

4 EXPERIMENTS

To evaluate the proposed system and investigate the limitation of the existing MLP-based method under high-level noisy environments, we conducted experiments where both of the systems were used to detect voice regions in noisy signals recorded by a hose-shaped rescue robot.

4.1 Hose-shaped rescue robot

The hose-shaped rescue robot is one of the robots developed by ImPACT project [4] for search and rescue operations during large-scale disasters such as earthquakes. This robot is long and slim like a snake, as shown in the left of Fig. 4, which allows it to investigate narrow spaces where are impossible for a human to enter. With the microphones attached around the body of the robot, voices of disaster victims can be captured so that

Table 1. Details of training and test datasets.

	Training data (SIM)	**Test data (SIM)**	**Test data (REC)**
Directions of arrival	front, right	left	left
Speakers	2 male and 2 female (fkn, fks, mho, mht)	1 male and 1 female (ftk, mmy)	1 male and 2 female (fym, mae, msh)
Positions of robot	4 types	2 types	1 type
Vibration levels	2 types	2 types	3 types including no vibration
Number of frames	about 13.06 millions	about 1.55 millions	7512

rescuers can localize the positions of victims and master the conditions of victims. However, it is challenging to capture clear voices of victims since the self-created ego noise is always existing, involving the driving sound of the vibration motors, the fricative sound generated between the cilia and floor and the noise generated by microphone vibration. Moreover, the energy of ego noise is notably higher than those of voices because of the close distance between the microphones and noise sources and a relatively low-energy sound of a person seeking help. This means SNRs of speech signals recorded by the robot are generally low.

4.2 Datasets and experimental conditions

We created two datasets for experiments, namely, simulated dataset (SIM) and recorded dataset (REC), using the hose-shaped rescue robot and ATR503 speech database [16]. The data in the previous one were generated by first convolving dry speech sources and measured room impulse responses (RIR), then adding them into the recorded ego noise signals of the robot with SNR=$\{-30,-25,-20,-15,-10,-5\}$ dB. The room configuration for measuring RIRs is shown in the right figure in Fig. 4, and the ego noise signals were recorded with 2 types of vibration levels and 4 different positions of the robot. The ground truth was obtained by performing frequency domain power-based VAD to convolved speech signals following a hangover process. We divided SIM dataset into two sub-datasets as training dataset and test dataset so that the speakers, direction of arrival were different between the sub-datasets. REC dataset was generated by recording speech signals from loudspeakers using the microphones attached to the robot with 2 different vibration levels as well as without vibration. The sound levels of speech signals were set at $\{30,40,50,70\}$ dB so that the SNRs of the recorded signals were about [-30, -20] dB. The ground truth of this dataset was manually labeled. Those data with uncertain labels were excluded in the experiment. Note that the REC dataset was only used for the test. More details of training and test datasets are shown in Table 1.

All signals were recorded or generated using the single microphone at the front of the robot with a sampling rate of 16 kHz. The magnitude spectrograms were calculated with window length and window shift set at 32 ms and 16 ms, respectively. The temporal length for smoothing was set at 11, i.e., $L = 5$. The threshold was set at $\theta = 0.5$. We compared two varieties of the proposed systems, namely, gated CNN-based VAD classifier without/with the post-processing, with two MLP-based VAD classifiers with the post-processing. The inputs of MLP-based classifiers were segmental features combining 7 frames. Four criteria were used for evaluation. Namely, 1) Root mean square error (RMSE) of the probability series of speech between the ground truth and estimated value; 2) Percentage of the accurately detected frames (accuracy; ACC); 3) False acceptance rate (FAR); and 4) False rejection rate (FRR).

4.3 Results

Table 2 shows the results achieved with MLP-based classifiers with the different number of layers, the proposed gated CNN-based classifier without post-processing, and the proposed system. Note that RMSE results were calculated by taking an average of all the data without regard to SNRs and the other criteria were calculated based on the whole datasets. Although MLP-based systems obtained adequate results in terms of accuracy and FRR, FAR and some examples shown in the left of Fig. 6 show that MLP-based systems tended to predict all

Table 2. VAD results of SIM and REC test datasets.

	SIM				REC			
	RMSE	ACC	FAR	FRR	RMSE	ACC	FAR	FRR
MLP(3 layers)	0.432	0.700	0.866	0.068	0.60	0.700	0.868	**0.027**
MLP(5 layers, dropout)	0.439	0.692	0.766	0.119	0.466	0.711	0.785	0.050
Gated CNN	0.352	0.848	0.326	0.077	0.451	0.767	**0.534**	0.084
Gated CNN+smoothing	**0.318**	**0.863**	**0.325**	**0.056**	**0.424**	**0.770**	0.542	0.077

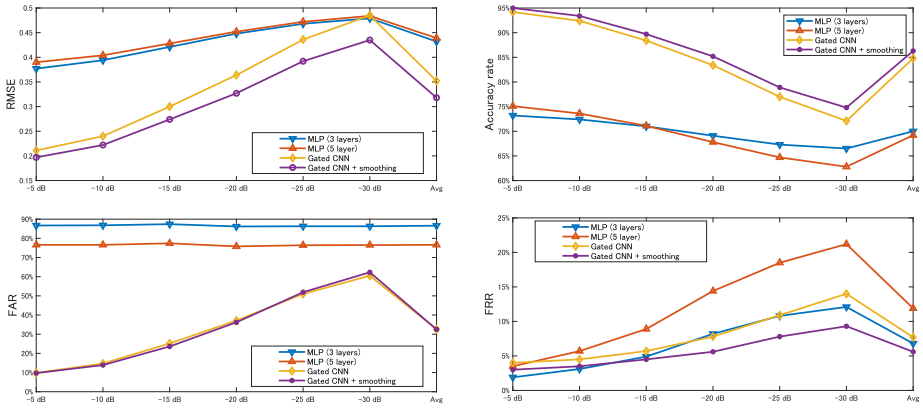

Figure 5. Results in terms of RMSE, ACC, FAR, and FRR under environments with different SNRs.

the frames as speech regions. In contrast, the proposed gated CNN-based systems achieved better performances in terms of all the criteria with SIM dataset. The detection accuracy was about 85% even in high-level noise environments. The results also show that smoothing is effective in improving the detection accuracy.

The performances achieved in various SNRs are shown in Fig. 5. The performances tended to decrease when SNRs decreased, which were reasonable. However, from Fig. 5 and the results obtained with REC dataset shown in Table 2, the results show that the VAD performances under situations with SNRs lower than -20 dB decreased rapidly, and the accuracy was lower than 80%, which was unsatisfactory. There is still a massive space in improving the detection accuracy under environments where SNRs are under -20 dB.

5 CONCLUSIONS

In this paper, we proposed a VAD system that adopts a neural network constructed with gated CNN as speech/non-speech classifier and applies smoothing as post-processing. We focused on high-level noisy environments where SNRs are lower than -5 dB and investigated the VAD performances of the proposed system and an existing MLP-based VAD system under such severe situations. The experimental results showed that 1) the proposed gated CNN-based VAD system outperformed MLP-based VAD systems; 2) utilizing temporal smoothing as post-processing was effective in improving the VAD accuracy; 3) the proposed method achieved satisfactory results under the situations where SNRs were about [-20, -5] dB whereas the performances under the situations with SNRs lower than -20 dB were unsatisfactory. To break through the limitation of VAD, more efforts need to be paid in developing systems for environments with extremely low SNRs.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 19H04131, SECOM Science and Technology Foundation, and Strategic Core Technology Advancement Program (Supporting Industry Program).

REFERENCES

[1] Ramirez, J., Górriz, J. M., Segura, J. C. Voice activity detection. fundamentals and speech recognition system robustness. Robust speech recognition and understanding, InTech, 2007.

[2] Verteletskaya, E., Simak, B. Speech distortion minimized noise reduction algorithm. Proceedings of the World Congress on Engineering and Computer Science, Vol. 1, 2010, pp. 20–22.

[3] Sahidullah, M., Saha, G. Comparison of speech activity detection techniques for speaker recognition. preprint arXiv:1210.0297, 2012.

[4] Impulsive Paradigm Change through Distributed Technologies Program (ImPACT), http://www.jst.go.jp/impact/en/program/07.html (2019/05/20).

[5] Embodied Audition for RobotS (EARS), https://robot-ears.eu/ (2019/05/20).

[6] Zhang, X. L., Wu, J. Deep belief networks based voice activity detection. IEEE Trans. on Audio, Speech, and Language Processing, Vol. 21(4), 2013, pp. 697–710.

[7] Tashev, I., Mirsamadi, S. DNN-based Causal Voice Activity Detector. In Information Theory and Applications Workshop, 2016.

[8] Kang, T. G., Kim, N. S. DNN-based voice activity detection with multi-task learning. IEICE Trans. on Information and Systems, Vol. 99(2), 2016, pp. 550–553.

[9] Sertsi, P., Boonkla, S., Chunwijitra, V., Kurpukdee, N., Wutiwiwatchai, C. Robust voice activity detection based on LSTM recurrent neural networks and modulation spectrum. in 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 2017, pp. 342–346.

[10] Chang, S. Y., Li, B., Simko, G., Sainath, T. N., Tripathi, A., van den Oord, A., Vinyals, O. Temporal Modeling Using Dilated Convolution and Gating for Voice-Activity-Detection. in In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 5549–5553.

[11] Sehgal, A., Kehtarnavaz, N. A Convolutional Neural Network Smartphone App for Real-Time Voice Activity Detection. IEEE Access, Vol. 6, 2018, pp. 9017–9026.

[12] Hochreiter, S., Schmidhuber, J. Long short-term memory. Neural computation, Vol. 9(8), 1997, pp. 1735–1780.

[13] Tai, K.S., Socher, R., Manning, C.D. Improved semantic representations from tree-structured long short-term memory networks. preprint arXiv:1503.00075, 2015.

[14] Dauphin, Y. N., Fan, A., Auli, M., Grangier, D. Language modeling with gated convolutional networks. in Proceedings of the 34th International Conference on Machine Learning, Vol. 70, 2017, pp. 933–941.

[15] Yu, F., Koltun, V. Multi-scale context aggregation by dilated convolutions. preprint arXiv:1511.07122, 2015.

[16] Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H., Shikano, K. ATR Japanese speech database as a tool of speech recognition and synthesis. Speech Communication, Vol. 9, 1990, pp. 357–363.

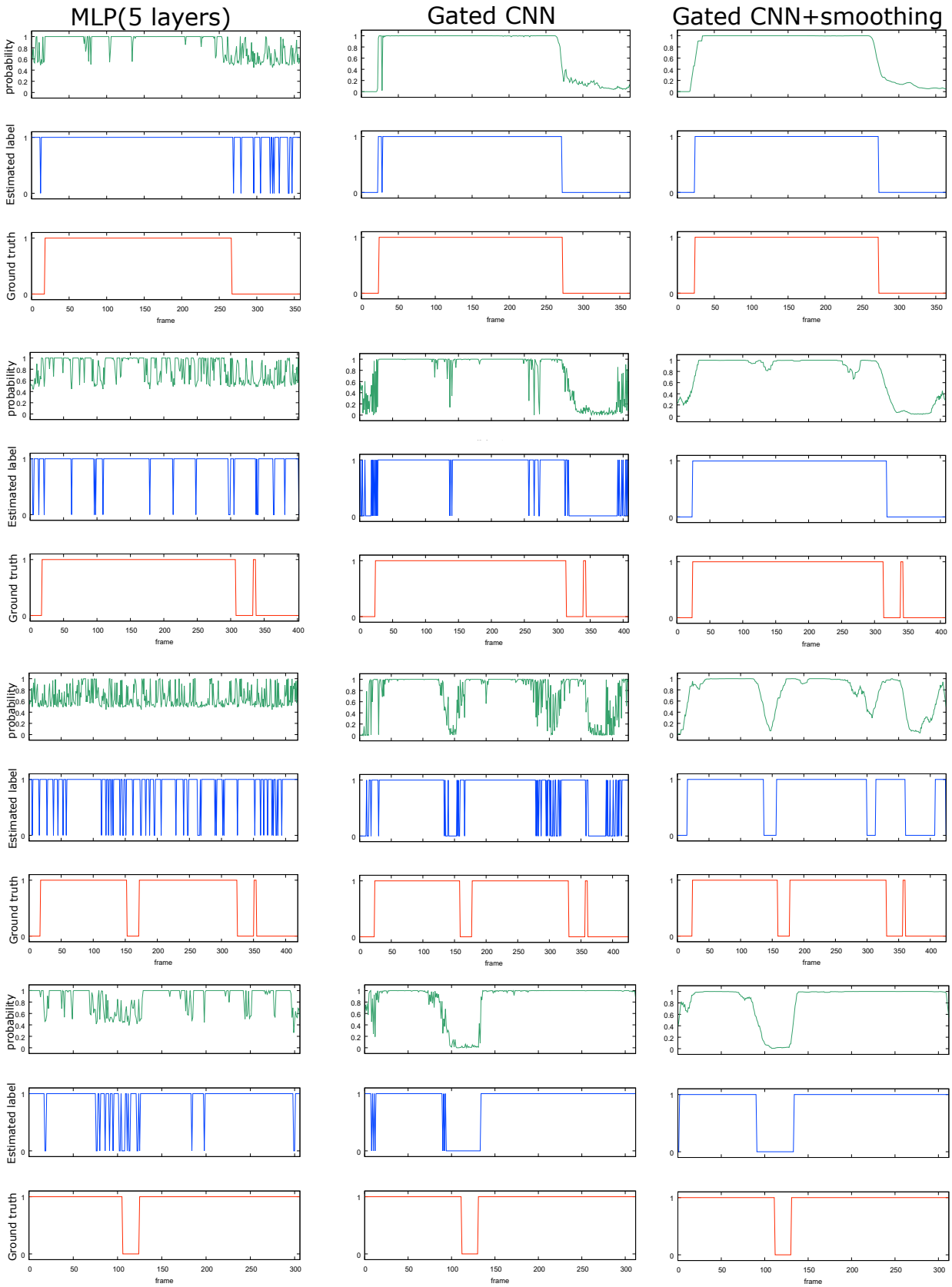

Figure 6. Examples of detected speech regions. From top to bottom are data in SIM dataset whose SNRs were −5 dB, −15 dB, and −30 dB, respectively, and an example from REC dataset.