目的話者抽出法 SpeakerBeam の実雑音・残響環境下での評価 *

○デルクロア・マーク [1], Zmolikova Katerina[2]，落合翼[1]，木下慶介 [1], 荒木章子[1]，中谷智広 [1],
[1] 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所,
[2] Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Czechia,

1 あらまし

近年，目的話者の特徴を表現した補助情報を用いて，混合音声から目的話者の音声を選択的に抽出する方法 SpeakerBeam を提案、検討してきた。本稿では、SpeakerBeam 処理の低演算化を目的とした新たなニューラルネットワーク構造を提案し、実雑音・残響環境下での実験により、その効果を確認する。

2 Introduction

Recently deep learning has been used intensively for processing mixtures of speech signals. For example, several approaches have been proposed for speech separation such as deep clustering [1], deep attractor networks [2] and permutation invariant training (PIT) [3]. Speech separation approaches separate a speech mixture into all its sources. Consequently, it usually requires knowledge or estimation of the number of sources in the mixtures. Moreover, there is a permutation ambiguity at the output, i.e. the mapping between the sources and the separation outputs is arbitrary. For some applications, these issues limit the practical uses of separation methods.

We have recently proposed SpeakerBeam, which is an alternative approach for processing mixtures of speech signals [4, 5]. SpeakerBeam extracts only the speech signal that corresponds to a target speaker, instead of separating all source signals. SpeakerBeam uses an auxiliary adaptation utterance containing only the speech of the target speaker to compute the characteristics of the voice of the target speaker. SpeakerBeam can then focus on extracting only speech of that speaker based on the speech characteristics. This framework alleviates the permutation ambiguity problem and does not need knowledge of the number of speakers in the mixtures.

The original SpeakerBeam [4] required a large number of parameters, which was not practical. In this paper, we present a compact version of SpeakerBeam and evaluate it with noisy and reverberant speech mixtures.

3 Compact SpeakerBeam

Figure 1 is a schematic diagram of SpeakerBeam. It consists of two neural networks, a sequence summary network, and a target speech extraction network. The sequence summary network inputs the auxiliary adaptation utterance and outputs a vector, $\mathbf{a}^{(s)}$, characterizing the target speaker as [6],

$$\mathbf{a}^{(s)} = \frac{1}{T}\sum_{t=1}^{T} G(\mathbf{a}_t^{(s)}), \tag{1}$$

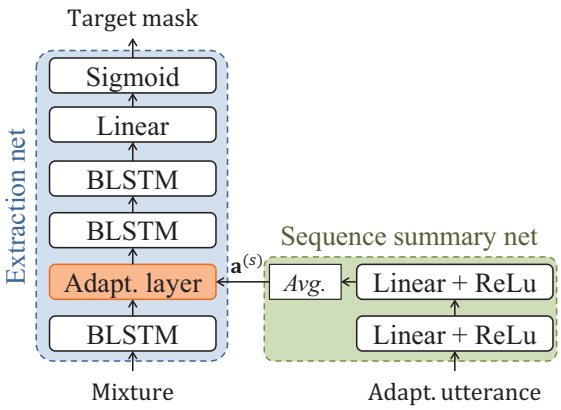

Fig. 1 Schematic diagram of SpeakerBeam.

where $\mathbf{a}_t^{(s)}$ represents the amplitude spectrum of the t^{th} frame of the adaptation utterance, T is the number of frames of the adaptation utterance, and $G(\cdot)$ consists of an auxiliary neural network.

The target speech extraction neural network inputs the amplitude spectrum features of the mixture and outputs a time-frequency mask that indicates regions where the target speaker is active. The extraction network contains an adaptation layer that inputs the vector characterizing the target speaker, $\mathbf{a}^{(s)}$, as auxiliary input. This adaptation layer makes the extraction network able to focus on extracting the speech of the target speaker only. The original version of SpeakerBeam uses a factorized adaptation layer [7]. Such a factorized layer requires a large number of parameters, making it difficult to train and deploy. Here we call this approach SpeakerBeam with factorized adaptation layer ("SpeakerBeam FA").

Here we proposed a simpler adaptation layer that simply scales the activation with weights obtained from the auxiliary network as,

$$\mathbf{h}^{\text{out}} = \mathbf{a}^{(s)} \odot \mathbf{h}^{\text{in}}, \tag{2}$$

where $\mathbf{h}^{\text{out}}$ and $\mathbf{h}^{\text{in}}$ are the output and input of the adaptation layer. This approach is related to subspace LHUC[8], that was recently proposed for speaker adaptation of acoustic models for automatic speech recognition. We call this new version SpeakerBeam with scaling adaptation layer ("SpeakerBeam SA").

4 Experiments

4.1 Data

We tested SpeakerBeam on two different artificially created corpora. The first corpus consists of English

* Evaluation of SpeakerBeam target speech extraction in real noisy and reverberant conditions by DELCROIX Marc[1], ZMOLIKOVA Katerina [2], OCHIAI Tsubasa[1], KINOSHITA Keisuke[1], ARAKI Shoko[1], NAKATANI Tomohiro[1], [1] NTT Communication Science Laboratories, NTT Corporation, [2] Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Czechia,

Table 1 SDR for English and Japanese mixtures.

	size	ENG	JP
Mixture	-	0.15 dB	-2.0 dB
PIT	134 M	8.8 dB	4.8 dB
SpeakerBeam FA	212 M	9.9 dB	-
SpeakerBeam SA	134 M	9.8 dB	5.0 dB

mixtures generated by mixing WSJ utterances without noise or reverberation [1]. The training set consists of 30 hours of speech. The second corpus consists of mixtures of Japanese utterances taken from CSJ [9]. This dataset includes background babble noise, music, and reverberation. The SNR ranges from 0 to 20 dB. The training set consists of 157 hours of speech.

We also present results of real recordings that mimic the mixing conditions of the Japanese noisy and reverberant mixtures. For all experiments, we used an utterance of the target speaker different from that in the mixture as adaptation utterance.

4.2 Settings

The network configuration consists of 3 BLSTM layers with 512 units for the forward and backward passes and a 512 × 1024 projection layer followed by a tanh activation at the output of each BLSTM layer. The output layer is a fully connected layer followed by a sigmoid to output mask values in the range $[0, 1]$.

For SpeakerBeam, we replaced the layer after the first BLSTM layer by a factorized (FA) or scaling adaptation (SA) layer. For "SpeakerBeam FA" we used 30 factors. For the sequence summary network, we used 2 fully connected layers with 200 units and ReLu activation functions. The output layer of the auxiliary network consists of a linear layer followed by a time averaging operation. The input features of the extraction and auxiliary networks consist of amplitude spectrum.

We compare SpeakerBeam with a PIT-based separation network with a similar network configuration, except that it outputs 2 masks (one for each source in the mixtures). In case of PIT, we assumed oracle target speaker selection after separation. Consequently, PIT performance should be understood as upper-bound performance for PIT-based target speech extraction.

We evaluate performance in terms of signal to distortion ratio (SDR) [10].

4.3 Results and discussions

Table 1 shows the performance of the English and Japanese mixtures. The proposed "SpeakerBeam SA" performs similarly to "SpeakerBeam FA" but with a much smaller network, making it more practical to use. SpeakerBeam also outperforms PIT based separation with oracle target speaker selection for both tasks. This confirms that SpeakerBeam can extract a target speaker even in severe noisy conditions.

The results of Table 1 were obtained with simulated data. We also tested using the model trained on simulated Japanese mixture for processing of real recordings. Figure 2 shows the spectrograms of mixture of 2 female speakers with babble background noise, the headset for both speakers and extracted speech for real recordings with one microphone and 8 microphones. When using multiple microphones, SpeakerBeam is combined with a beamformer as in [4, 11]. Looking at the spectrograms,

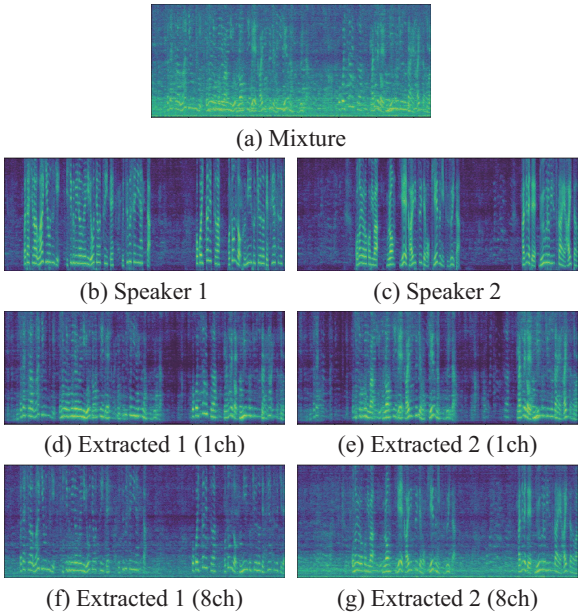

Fig. 2 Spectrograms of (a) a real recording of a mixture of two female speakers, (b)-(c) headset of recordings of each target speaker and SpeakerBeam outputs for (d)-(e) single (1ch) and (f)-(g) multi-microphone (8ch) processing.

we can confirm that SpeakerBeam successfully extracts the target speakers even for such challenging conditions. Interested readers can evaluate the speech extraction performance in our demo video [12].

参考文献

[1] J. R. Hershey, et al., "Deep clustering: Discriminative embeddings for segmentation and separation," in Proc. of ICASSP, 2016.

[2] Z. Chen, et al., "Deep attractor network for single-microphone speaker separation," in Proc. of ICASSP, 2017.

[3] Y. Qian, et al., "Single-channel multi-talker speech recognition with permutation invariant training," arxiv, 2017.

[4] K. Zmolikova, et al., "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in Proc. of Interspeech, 2017.

[5] K. Zmolikova, et al., " Learning speaker representation for neural network based multichannel speaker extraction, " in Proc of ASRU ' 17, Dec 2017.

[6] K. Vesely et al., "Sequence summarizing neural network for speaker adaptation," in Proc. of ICASSP, 2016.

[7] M. Delcroix, et al., "Context adaptive deep neural networks for fast acoustic model adaptation in noisy conditions," in Proc. of ICASSP, 2016.

[8] L. Samarakoon and K. C. Sim, "Subspace LHUC for fast adaptation of deep neural network acoustic models," in Proc. of Interspeech, 2016.

[9] K.Maekawa et al. "Spontaneous speech corpus of Japanese," in Proc. of LREC, 2000.

[10] E. Vincent et al., "Performance measurement in blind audio source separation," IEEE trans. ASLP, vol. 14, no. 4, pp. 14621469, 2006.

[11] J. Heymann et al., "Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition," in Proc. CHiME 4, 2016.

[12] "SpeakerBeam," `https://youtu.be/7FSHgKip6vI`, (English), `https://youtu.be/BM0DXWgGY5A` (Japanese), Cited Dec. 10 2018.