End-to-end SpeakerBeam for single channel target speech recognition

Marc Delcroix¹, Shinji Watanabe², Tsubasa Ochiai¹, Keisuke Kinoshita¹, Shigeki Karita¹, Atsunori Ogawa¹, Tomohiro Nakatani¹

¹NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
²Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

Abstract

End-to-end (E2E) automatic speech recognition (ASR) that directly maps a sequence of speech features into a sequence of characters using a single neural network has received a lot of attention as it greatly simplifies the training and decoding pipelines and enables optimizing the whole system E2E. Recently, such systems have been extended to recognize speech mixtures by inserting a speech separation mechanism into the neural network, allowing to output recognition results for each speaker in the mixture. However, speech separation suffers from a global permutation ambiguity issue, i.e. arbitrary mapping between source speakers and outputs. We argue that this ambiguity would seriously limit the practical use of E2E separation systems. SpeakerBeam has been proposed as an alternative to speech separation to mitigate the global permutation ambiguity. SpeakerBeam aims at extracting only a target speaker in a mixture based on his/her speech characteristics, thus avoiding the global permutation problem. In this paper, we combine SpeakerBeam and an E2E ASR system to allow E2E training of a target speech recognition system. We show promising target speech recognition results in mixtures of two speakers, and discuss interesting properties of the proposed system in terms of speech enhancement and diarization ability.

Index Terms: end-to-end speech recognition, target speech extraction, SpeakerBeam

1. Introduction

End-to-end (E2E) automatic speech recognition (ASR) models directly map a sequence of speech features to a sequence of characters using a single neural network [1–5]. These models enable training of the whole ASR system for the ASR criterion. Moreover, they greatly simplify the ASR training and decoding pipelines. Recently, there have been several investigations on E2E systems robust to noise or interfering speakers [6–11]. In [6], a neural network mask-based beamformer front-end was combined with a joint connectionist temporal classification (CTC)/attention-based E2E ASR system, enabling training the enhancement front-end with an ASR training criterion, and avoiding the need for parallel noisy/clean speech data that are usually required when training a speech enhancement neural network front-end. There have been also several proposals to incorporate speech separation functionality into an ASR system [7–10, 12, 13]. For example, [7] proposed connecting a separation front-end module with an E2E ASR system [8–10] directly included speech separation capabilities into the encoder/decoder modules. These models output two character sequences, one for each speaker in the mixture of two speakers. However, these approaches suffer from two limitations. First, speech separation has a global permutation ambiguity issue, i.e. the mapping between the speakers in the mixture and the outputs is arbitrary. Arguably, this ambiguity would be particularly severe for E2E systems since the output is a character sequence that has thus lost any information that could help identify speakers after separation such as in [14]. Second, the E2E training of such models requires using a permutation invariant training (PIT) loss [8, 15] that potentially has severe computation and memory requirements proportional to the number of possible source permutations. Therefore, it would be challenging to extend E2E training to mixtures of more than two speakers.

We have recently proposed SpeakerBeam [16], which is an alternative to speech separation approaches for processing speech mixtures that focuses on extracting only speech of a target speaker in a mixture. The system is informed about the target speaker given his/her speech characteristics that are derived from an adaptation (or enrollment) utterance of that speaker [16–19]. SpeakerBeam is a step toward mimicking human selective hearing ability (i.e. cocktail party problem) [20, 21]. SpeakerBeam presents a competitive alternative to separation approaches [22] and does not suffer from the global permutation ambiguity issue by its nature. Moreover, the model is independent of the number of sources in the mixtures.

In this paper, we combine SpeakerBeam and E2E ASR to realize E2E training of a target speech recognition system. Following our previous investigations on SpeakerBeam combined with hybrid ASR systems [23], we propose two implementations of E2E SpeakerBeam, i.e. an adaptive encoder and a cascade connection. The adaptive encoder inserts the speech extraction capability in the encoder of an E2E ASR system by including an adaptation layer that adapts the encoder to the target speaker based on the speech characteristics. The cascade connection scheme interconnects a time-frequency mask estimation network for extracting the target speech and an E2E ASR module, enabling better interpretability of the network behavior. A related approach was proposed recently [11], where the speaker characteristics were added to the attention module. However, that approach could only recognize speech of a target speaker in mixtures of non-overlapping speech since the attention is performed only over time and not frequency.

In experiments with mixtures of two speakers, we confirm that our proposed schemes achieve high level of recognition performance, while maintaining the advantages of conventional SpeakerBeam (i.e. no global permutation ambiguity, constant complexity with increased number of speakers in the mixture) and E2E training (i.e. no need for parallel mixture/clean data during training). Moreover, we show that E2E SpeakerBeam with cascade connection realizes limited but perceivable enhancement of the target speaker and that it can detect the activity periods of the target speaker, demonstrating the potential for joint speech recognition and diarization of speech mixtures.
2. E2E target speech recognition

We first introduce the problem and review the baseline E2E ASR module for a single speaker recognition that serves as a basis for our proposed method. We then describe the two proposed E2E SpeakerBeam schemes, i.e. adaptive encoder and cascade connection.

2.1. Problem formulation

Let us consider a model of observed speech mixtures as,

\[ Y^{\text{STFT}} = X^{\text{STFT}} + \sum_{i \neq s} X_i^{\text{STFT}} + N^{\text{STFT}}, \quad (1) \]

where \( Y^{\text{STFT}}, X_s^{\text{STFT}}, X_i^{\text{STFT}}, N^{\text{STFT}} \in \mathbb{C}^{T \times F} \) are sequences of short time Fourier transform (STFT) coefficients associated with a speech mixture, a target speech, interference speakers and noise, respectively. \( T \) and \( F \) are the number of time frames and frequency bins, respectively. In this work, we do not consider noise, but we have confirmed that SpeakerBeam can work in noisy and reverberant conditions [22]. We use either amplitude spectrum coefficients (e.g. \( Y^{\text{Amp}} \)) or log Mel filterbank coefficients (e.g. \( Y^{\text{Mel}} \)) as input features.

Let \( W_s \in \mathbb{U}^K \) be the sequence of characters associated with the target speech signal, \( X_s \), where \( \mathbb{U} \) is the set of possible characters and \( K \) is the length of the output character sequence.

Let \( A_s \in \mathbb{R}^{T' \times F} \) be the sequence of speech features of an adaptation utterance of the target speaker, where \( T' \) is the number of time frames of that utterance. Note that the adaptation utterance includes speech spoken only by the target speaker but differs from the utterance in the mixture.

We aim at recognizing only the character sequence, \( W_s \), of the target speaker in the mixture, by exploiting the adaptation utterance, \( A_s \), to inform the system about the target speaker.

2.2. E2E ASR module for single speaker

Fig. 1 shows a schematic diagram of our baseline E2E ASR module for recognition of a single speaker. It consists of an encoder module composed of a VGG-motivated CNN [24, 25] and several BLSTM layers followed by two parallel decoders, i.e. a CTC and an attention-based decoder [5]. Both modules predict the output character sequence probability given the encoder output sequence. The input features consist of log Mel filterbank coefficients. The model is trained with a multi-task objective function including both CTC and attention losses [5]. In addition, both CTC and attention decoder outputs are used during decoding.

2.3. E2E SpeakerBeam with adaptive encoder

Fig. 2-(a) shows a schematic diagram of the E2E SpeakerBeam with an adaptive encoder. In [23], we showed that we could recognize speech of a target speaker in a mixture simply by adapting an acoustic model of a DNN-hybrid ASR system to the target speaker. In this paper, we propose a similar approach for E2E ASR. We insert an adaptation layer to the encoder of the E2E ASR module to make it adaptive to the target speaker, thus enabling it to focus on encoding speech of the target speaker only. The adaptation layer receives auxiliary features computed with an auxiliary network. The details of the auxiliary network and adaptation layer are described below. As for the ASR module for a single speaker, we use log Mel filterbank coefficients as input features for both the encoder and the auxiliary networks, i.e. \( Y^{\text{Mel}} \) and \( A_s^{\text{Mel}} \).

2.3.1. Auxiliary network

The auxiliary network aims at extracting the characteristics of the target speaker from the adaptation utterance. It consists of a sequence summary network [26, 27], which is a neural network followed by a time averaging operation that inputs the adaptation utterance, \( A_s \), and outputs a vector, \( \alpha_s \), representing the target speaker characteristics as,

\[ \alpha_s = \frac{1}{T'} \sum_{\tau=1}^{T'} G(\alpha_s, \tau), \quad (2) \]

where \( G(\cdot) \) is a fully connected neural network and \( \alpha_s, \tau \) is the \( \tau \)-th frame of the adaptation utterance, \( A_s \).

2.3.2. Adaptation layer

The adaptation layer guides the encoder to focus on only the target speaker in the mixture and to neglect interfering speakers, using the auxiliary features, \( \alpha_s \). Various approaches for the adaptation layer have been investigated [16, 18, 22, 28]. In this paper, we employ the scaling activation adaptation layer [22],

\[ h_{\text{out}} = h_{\text{in}} \odot \alpha_s, \quad (3) \]

where \( \odot \) is an element-wise product, and \( h_{\text{in}} \) and \( h_{\text{out}} \) are the input and output of the adaptation layer, respectively. Equation (3) is similar to subspace learning hidden unit contributions (LHUC) that was proposed for acoustic model adaptation [29]. It offers a powerful yet compact approach for adaptation [22].

2.4. E2E SpeakerBeam with cascade connection

Fig. 2-(b) shows a schematic diagram of the E2E SpeakerBeam system with cascade connection. It consists of a target speaker extraction network that computes a time-frequency mask to extract the target speaker out of the mixture [30, 31], and an E2E ASR module for a single speaker to perform recognition based...
3.1. Experimental settings

We perform experiments using the MERL two speaker mixture corpus [33], which is created by mixing clean WSJ0 [34] utterances at SNRs between 0 and 5 dB with full overlap (FO) conditions. The training set consists of 20000 mixtures and the test set of 3000 mixtures of speakers different from those in the training set. In addition, we created a second test set that consists of similar mixtures to MERL mixtures, but with a shifted starting time of one of the utterances to create somewhat more realistic partial overlap (PO) conditions. For both data sets, we used a sampling frequency of 16 kHz.

3.1. Experimental settings

3.2. ASR evaluation

Table 1 shows the CER and WER for the baseline systems and the E2E SpeakerBeam with and without MTL. Not surprisingly, without informing the system about the target speaker, baseline recognition performance is very poor. The performance in PO conditions is also bad because the system also recognizes speech of the interfering speaker causing many insertion errors. Note that recognition of clean speech with the clean baseline system provides a lower-bound WER of 5.5 %.

We observe that both E2E SpeakerBeam schemes significantly improve ASR performance for both FO and PO conditions. The cascade connection provides slightly better results. MTL brings a small improvement in all cases but for the PO condition when using the cascade connection. This is due to a significant increase in insertion errors due to the remaining speech of the interference speaker. We believe that this behavior could be mitigated by retraining the system with PO data.

Table 1: Target speech recognition CER and WER [%].

<table>
<thead>
<tr>
<th>Model</th>
<th>MTL</th>
<th>Full overlap CER</th>
<th>Full overlap WER</th>
<th>Partial overlap CER</th>
<th>Partial overlap WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean baseline</td>
<td>-</td>
<td>75.6</td>
<td>114.7</td>
<td>93.2</td>
<td>106.7</td>
</tr>
<tr>
<td>Dominant baseline</td>
<td>-</td>
<td>57.2</td>
<td>75.7</td>
<td>73.7</td>
<td>87.3</td>
</tr>
<tr>
<td>SpkBeam adapt enc</td>
<td>✓</td>
<td>13.4</td>
<td>21.1</td>
<td>11.6</td>
<td>16.5</td>
</tr>
<tr>
<td>SpkBeam cascade</td>
<td>✓</td>
<td>11.1</td>
<td>18.4</td>
<td>8.9</td>
<td>13.6</td>
</tr>
</tbody>
</table>

VGG-BLSTM encoder with 1024 units for the BLSTMs and sub-sampling by a factor 4, and a joint CTC/attention decoder. The CTC decoder consists of a softmax layer appended directly after the encoder. The attention decoder consists of a location-based attention module, an LSTM layer with 300 units followed by a softmax layer.

The adaptive encoder has the same configuration as the baseline system except that we inserted an adaptation layer after the VGG module and used an auxiliary network consisting of two linear layers with 200 units with ReLU activation functions, followed by a linear layer without non-linear activation function.

The speaker extraction network of the cascade connection consists of 3 BLSTM layers with 512 units. The adaptation layer is inserted after the first BLSTM layer. The auxiliary network has a similar configuration as for the adaptive encoder.

The input features consist of 80 dimension log Mel filterbank coefficients for the baseline and adaptive encoder, and 201 dimension amplitude spectrum coefficients for the cascade connection. We normalized the log Mel filterbank coefficients with utterance-level mean and variance normalization.

All models were trained with adadelta for up to 20 epochs. We trained two baseline single speaker E2E ASR modules one trained with clean speech (“Clean baseline”) and one with the dominant speaker in the mixtures (“Dominant baseline”). Both E2E SpeakerBeam schemes were trained with FO training data (i.e. without PO data), and with and without MTL loss, i.e. $\{\mu = 0.5, \nu = 0.5\}$ (MTL) or $\{\mu = 1, \nu = 0\}$ (no MTL).

In all experiments, we used attention/CTC joint decoding with score combination with a word-level RNN-LM trained on the text data of the WSJ1 corpus [36]. We present results averaged over both target speakers in the mixture in terms of character error rate (CER), word error rate (WER), signal-to-distortion-ratio (SDR) and diarization error rate (DER).
Table 2: Speech enhancement results in terms of SDR [dB].

<table>
<thead>
<tr>
<th></th>
<th>Full overlap</th>
<th>Partial overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MTL Same Diff Avg</td>
<td>Same Diff Avg</td>
</tr>
<tr>
<td>Mixture</td>
<td>- 0.1 0.1 0.1</td>
<td>0.0 0.0 0.0</td>
</tr>
<tr>
<td>SpkBeam</td>
<td>✓ 0.9 3.4 2.2</td>
<td>1.3 3.8 2.7</td>
</tr>
<tr>
<td></td>
<td>2.6 5.8 4.3</td>
<td>3.4 7.7 5.7</td>
</tr>
</tbody>
</table>

![Image](image1)

(a) Mixture
(b) Clean Spk 1
(c) Clean Spk 2
(d) Extracted Spk 1
(e) Extracted Spk 2
(f) Diarization Spk 1
(g) Diarization Spk 2

Figure 3: Example of speech enhancement and diarization outputs using E2E SpeakerBeam with MTL for PO condition.

3.3. Speech enhancement evaluation

E2E SpeakerBeam with cascade connection can output enhanced speech of the target speaker. Table 2 shows the SDR of the mixture and with SpeakerBeam with and without MTL. The table shows the average results and results for the mixtures of the same and different genders. Although the system is only trained with the E2E ASR loss (without any speech enhancement objective function), it still performs some speech enhancement (about 2 dB SDR improvement). As expected, using MTL further improves SDR by an additional 2 dB.

Fig. 3 show examples of enhanced speech for a PO mixture. Sound samples can be found on our webpage [37]. We can perceive that speech of the target speaker is clearly dominant at the output of the speaker extraction module but the interference speaker is still audible. Interestingly, by training the system E2E, the ASR module learns to ignore the remaining interference signals. Note that SpeakerBeam with a similar configuration trained with an enhancement objective and for about 200 epochs can achieve about 9.7 dB on the FO set [23].

The results of Table 2 also show that processing same gender mixtures remains more challenging as the speakers may have more similar voice characteristics.

Table 3: Diarization results in terms of DER [%].

<table>
<thead>
<tr>
<th></th>
<th>Full overlap</th>
<th>Partial overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Same Diff Avg</td>
<td>Same Diff Avg</td>
</tr>
<tr>
<td>Mixture</td>
<td>31.1 31.2 31.1</td>
<td>84.2 84.6 84.4</td>
</tr>
<tr>
<td>(1) Enhanced</td>
<td>28.3 23.4 25.7</td>
<td>73.2 57.9 64.9</td>
</tr>
<tr>
<td>(2) Attention</td>
<td>15.3 8.4 11.6</td>
<td>36.5 18.2 26.6</td>
</tr>
<tr>
<td>(3) CTC</td>
<td>10.9 4.9 7.6</td>
<td>18.1 6.1 11.6</td>
</tr>
</tbody>
</table>

3.4. Diarization evaluation

We experimented the capability of the E2E SpeakerBeam to detect periods where the target speaker is active in the mixture, i.e. further called diarization. We used here E2E SpeakerBeam with cascade connection without MTL (a similar level of performance was achieved with MTL). We tested three simple approaches for diarization based on thresholding a frame level score given by either (1) the energy of the enhanced speech (“Enhanced”), (2) the sum of the attention weights for each input time frame (“Attention”), (3) the blank symbol [38] posterior probability of the CTC decoder output (“CTC”). Table 3 shows the DER evaluated using the NIST toolkit [39]. In the evaluation, we removed the speaker mapping optimization step of the NIST toolkit to also account for potential errors in speaker label assignments. Reference diarization labels were obtained by simply thresholding the energy of the clean signal. The baseline mixture results were obtained using a similar threshold on the mixture (“Mixture”). For all diarization, we performed hangover smoothing before evaluation.

Thresholding the energy of the enhanced speech does not perform well because of the remaining interference speakers. Using the sum of the attention weights performs better, but DER degrades significantly in the PO case. The posterior probability of the blank symbol correlates well with the activity of the target speaker and achieved best DER. Figs. 3(f)-(g) show examples of diarization outputs.

Diarization performance also degrades for same-gender mixtures, mostly for the PO case, because SpeakerBeam may confuse the target speaker when speakers have similar voice characteristics. Improving target speaker discrimination will be a part of our future works.

3.5. Performance comparison with prior works

There have been several works reporting results with the MERL two mixture data set using hybrid and E2E systems [7–10, 13, 40]. A summary of recent scores on this data set is provided in Table 3 of [10], showing that the best performance (WER of 25.4%) for recognition was achieved with an E2E system combining an encoder with separate modules for each speaker in the mixture and separate attention modules. This system performs recognition of both speakers but cannot identify a target speaker as our proposed E2E SpeakerBeam schemes do.

E2E SpeakerBeam has a similar configuration but achieved significantly better recognition performance (WER of 18.0%) using the adaptation utterance of the target speaker. Moreover, it can identify speakers in a mixture.

4. Conclusions

We have proposed two schemes for E2E target speaker extraction and recognition that achieved high recognition performance on the MERL two mixture task. Moreover, we demonstrated that the E2E SpeakerBeam can enhance the target speaker and also detect the speaker activity in the mixture, showing potential for diarization. Future work will include investigations of the proposed schemes in more realistic conditions with noise and reverberation [22, 41].

5. Acknowledgements

We would like to thank Xuankai Chang of Shanghai Jiao Tong University and Katerina Zmolikova of Brno University of Technology (BUT) for their supports in setting-up the baselines.
6. References


