Robust Multichannel End-to-End Speech Recognition Based on Multi-Output Densenet

Zheng Chonghui[1] Shinozaki Takahiro[1]

Abstract: We propose DenseNet-based robust multi-channel speech recognition in noisy environments. Recently, DenseNet has shown its efficiency in mask generation for single-channel speech enhancement tasks. In this paper, we propose a multi-output DenseNet for the multi-channel situation. In the structure, beamforming frontend and acoustic model back-end share a same DenseNet block, which can generate high-quality masks for beamforming, and also perform feature extraction at the acoustic model part. At the training step, beamforming front-end and acoustic model back-end are trained jointly with the ASR target. The experimental result shows that lower character error rate and word error rate are obtained by the proposed method compared with conventional BLSTM based beamformer front-end.

Keywords: multichannel speech recognition, end-to-end training, DenseNet, beamforming

1. Introduction

Recently, multichannel speech recognition is getting more and more attention. A neural-network-based beamformer as front-end becomes a popular option to handle noisy speech. In [1][2][3], a BLSTM network is trained to generate masks of time-frequency(TF) bins to estimate signal statistics for beamformers like Generalized Eigenvalue(GEV) and Minimum Variance Distortionless Response(MVDR). In [4][5][6], joint optimization of the neural beamformer front-end and ASR back-end has also been considered.

In this paper, we focus on joint training of front-end beamforming and back-end ASR and propose a new method to jointly train the two parts based on DenseNet.

2. DenseNet Architecture

DenseNet is reported to have excellent performance in image recognition tasks[7] and give improvement to single-channel speech enhancement tasks[8]. The basic idea of DenseNet is to improve the information flow between layers by concatenating all preceding layers as $X_L = H_L([X_0, X_1, ..., X_{L-1}])$, where [...] denotes the concatenation and the composite functions $H_L()$ typically consists of a normalization layer, an activation layer and a convolutional layer with k feature maps. A transition layer, which consists of a normalization layer, a convolution layer and a pool layer, is used between DenseBlocks to reduce the dimension and number of the feature maps passed on to the following layers. Such a dense connected structure enables all layers to receive the gradient directly and also reuse features computed in preceding layers. Figure 1 illustrates the DenseBlock.

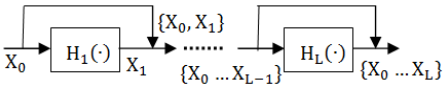

Fig. 1 DenseBlock Structure.

3. DenseNet-based Front-end Beamforming

The process of front-end beamforming is similar to the one in ESPnet[9]. The difference is that we use DenseNet for mask estimation instead of BLSTM. The network outputs the time-frequency masks as follows:

$$Z_c = DenseNet(x_{t,f,c}{}_{t=1}^{T})$$

$$m_{t,c}^{S} = sigmoid(linear_S(Z_c)$$

$$m_{t,c}^{N} = sigmoid(linear_N(Z_c)$$

where $x_t, f, c \in C$ is an STFT coefficient of c-th channel noisy signal at a time-frequency bin(t,f), Z is the output sequence of DenseNet, while $linear_S$ and $linear_N$ are used to obtain speech noise masks respectively. These masks are averaged over channels and used to compute the power spectral density (PSD) matrices of speech Φ_S and noise Φ_N at frequency bin b as follows:

$$\Phi_S(f) = \frac{1}{\sum_{t=1}^{T} m_{t,f}^{S}} \sum_{t=1}^{T} m_{t,f}^{S} x_{t,f} x_{t,f}^{H}$$

$$\Phi_N(f) = \frac{1}{\sum_{t=1}^{T} m_{t,f}^{N}} \sum_{t=1}^{T} m_{t,f}^{N} x_{t,f} x_{t,f}^{H}$$

From the speech PSD and noise PSD, MVDR beamforming filter is computed as follows:

$$g(f) = \frac{\Phi^N(f)^{-1}\Phi^S(f)}{Tr(\Phi^N(f)^{-1}\Phi^S(f))}u$$

[1] Tokyo Institute of Technology, Tokyo, Japan
www.ts.ip.titech.ac.jp

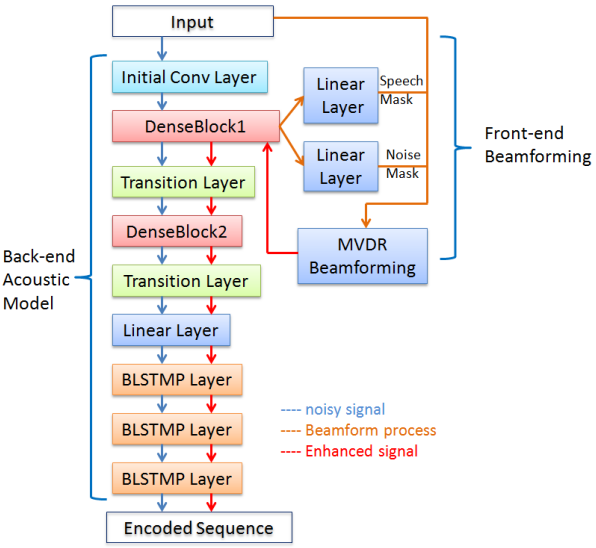

Fig. 2 Multi-output DenseNet Structure.

where $u \in [0, 1]$ is a weight vector to microphone channels, and $Tr(\cdot)$ is the matrix trace operation.

Finally, the enhanced STFT signal is obtained as follows:

$$\hat{x}_{t,f} = \sum_{c=1}^{C} g_{t,f,c} x_{t,f,c}$$

4. Multi-output DenseNet

Our proposed multi-output DenseNet network consists of two DenseBlocks and is shown in Figure 2. The size of the input tensor to the network is provided as B, C, T, F, with B, C, T, F being batch size, number of microphone channels, numbers of frame and frequency respectively. The initial convolutional layer transfers microphone channel axis to the initial number of feature maps. The first DenseBlock is shared by the front-end beamforming and back-end acoustic model, while the second DenseBlock is used only for further feature extraction for the acoustic model part. Each of the DenseBlock consists of 3 sets of one batch normalization layer, ReLU activation layer and convolutional layer with the growth rate of 16 and therefore add 3*16 feature maps to the input.

For the front-end beamforming part, the output sequence is obtained from the first DenseBlock and passed on to two linear layers for mask generation. Note that the linear layers are apart from the second DenseBlock and are used for generating the masks for speech and noise separately. After the enhanced signal is obtained using masks and noisy input STFT, it is passed back to the first DenseBlock for feature extraction. To make the input size uniformed to the first DenseBlock, we add a C=1 axis to the single-channel enhanced STFT signal.

For the acoustic model part, two DenseBlocks are used for feature extraction. A transition layer parametrized with 0.5 is connected to each DenseBlock to reduce the dimension of the feature map to half. Then a linear layer is used to transit the DenseNet subnet to the BLSTMP subnet. Then following 3 BLSTMP layers generate the final encoded sequence. During the training step, we also adopt multi-condition training strategy, the noisy signal will randomly skip the front-end beamforming part and be passed straight to the acoustic model part.

Table 1 Details of Chime-4 Dataset)

	Train Set	Development Set	Evaluation Set
Hour_real	3	2.9	2.2
Hour_simu	15	2.9	2.2
Speaker_real	4	4	4
Speaker_simu	83	4	4
Channel	6	6	6

Table 2 CER and WER Results of Evaluation Set

	CER		WER	
	real	simu	real	simu
Baseline	11.4	8.9	20.7	16.3
DenseNet Backend	11.0	8.9	19.8	16.4
DenseNet Frontend	11.1	8.8	20.6	16.5
DenseNet Frontend+Backend	10.9	9.0	20.9	17.6
Multi-output DenseNet	**10.4**	**8.3**	**19.6**	**15.9**

5. Experimental Evaluation

We train and evaluate our proposed method on the Chime-4 dataset[10]. The details of the dataset is shown in table 1.

ESPnet toolkit[9] is used for end-to-end speech recognition, which is based on a hybrid combination of connectionist temporal classification(CTC) and attention-based encoder-decoder model. In the Chime-4 baseline of ESPnet, BLSTMP is used as frontend mask estimation and VGGBLSTMP is used as encoder.The front-end beamforming and back-end acoustic model is trained jointly. For front-end beamforming and acoustic model, 80-dimensional log Mel filterbank energy is used as input feature, which is computed from 400-dimensional STFT coefficients with window length as 400, the number of window shift as 160, and window function as hanning. In the encoder, we use 3-layer BLSTMP with 1024 units and 1-layer LSTM with 1024 units in the decoder. The CTC-attention weight is fixed as 0.3. The word-based RNN language model is used. We trained the proposed model under the adadelta optimizer for 20 epochs.

The experiments compare Character Error Rate(CER) of the baseline and our multi-ouput DenseNet for the 6-channel track. The results are shown in table 2.

Our multi-output DenseNet with BLSTMP performs the best in both real and simulated condition and improved the Chime-4 baseline in terms of CER by 10.5% relative on real condition and 8.0% on the simulated condition, which suggests that our proposed multi-ouput DenseNet can generate high-quality masks in the front-end beamforming while having the ability to extract effective features from noisy and enhanced signal in the back-end acoustic model.

6. Summary

In this paper, we present a new model topology for handling front-end beamforming and back-end acoustic model based on multi-output DenseNet. Experimental results on the Chime-4 dataset shows that both CER and WER are reduced by our proposed model on both real and simulation condition.

References

[1] Erdogan, Hakan, et al. "Improved mvdr beamforming using single-channel mask prediction networks." Interspeech. 2016.
[2] Heymann, Jahn, Lukas Drude, and Reinhold Haeb-Umbach. "Neural

network based spectral mask estimation for acoustic beamforming." 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

[3] Pfeifenberger, Lukas, Matthias Zöhrer, and Franz Pernkopf. "DNN-based speech mask estimation for eigenvector beamforming." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[4] Ochiai, Tsubasa, et al. "Unified architecture for multichannel end-to-end speech recognition with neural beamforming." IEEE Journal of Selected Topics in Signal Processing 11.8 (2017): 1274-1288.

[5] Xiao, Xiong, et al. "On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[6] Heymann, Jahn, et al. "Beamnet: End-to-end training of a beamformer-supported multi-channel asr system." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[7] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[8] Du, Xingjian, et al. "End-to-End Model for Speech Enhancement by Consistent Spectrogram Masking." arXiv preprint arXiv:1901.00295 (2019).

[9] Watanabe, Shinji, et al. "Espnet: End-to-end speech processing toolkit." arXiv preprint arXiv:1804.00015 (2018).

[10] Barker, Jon, et al. "The third 'CHiME'speech separation and recognition challenge: Dataset, task and baselines." 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015.