Dynamic-attention based Encoder-decoder model for Speaker Extraction with Anchor speech

Hao Li, Xueliang Zhang, Guanglai Gao
College of Computer Science, Inner Mongolia University, China
E-mail: lihao@mail.imu.edu.cn cszxl@imu.edu.cn csggl@imu.edu.cn

***Abstract*—Speech plays an important role in human-computer interaction. For many real applications, an annoying problem is that speech is often degraded by interfering noise. Extracting target speech from background interference is a meaningful and challenging task, especially when interference is also human voice. This work addresses the problem of extracting target speaker from interfering speaker with a short piece of anchor speech which is used to obtain the target speaker identify. We propose a encoder-decoder neural network architecture. Specifically, the encoder transforms the anchor speech to a embedding which is used to represent the identity of target speaker. The decoder utilizes the speaker identity to extract the target speech from mixture. To make a acoustic-related speaker identity, The dynamic-attention mechanism is utilized to build a time-varying embedding for each frame of the mixture. Systematic evaluation indicates that our approach improves the quality of speaker extraction.**

Index Terms: speaker extraction, dynamic-attention, Encoder-decoder.

I. Introduction

Speech separation is to separate target speech from background interference, which is a meaningful work for many applications, such as mobile telecommunication, robust automatic speech recognition (ASR). Speaker separation is a special case of speech separation when interference is also speech from other speakers. Another case is speech denoising focusing on speech and non-speech separation. In general, speaker separation is more challenging than speech denoising.

Speaker separation has a wide range of application in many scenarios, e.g. , teleconference and meeting transcription, where simultaneous speaking is common. Several traditional multichannel signal processing methods can be applied for speaker separation, e. g. , beamforming [1] [2], independent component analysis (ICA) [3]. The effect of these methods is directly related to the number of microphones. To achieve a good performance, a large number of microphones are needed. In practice, single-microphone scenario is the most common case.

Recently, speech separation is treated as a supervised learning problem and has achieved great progress by using deep learning, especially for single-channel speech enhancement[4]. Learning-based algorithms are also invented for single-channel speaker separation including deep clustering [5], permutation-invariant training (PIT) [6] and deep attractor network [7]. These methods attempt to separate all the speakers in single-channel mixture.

In some cases, such as, teleconference and call center, only the target speaker is transmitted in real time and there is no need to separate all speakers. In [8], Delcroix *et al.* proposed SpeakerBeam to separate the target speech from interfering talker with an additional slice of speech of the target speaker. The authors embedded a speaker adaptive layer into a DNN to adapt the target speaker, where the speaker adaptive layer produces a weighted sum of the contribution of sub-layers.

King *et al.* [9] used an Encoder-decoder model for speaker extraction. The encoder is a long short term memory (LSTM), which projects an anchor speech to a fixed-size embedding. The output of the last frame is used as identity of the target speaker, and fed into the decoder to predict the target speech. The encoder and the decoder are jointly trained to improve performance. One problem is that the output of the last frame in encoder does not represent the speaker well [10][11].

In this paper, we introduce attention mechanism [11] for target speaker separation. Attention mechanism weights the encoder output by a set of coefficients to obtain a embedding. Then the coefficients are fixed in decoding phase. Intuitively, the coefficients should consider both anchor speech and mixture. Hence, we propose a dynamic-attention based Encoder-decoder to obtain a more effective speaker representation. The dynamic attention provides a soft decision to select frames of the anchor speech according to the current frame of the mixture. Therefore, each frame of the mixture will conduct a different embedding which is called the acoustic-related speaker identity.

The rest of the paper is organized as follows. We will describe the related work in Section II. The proposed algorithm is described in detail in Section III. The experimental setup and evaluation results are presented in Section IV. We conclude this paper in Section V.

II. Related work

A. Encoder-decoder speaker extraction model

Before introducing the proposed model, we first review a conventional Encoder-decoder (EncDec) model. Fig. 1 is the schematic diagram of the EncDec model. The encoder network is an LSTM model consuming a variable length sequence of features $(\boldsymbol{s}_1 \cdots \boldsymbol{s}_T)$ from the anchor speech, and generates an embedding $\boldsymbol{e}$ of the desired speech. This embedding is appended to the acoustic feature vector $(\boldsymbol{x}_1 \cdots \boldsymbol{x}_N)$, then the combined vector is fed into a decoder LSTM, which is trained to predict senone posteriors for the frame. The parameters are

jointly optimized by minimizing the mean square error (MSE) objective function via ADAM optimizer[12].

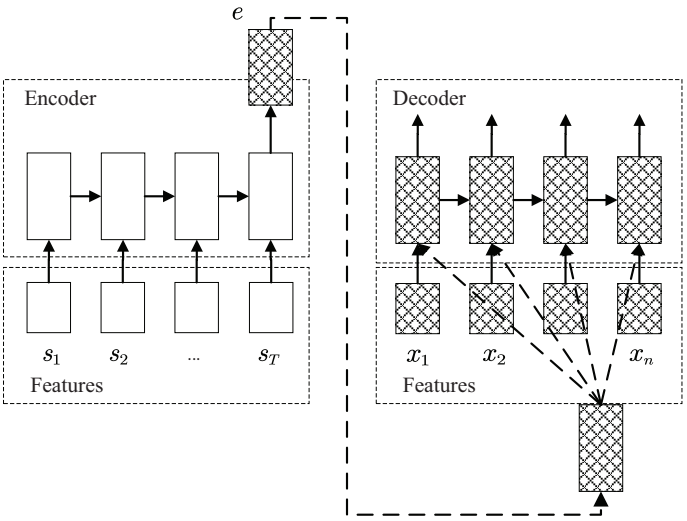

Fig. 1. EncDec speaker extraction model.

B. Attention mechanism

In EncDec, the embedding feature for the target speaker is not well captured [10]. Ideally, the speaker embedding should be built only using the frames corresponding to phonemes. Recently, attention mechanism has shown great performance on a range of tasks, such as, speech recognition [11] and machine translation [13][14]. The structure of the attention mechanism is shown in Fig. 2. The attention mechanism learns a scalar score $\beta_t \in \mathbb{R}$ for the LSTM output $\boldsymbol{z}_t$ at frame t,

$$\beta_t = f\left(\mathbf{z}_t\right),\ t \in [1,2,\cdots,T], \tag{1}$$

where $f(\cdot)$ is the attention function. Then, the normalized weights $h_t \in [0,1]$ can be calculated using these scores,

$$h_t = \frac{\exp\left(\beta_t\right)}{\sum_{i=1}^{T}\exp\left(\beta_i\right)} \quad t \in [1,2,\cdots,T], \tag{2}$$

such that $\sum_{t=1}^{T} h_t = 1$. Finally, as shown in Fig. 2, the embedding is formed as the weighted sum of the LSTM's output as,

$$\boldsymbol{e} = \sum_{t=1}^{T} \boldsymbol{z}_t h_t. \tag{3}$$

III. Dynamic-attention based EncDec model

For the attention mechanism, the embedding is fixed for a certain anchor speech. Intuitively, the coefficients should consider both anchor speech and mixture. To obtain a more effective speaker representation, in this paper, we propose a dynamic-attention based EncDec where the weight coefficients in attention are time-varying with the mixture speech. The structure of the proposal is shown in Fig. 3.

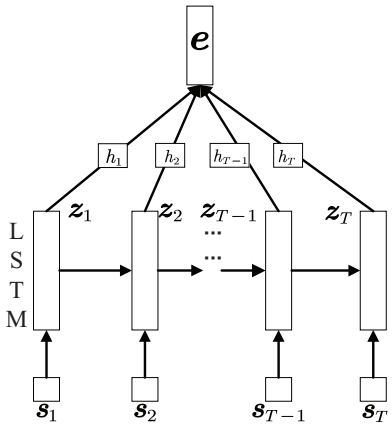

Fig. 2. Traditional attention mechanism.

In the time frame m, $\mathbf{r}_m$ (m-th frame output of the LSTM II) is used as the additional input to learn the scalar score $\beta_{t,m}$,

$$\beta_{t,m} = f\left([\boldsymbol{z}_t, \boldsymbol{r}_m]\right), \tag{4}$$

where $t \in [1,2,...,T]$ is anchor speech frame index.

In this paper, we choose the *Bahdanau Attention* [15] as the attention function for our experiments. Equation (4) can rewritten as:

$$\beta_{t,m} = \boldsymbol{V}^T \tanh(\mathbf{W}^z \boldsymbol{z}_t + \mathbf{W}^r \boldsymbol{r}_m + \boldsymbol{b}), \tag{5}$$

where $\boldsymbol{V}^T$, $\mathbf{W}^z$, $\mathbf{W}^r$ and $\boldsymbol{b}$ are trainable parameters. Then, the normalized weights $h_{t,m} \in$ [0,1] can be calculated as,

$$h_{t,m} = \frac{\exp\left(\beta_{t,m}\right)}{\sum_{i=1}^{T}\exp\left(\beta_{i,m}\right)}. \tag{6}$$

Finally, the embedding $\boldsymbol{e}_m$ at m-th frame of mixture is formed as a weighted sum of encoder outputs, as follows,

$$\boldsymbol{e}_m = \sum_{t=1}^{T} \mathbf{z}_t h_{t,m}. \tag{7}$$

Finally, $\boldsymbol{e}_m$ and $\boldsymbol{r}_m$ are concatenated as the input to LSTM III to separate target speech. The encoder and the decoder are jointly trained.

IV. Experimental

A. Dataset

The proposed system is evaluated by using the WSJ0-2mix datasets[1]. In WSJ0-2mix, each sentence contains two speakers. The WSJ0-2mix dataset introduced in [5] is derived from the WSJ0 corpus [16]. The 30h training set and the 10h validation set contain two-speaker mixtures generated by randomly selecting from 49 male and 51 female speakers from si_tr_s. The Signal-to-Noise Ratios (SNRs) are uniformly chosen between 0 dB and 5 dB. The 5h test set is generated similarly by using utterances from 16 speakers from si_et_05, which do not appear in the training and validation sets. The

[1]Available at: http://www.merl.com/demos/deep-clustering

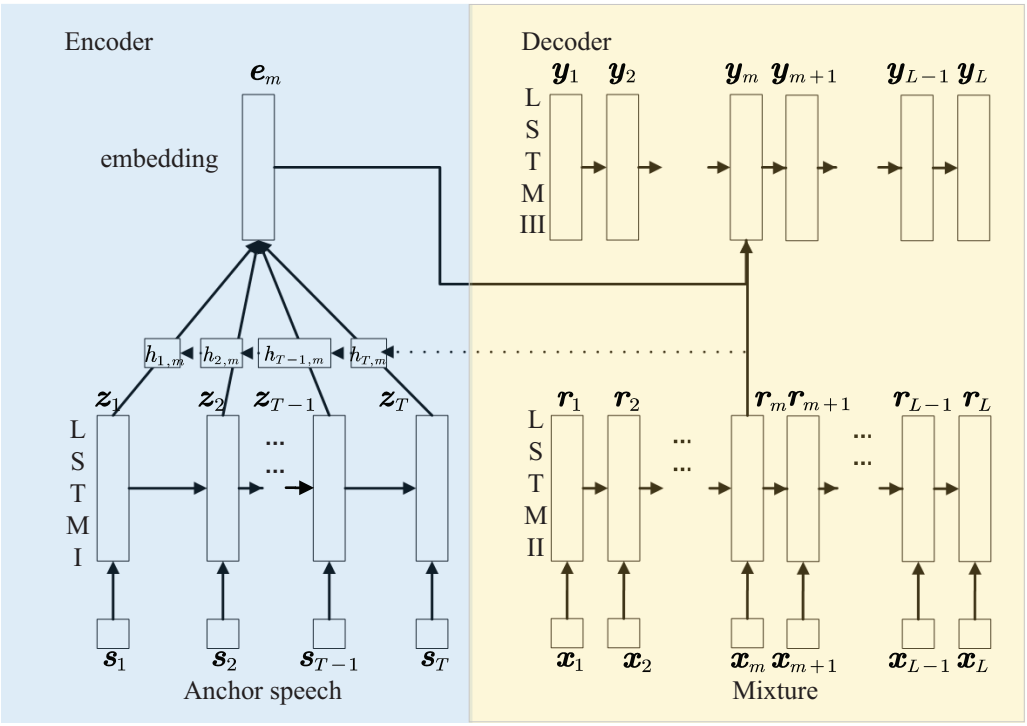

Fig. 3. Dynamic-attention based EncDec model.

test set includes 1603 F&M sentences, 867 M&M sentences, and 530 F&F sentences.

For each mixture, we randomly choose an anchor utterance from the target speaker (different from the utterance in the mixture). The length of the anchor speech is 1 second on average.

B. Metrics and parameters

The performance is evaluated by two objective metrics: perceptual evaluation of speech quality (PESQ) [17] and Signal-to-Distortion Ratio (SDR) [18], both of which are widely used to evaluate speech enhancement performance for multi-talker speech separation tasks. The PESQ measures the speech quality by computing the disturbance between clean and processed speech. The range of PESQ score is from -0.5 to 4.5. For both of the PESQ and SDR metrics, the higher number indicates the better performance.

C. Baseline model and algorithm settings

We compare the proposed method with EncDec [9] and SpeakerBeam [8]. The structure used in EncDec is shown in Fig. 1. The original SpeakerBeam used Bidirectional Long Short Term Memory (BLSTM) to get better performance. In this paper, the BLSTM is replaced by LSTM to ensure that the system is causal. The number of sub-layers used in SpeakerBeam is 30.

The short time Fourier transform (STFT) is utilized as the input feature. The length of the STFT analysis window is 32 ms, and the window shift is 16 ms. Thus, the number of FFT points is 256 for 8 kHz sampling rate.

TABLE I
MODEL CONFIGURATION.

name	size	layers
LSTM I	512	3
LSTM II	512	1
LSTM III	1024	2

The embedding size for the proposed and EncDec model is 512. The configuration of the proposed method is shown in Table I. For EncDec, the encoder has 3 LSTM layers with 512 units in each layer, and the decoder has 3 LSTM layers with 1024 units in each layer. The cost function is MSE. Weights of the networks are randomly initialized. The ADAM optimizer is utilized for back propagation. We also use dropout [19] in LSTM layers to avoid overfitting. The dropout rate is 0.2. All models are trained using Pytorch [20].

D. Evaluation results

TABLE II
AVERAGE SDR SCORE ON TEST SET.

	F&M	F&F	M&M	Average
unprocessed	2.58	2.71	2.65	2.62
EncDec	10.04	5.43	5.30	7.85
SpeakerBeam	9.93	5.43	5.28	7.80
Proposed	**10.24**	**5.85**	**5.71**	**8.16**

Table II and Table III list the average PESQ and SDR score on test set, respectively. It can be seen that the proposed method outperforms the EncDec and SpeakerBeam. Compared with the mixture, extracted speech improves the PESQ score

TABLE III
AVERAGE PESQ SCORE ON TEST SET.

	F&M	F&F	M&M	Average
unprocessed	2.15	2.13	2.25	2.17
EncDec	2.62	2.28	**2.35**	2.48
SpeakerBeam	2.47	**2.29**	2.33	2.39
Proposed	**2.65**	**2.29**	**2.35**	**2.50**

around 0.33, and the SDR nearly 5.54 dB, indicating the effectiveness of proposed method to perform speaker extraction.

The EncDec extracts the embedding feature uses the anchor speech in encoding phase, then fixes the embedding serves an additional input to the decoder. Compared with the EncDec, the proposed method extracts embedding dynamically according to the mixture, which can track the target speaker better.

The proposed method also outperforms another recent method SpeakerBeam[8]. The SpeakerBeam determines the expression of the 30 sub-layers through a set of adaptive coefficients generated by the anchor speech. Although the number of parameters in SpeakerBeam is bigger than the proposed method. The number of sub-layers is still insufficient to characterize the target speaker. Compared with SpeakerBeam, the proposed method still obtains 0.36 SDR gain (from 7.80 to 8.16) and 0.11 PESQ gain (from 2.39 to 2.50).

Fig. 4 illustrates the spectrograms of the extracted speech using different methods on a test utterance, where both target and interfering speaker are female. We can find that the proposed method can extract the target speaker from the mixture and suppress the interfering speaker better than compared methods.

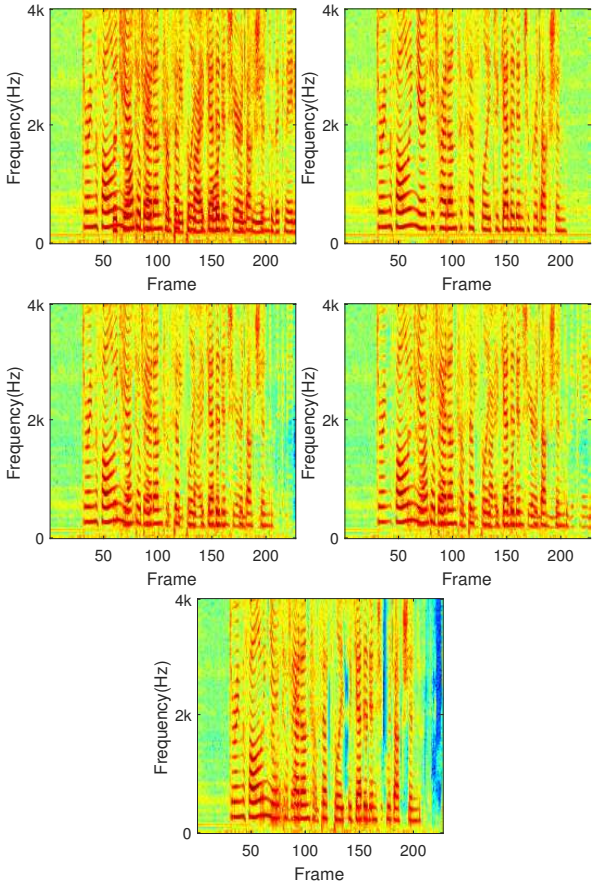

Fig. 4. Spectrograms of extracted speech using different methods. Top left is the spectrogram of the mixture. Top right is the spectrogram of the target speech. Center left is the spectrogram of the EncDec. Center right is the spectrogram of the SpeakerBeam. Bottom center is the spectrogram of the proposed method.

V. CONCLUSION

In this paper, a dynamic-attention based EncDec model for the speaker extraction with anchor speech is proposed. A dynamic attention mechanism is used to take advantage of anchor speech, and extract target speaker better. According to the experiments results, the proposed method achieves better performance than the other related methods. In the future, we will explore the effect of anchor length on speaker extract and the robustness problem in the presence of noise and reverberation.

ACKNOWLEDGMENT

This research was partly supported by the China National Nature Science Foundation (No. 61876214, No.61773224).

REFERENCES

[1] S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov. A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 25(4):692–730, 2017.

[2] O. Schwartz, S. Gannot, and E.A. Habets. Multispeaker LCMV beamformer and postfilter for source separation and noise reduction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(5):940–951, 2017.

[3] S. Araki, S. Makino, H. Sawada, and R. Mukai. Underdetermined blind speech separation with directivity pattern based continuous mask and ICA. In *Signal Processing Conference, 2004 12th European*, pages 1991–1994. IEEE, 2004.

[4] D. Wang and J. Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018.

[5] J.R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 31–35. IEEE, 2016.

[6] 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*, pages 241–245. IEEE, 2017.

[7] Z. Chen, Y. Luo, and N. Mesgarani. Deep attractor network for single-microphone speaker separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 246–250. IEEE, 2017.

[8] M. Delcroix, K. likova, K. Kinoshita, A. Ogawa, and T. Nakatani. Single channel target speaker extraction and recognition with Speaker Beam. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5554–5558. IEEE, 2018.

[9] B. King, I.-F. Chen, Y. Vaizman, Y. Liu, R. Maas, B.H. Parthasarathi, and B. Hoffmeister. Robust speech recognition via anchor word representations. *INTERSPEECH-2017*, pages 2471–2475, 2017.

[10] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani. Speaker-Aware Neural Network Based Beamformer for Speaker Extraction in Speech Mixtures. In *Interspeech*, pages 2655–2659, 2017.

[11] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-based models for speech recognition. In *Advances in neural*

information processing systems, pages 577–585, 2015.

[12] D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] M.-T. Luong, H. Pham, and C.D. Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[14] O. Firat, K. Cho, and Y. Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*, 2016.

[15] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[16] J. Garofolo, D. Graff, D. Paul, and D. Pallett. Csr-i (wsj0) complete ldc93s6a. *Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[17] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *ICASSP*, volume 2, pages 749–752. IEEE, 2001.

[18] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469, 2006.

[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[20] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.