Robust Speech Command Recognition using Label-Driven Time-Frequency Masking

Anonymous Author(s)
Affiliation
Address
email

Abstract

Speech enhancement driven robust Automatic Speech Recognition (ASR) systems typically require a parallel corpus with noisy and clean speech utterances for training. Moreover, many studies have reported that such front-ends, even though improve speech quality, do not always improve the recognition performance. On the other hand, multi-condition training of ASR systems provides little visualization or interpretability capabilities of how these systems achieve robustness. In this paper, we propose a novel neural architecture with unified enhancement and sequence classification block, that is trained in an end-to-end manner only using noisy speech without having information of clean speech. The enhancement block is a fully convolutional network that is designed to perform Time Frequency (T-F) masking like operation, followed by an LSTM sequence classification block. The T-F masking formulation enables visualization of learned mask and helps us to visualize the T-F points important for classification of a speech command. Experiments performed on Google Speech Command dataset show that our proposed network achieves better results than the baseline model without an enhancement front-end.

1 Introduction

Performance degradation of Deep Neural Network (DNN) based Automatic Speech Recognition (ASR) systems in the presence of channel distortions, reverberation, and additive noise is still a well known issue [1]. There are two major paradigms to achieve robustness to various noise conditions, (a) use of model adaptation techniques to achieve robustness against various degradation conditions [1, 2, 3, 4], (b) use of enhancement front-end to map noisy speech features to clean features [5, 6, 7, 8, 9, 10, 11]. Model adaptation techniques majorly use representation power of DNNs to train the model with various degradation conditions. This approach is reported to work well in wide range of degradation conditions [2, 4], without using information of clean speech. However, they do not give much insights regarding their inner workings. Other popular approach to achieve robustness is to employ an enhancement front-end using De-noising Autoencoder (DAE) based on various DNN architectures such as DNN-DAE [12], Time-Delay Neural Network (TDNN)-DAE [8], Recurrent Neural Network (RNN)-based DAE [6, 7], or Time-Frequency (T-F) masking-based approaches to enhance the noisy signal [10, 9]. To train such front-ends, a parallel corpus containing noisy and clean speech pairs is required. However, it is reported that such front-ends do not always yield improvement in performance in unseen noise conditions [12].

In this paper, we propose a novel neural network architecture that can leverage advantages of both these approaches. We propose a network with an enhancement front-end block that has a T-F masking like formulation in such that it learns feature detectors to locate T-F regions important for classification. The output of this enhancement block is given to an LSTM-based sequence classification block.

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

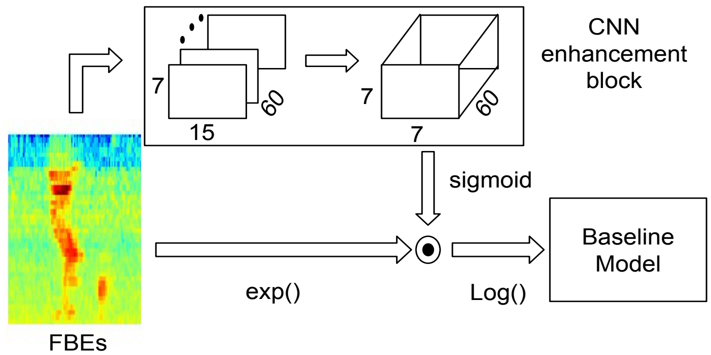

Figure 1: Architecture of proposed model that consist of a convolutional enhancement block and baseline LSTM classification block.

The entire network is jointly trained in an end-to-end manner using only noisy data with random parameter initialization and without providing the network any information about clean speech data.

The proposed architecture enables an easy visualization of the enhancement process by inspecting the T-F mask applied by the enhancement block and activation maps of convolution filters in the first layer of the enhancement block. This visualization gives insights on what T-F regions are important for classification of a speech command. Experiments on Google Speech Command dataset [13] demonstrate the effectiveness of the proposed model and its visualization capabilities.

2 Proposed model architecture

Our proposed model based on label-driven T-F masking is shown in Figure 1. It consist of two fully convolutional layers in the enhancement block. The output of convolutional block is then applied to the input T-F representation by treating the output of the convolutional block as a T-F mask. We constrain the mask values between $0 - 1$ by applying the sigmoid activation function at the output of the convolutional block. Here, the input is log-magnitude domain T-F representation such as log Mel-Filterbank Energies (FBEs). Mathematically, operations of the enhancement block can be summarized as follows:

$$Y(t, f) = log(exp(X(t, f)) \circ M(t, f)), \tag{1}$$

where $X(t, f)$ is the input T-F representation (e.g. FBEs), $M(t, f)$ is the T-F mask taken at the output of the enhancement block, and $Y(t, f)$ is the enhanced T-F representation.

An LSTM layer (referred as the baseline model) takes the enhanced T-F representation $Y(t, f)$ as an input and the final hidden state of the LSTM cell is then propagated to fully connected and softmax classification layer. The parameters of classification and enhancement block are optimized using the final hidden state of the LSTM layer. This enhancement block tries to enhance the entire T-F sequence, in contrast with the frame-level enhancement. The enhancement block, along with the baseline model is trained to maximize the target class probability. Hence, the T-F mask is learned in a manner that will increase correct classification probability.

3 Experiments and Results

3.1 Database description

We use Google Speech Command dataset for our experiments [13]. The database consists of 64,727 audio files, each of 1 second duration, and consisting of one spoken command. Each utterance is labelled with one of the possible 30 commands. The splits for train (80%), validation (10%), and test (10%) datasets are provided in the database. The dataset also provides background noise audio files with six types of noise. In the initial observation we found that the audio files were already containing

Table 1: Classification accuracy (%) of various models on validation and test dataset.

Model name	Validation	Test	Test (20 classes)
LSTM baseline	90.93	90.76	91.12
Direct enhancement	91.5	91.47	91.97
T-F masking enhancement	**92.92**	**92.9**	**93.24**

Table 2: Classification accuracy (%) of all the models on noisy test set. All the available noises in the database were added with 15 dB, 10 dB, and 5 dB SNR.

Test	LSTM baseline			T-F masking enhancement			Direct enhancement		
Noise	15 dB	10 dB	5 dB	15 dB	10 dB	5 dB	15 dB	10 dB	5 dB
running_tap	75.42	64.74	47.24	82.25	73.46	56.12	81.59	72.60	54.32
dude_miaowing	76.14	65.53	49.26	82.38	73.87	57.41	81.78	73.24	57.82
exercise_bike	76.40	64.30	42.83	83.70	73.91	54.00	81.95	71.60	49.74
doing_dishes	82.53	73.01	56.27	86.26	79.90	67.90	83.96	75.62	59.25
pink_noise	85.44	7.15	68.31	89.20	85.62	77.51	86.83	83.07	73.72
white_noise	72.07	58.26	35.48	79.75	68.01	45.94	79.41	67.90	44.92
Average	78.00	55.50	49.90	**83.92**	**75.80**	**59.81**	82.59	74.01	56.63

little noise. To evaluate the robustness of the proposed model and visualization of enhancement process, we add the provided noises at 15 dB, 10 dB and 5 dB SNR to test utterances.

3.2 Model architectures and results

In our proposed model, the first convolution layer in the enhancement block had 60 convolutional filters of size 15×7 followed by ReLU activation. The second convolution layer had one convolutional filter of size $7 \times 7 \times 60$ followed by sigmoid activation. The number of filters and filter dimensions were optimized on validation set. The model was trained to jointly optimize the parameters of enhancement block and the baseline model. Results of this model are tabulated under the label “T-F masking enhancement".

The baseline model has an LSTM layer with 128 units and ReLU activation. The LSTM layer was followed by a fully connected layer with 128 units ReLU activation. The output layer had 30 softmax units for 30 class classification. Input to our models were 40 dimensional FBEs of 1 second utterance extracted by taking the frames of 25 ms with 10 ms overlap. The model was trained using cross-entropy objective and ADAM optimizer with learning rate of 0.001 for 10 epochs and model that gives the best accuracy on validation dataset was used for testing.

We train one more model to compare results with T-F masking based formulation. In this model we use the same enhancement block as used earlier. However, rather than treating output of the enhancement block as T-F mask, we treat the output of the enhancement block as an enhanced T-F representation and directly feed it to the baseline model. The model parameters and training scheme for the model with the direct enhancement block same as the model with T-F masking enhancement. In this case we used linear activation instead of sigmoid activation. Results of this model are tabulated under the label “Direct enhancement".

Results for all the models are shown in Table 1. The baseline results are better than the CNN and Capsule Network models trained on the same database with same training/testing condition [14]. On 20 commands evaluation our LSTM baseline (91.12 %) performed better than CNN (77.9 %) and CapsNet (87.3 %) for 20-class decoding [14]. Model with direct enhancement gave improvement over baseline system in both validation (91.5%) and test(91.47%) dataset. While the proposed model gave the best results on the original validation (92.92%) and test (92.9%) dataset. Table 2 shows the results of evaluating the trained models on noisy dataset. Performance of LSTM baseline degraded greatly in the presence of noise. By employing CNN enhancement block, the results on noisy database improved significantly. While in this case also proposed model with label-driven T-F masking enhancement gave the best results.

3.3 Visualizing the enhancement process

To visualize the enhancement process we show input, output, and T-F mask applied by enhancement block in Figure 2-I for an original test utterance as well as noisy versions of it. It can be observed

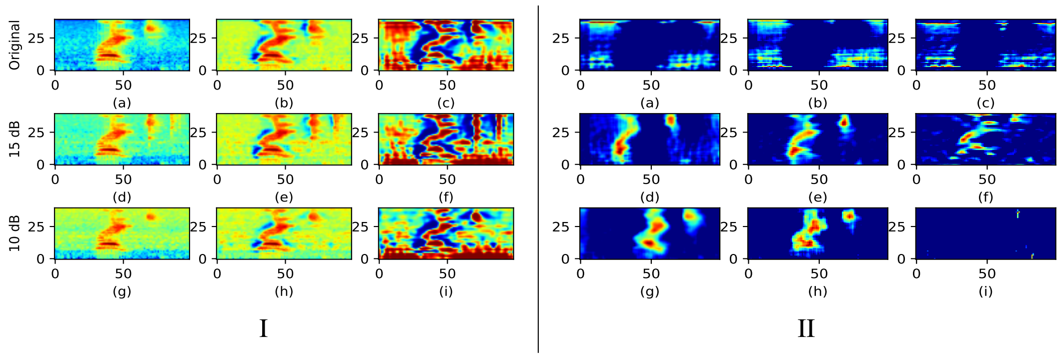

Figure 2: I: Visualization of the enhancement process applied by the propose model. II: Visualization of the activation maps corresponding to some of the filters learned by proposed model. (Vertical axis corresponds to frequency and horizontal axis is the time axis in both figures.)

that the original utterance shown in Figure 2-I (a) taken from test dataset already has some noise. The enhancement block gives the output shown in Figure 2-I (b). T-F mask learned by the enhancement block is shown in figure 2-I (c). Figure 2-I (d)-(f) and (g)-(i) show the similar plots for the same utterance with 15 dB and 10 dB additive noise of type running_tap, respectively.

Visual inspection of the enhanced representation and T-F mask suggests that the enhancement block tries to achieve two things : (1) finding out the important T-F points in the input T-F representation where acoustic information about spoken word is present, (2) finding out the boundary between T-F regions where acoustic information about spoken word is present and silence regions. Additionally, the T-F masks for original as well as noisy utterances are fairly similar except for some T-F regions with very less SNR.

Figure 2-II shows the activation maps of 9 selected filters (out of 60 filters) for the utterance in Figure 2-I (a) as input. Figure 2-II (a)-(c) suggest that the underlying filters tries to locate the T-F regions where spoken command is present. Figure 2-II (d)-(f) show that the underlying filters are locating the boundaries between acoustic information of spoken command and non speech T-F regions. While Figure 2-II (g)-(h) show the activation of filters that are finding out the important T-F points in the area where acoustic information of spoken command is present. Figure 2-II (i) shows the activation of filter that is not significant for classifying the utterance. We found that majority of the activation maps resembled Figure 2-II (a)-(c), i.e. trying to locate the T-F regions where spoken command is present. Other significant number of filters resembled the filters shown in Figure 2-II (d)-(f).

These visualizations suggest that enhancement for robust speech classification is different than traditional enhancement. While traditional enhancement front-ends try to remove or suppress noise, the label driven enhancement approach focuses on finding out important regions in T-F representation that are significant to increase the correct classification probability.

4 Summary and Conclusions

In this paper, we propose a novel neural network architecture with a fully convolutional enhancement block and LSTM-based classification block for robust speech command recognition. We trained our model directly on noisy data to jointly train the enhancement and classification block. The enhancement block had T-F masking like formulation for enhancement purpose. Our proposed model gave significantly better classification accuracy for both original and noisy test set. The visualization of enhancement process for improving classification accuracy gave significant insights on the working of the proposed network. We observed that instead of removing or suppressing noise present in the noisy T-F representation, the enhancement block locates the important regions in the input T-F representation.

References

[1] Vikramjit Mitra, Horacio Franco, Richard M Stern, Julien Van Hout, Luciana Ferrer, Martin Graciarena, Wen Wang, Dimitra Vergyri, Abeer Alwan, and John HL Hansen. Robust features in deep-learning-based speech recognition. In *New Era for Robust Speech Recognition*, pages 187–217. Springer, 2017.

[2] Michael L Seltzer, Dong Yu, and Yongqiang Wang. An investigation of deep neural networks for noise robust speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7398–7402. IEEE, 2013.

[3] Vikramjit Mitra and Horacio Franco. Leveraging deep neural network activation entropy to cope with unseen data in speech recognition. *arXiv preprint arXiv:1708.09516*, 2017.

[4] Yanmin Qian, Mengxiao Bi, Tian Tan, and Kai Yu. Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(12):2263–2276, 2016.

[5] Killian Janod, Mohamed Morchid, Richard Dufour, Georges Linares, and Renato De Mori. Denoised bottleneck features from deep autoencoders for telephone conversation analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(9):1809–1820, 2017.

[6] Andrew L Maas, Quoc V Le, Tyler M O'Neil, Oriol Vinyals, Patrick Nguyen, and Andrew Y Ng. Recurrent neural networks for noise reduction in robust asr. In *Proc. Interspeech*, 2012.

[7] Erik Marchi, Fabio Vesperini, Florian Eyben, Stefano Squartini, and Björn Schuller. A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 1996–2000. IEEE, 2015.

[8] Cong-Thanh Do and Yannis Stylianou. Improved automatic speech recognition using subband temporal envelope features and time-delay neural network denoising autoencoder. *Proc. Interspeech 2017*, pages 3832–3836, 2017.

[9] Zhong-Qiu Wang and DeLiang Wang. A joint training framework for robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):796–806, 2016.

[10] Arun Narayanan and DeLiang Wang. Joint noise adaptive training for robust automatic speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2504–2508. IEEE, 2014.

[11] Yanmin Qian, Maofan Yin, Yongbin You, and Kai Yu. Multi-task joint-learning of deep neural networks for robust speech recognition. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 310–316. IEEE, 2015.

[12] Jun Du, Qing Wang, Tian Gao, Yong Xu, Li-Rong Dai, and Chin-Hui Lee. Robust speech recognition with speech enhanced deep neural networks. In *Proc. Interspeech*, 2014.

[13] Pete Warden. Speech commands: A public dataset for single-word speech recognition. *Dataset available from http://download.tensorflow.org/data/speech_commands_v0*, 1, 2017.

[14] Jaesung Bae and Dae-Shik Kim. End-to-end speech command recognition with capsule network. *Proc. Interspeech 2018*, pages 776–780, 2018.