A UNIFIED FRAMEWORK FOR NEURAL SPEECH SEPARATION AND EXTRACTION

Tsubasa Ochiai, Marc Delcroix, Keisuke Kinoshita, Atsunori Ogawa, Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan

ABSTRACT

The development of deep learning techniques has triggered the active investigation of neural network-based speech enhancement approaches. In particular, single-channel blind (*uninformed*) speech separation and speaker-aware (*informed*) speech extraction have received increased interest. Blind speech separation separates a speech mixture into all source signals without requiring any auxiliary information about the speakers. In contrast, speaker-aware speech extraction focuses on extracting speech from a target speaker using prior knowledge, such as an utterance spoken by the target speaker. Speaker extraction is therefore not fully blind, but it can mitigate the source permutation problem faced by blind source separation, and potentially achieve better speech quality by exploiting the auxiliary information. In this paper, to take advantage of both approaches, we propose a unified framework for both speech separation and speech extraction using a single model. This is realized by incorporating a speaker attention mechanism within a generalized permutation invariant training (PIT)-based blind speech separation model, and introducing a multitask separation/extraction objective for training the model. Experiments on the WSJ0-2mix dataset show that our proposed framework realizes both uninformed separation and informed extraction, and achieves better separation/extraction performance than a baseline PIT-based model.

***Index Terms*—** Speech separation/extraction, neural network, speaker attention

1. INTRODUCTION

With the advent of deep learning techniques, the performance of automatic speech recognition (ASR) systems has significantly improved [1, 2] and research interest has moved towards more challenging tasks, e.g., speech inputs contaminated by background noise and overlapping speakers [3]. To address such challenging tasks, deep learning-based single-channel source separation approaches have been actively investigated [4–11].

There are currently two main research directions; 1) blind (*uninformed*) speech separation [4–8] and 2) speaker-aware (*informed*) speech extraction [9–11]. The purpose of uninformed speech separation is to separate an input speech mixture into as many signals as there are in the mixture, without using any auxiliary information about the speakers in the mixture. Uninformed separation can work in a fully blind fashion, which may be required in many real-world applications when the identity of the speaker in the mixture cannot be obtained in advance. However, although recent uninformed separation approaches [4, 7] solve the permutation problem within an utterance, they generally suffer from a block permutation problem [12], i.e., a permutation problem across utterances. On the other hand, the aim of informed speech extraction is to extract only a target speaker's speech from the mixture, using auxiliary information about the target speaker. Although information about the target speaker is required for informed extraction to work, informed extraction has the potential to solve the source permutation problem faced by uninformed separation and track speakers across utterances. In addition, by exploiting the auxiliary information in the extraction stage, informed extraction also has the potential to achieve better speech quality than uninformed separation.

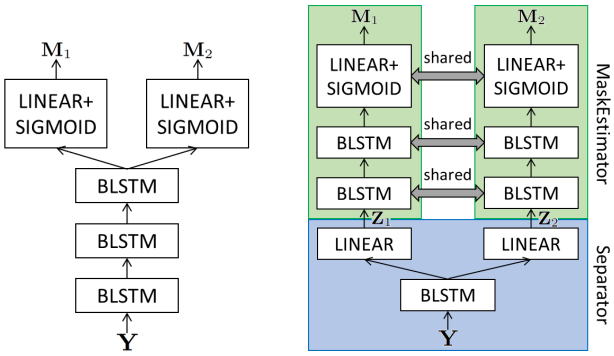
To take advantage of both approaches, it would be desirable to develop a framework that encompasses the capabilities of both uninformed speech separation and informed speech extraction, which would enhance the utility and both functionalities. To enable informed extraction, one could adopt an extra speaker identification step to identify the target speaker from the outputs of the uninformed separation network. However, such a speaker identification step inevitably suffers from identification errors that would limit the speech extraction performance. Recently, [13] proposed introducing an additional speaker identification layer to an uninformed separation network to jointly learn speaker identification embeddings and enable informed extraction. However, the separation network does not have access to the auxiliary information that could be beneficial for improving the extracted speech quality if it could be properly exploited.

In this paper, we propose a unified framework for speech separation and extraction using a single model, which we call the attention-based speech separation and extraction network (ASENet). We extend an uninformed speech separation approach to include an informed speech extraction capability. Recent deep learning-based speech separation approaches can be categorized into embedding space-based approaches [4, 5] such as deep clustering (DC), and mask estimation network-based approaches [6, 7] such as permutation invariant training (PIT). We adopt the PIT-based approach as the basis of our proposed framework because of the simplicity of its separation procedure, which does not require an additional clustering step, and the consistency between the training and test stages.

To realize a unified separation/extraction framework, we first generalized a PIT-based uninformed separation model so that it incorporated the separation mechanism internally and extended it by introducing 1) a speaker attention mechanism that can perform speaker selection inside the network and 2) a multitask learning-based training procedure that considers uninformed PIT-based loss and informed speaker-aware loss, simultaneously. With the proposed attention mechanism, the system can perform blind speech separation when auxiliary information is unavailable, or speaker-aware speech extraction when auxiliary information is available. Moreover, uninformed speech separation performance would also be improved thanks to the multitask objective that adds speaker-aware loss as an auxiliary task to the conventional PIT-based loss [13].

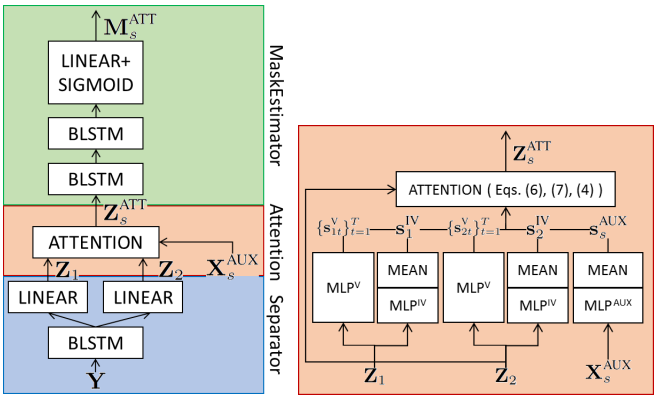
2. PROPOSED METHOD

The proposed framework, i.e., ASENet, operates in two modes depending on the availability of the auxiliary speaker information: 1) uninformed separation mode and 2) informed extraction mode. Firstly in Section 2.1, we generalize the conventional PIT-based separation network by introducing an internal separation model as the basis of the proposed architecture. Then in Sections 2.2-2.4, we describe in detail the architecture of the proposed ASENet and its



(a) Output separation (b) Internal separation

Fig. 1. Overview of speech separation network.



(a) Overall architecture (b) Details of attention module

Fig. 2. Overview of attention-based speech separation and extraction network (ASENet).

multitask learning-based training procedure.

2.1. Generalization of separation network architecture

A conventional PIT-based separation network is shown in Fig. 1-(a). It consists of a bidirectional long short-term memory (BLSTM)-based architecture with multiple output layers corresponding to each speaker in the mixture. Given a sequence of short-time Fourier transform (STFT)-based amplitude spectrum features of the mixture $\mathbf{Y}$, the network estimates a sequence of time-frequency (TF) masks $\mathbf{M}_i$ associated with each separate output $i = 1, \ldots, I$, where I is the number of separate output layers. The i-th separated speech signals $\hat{\mathbf{X}}_i$ are obtained by applying the estimated TF masks to the mixture signal as $\hat{\mathbf{X}}_i = \mathbf{M}_i \odot \mathbf{Y}$, where $\odot$ denotes element-wise product. In the following, to simplify the discussion, we consider the two source cases (i.e., $I = 2$), although the framework can be generalized to more sources.

Our aim is to include the speaker selection mechanism in the PIT-based separation framework. We could undertake speaker selection after the output layers; however, it is unclear whether this would offer the best speaker representation. Therefore, we first generalize the PIT-based separation to include the separation mechanism internally, as shown in Fig. 1-(b). With such a generalization, we can interpret the separation mechanism as being in two blocks, a separation block that generates separate internal embedding vectors $\{\mathbf{Z}_i\}_{i=1}^I$ associated with each source, and a mask estimation block that generates TF masks $\mathbf{M}_i$ from the internal embedding vectors, as the following functional forms:

$$\{\mathbf{Z}_i\}_{i=1}^I = \text{Separator}(\mathbf{Y}), \tag{1}$$

$$\mathbf{M}_i = \text{MaskEstimator}(\mathbf{Z}_i) \qquad (i = 1, ..., I), \tag{2}$$

where i is the index of the output of the Separator$(\cdot)$. Note that we assume that the MaskEstimator$(\cdot)$ can be shared by the I sources and thus use shared parameters.

2.2. Attention-based Separation and Extraction Network

In the informed extraction mode, it is assumed that the auxiliary (target speaker) information is available when extracting the signal of the target speaker from the mixture. Here, the auxiliary information consists of a sequence of STFT-based amplitude spectrum features $\mathbf{X}_s^{\text{AUX}}$ derived from an utterance spoken by the target speaker and different from that in the mixture, where s is the index of the target speaker.

Figure 2-(a) shows the overall architecture of the proposed ASENet. We incorporate the speaker-aware extraction functionality in the internal separation network by adding a speaker attention module. The role of the speaker attention module is to select which of the separate internal embedding vectors $\{\mathbf{Z}_i\}_{i=1}^I$ corresponds to the target speaker. We represent the entire attention-based extraction procedure in the following functional forms:

$$\{\mathbf{Z}_i\}_{i=1}^I = \text{Separator}(\mathbf{Y}), \tag{3}$$

$$\mathbf{z}_{st}^{\text{ATT}} = \underbrace{\sum_{i=1}^{I} a_{sti}\mathbf{z}_{it}}_{\text{Attention}}, \qquad (t = 1, ..., T) \tag{4}$$

$$\mathbf{M}_s^{\text{ATT}} = \text{MaskEstimator}(\mathbf{Z}_s^{\text{ATT}}), \tag{5}$$

where $\{a_{sti}\}_{i=1}^I$ is the attention weight vector at time step t for the target speaker s, and i indicates an index of the internal embedding vectors. Here, Separator$(\cdot)$ and MaskEstimator$(\cdot)$ are the same module in Eqs. (1) and (2). Separator$(\cdot)$ converts an input sequence of the mixture $\mathbf{Y}$ to a separate internal embedding vectors $\{\mathbf{Z}_i\}_{i=1}^I$, where $\mathbf{Z}_i = \{\mathbf{z}_{it}\}_{t=1}^T$. The speaker attention module reconstructs the attended internal embedding vectors for the target speaker $\mathbf{Z}_s^{\text{ATT}} = \{\mathbf{z}_{st}^{\text{ATT}}\}_{t=1}^T$ by interpolating the separate internal embedding vectors $\{\mathbf{z}_{it}\}_{i=1}^I$ over the I sources for every time frames. Subsequently, MaskEstimator$(\cdot)$ estimates the TF masks $\mathbf{M}_s^{\text{ATT}}$ for the attended internal embedding vectors $\mathbf{Z}_s^{\text{ATT}}$.

Equation (4) corresponds to the speaker attention mechanism, which performs a soft alignment (weighted averaging) of the separate internal embedding vectors $\{\mathbf{Z}_i\}_{i=1}^I$ over the I sources. The speaker attention mechanism is described in more detail in the following subsection.

The proposed ASENet can function in both informed and uninformed modes. In the informed mode (i.e., $\mathbf{X}_s^{\text{AUX}}$ is given), the estimated attention weights $\{a_{sti}\}_{i=1}^I$ in Eq. (4) are derived from the auxiliary information. On the other hand, in the uninformed mode (i.e., $\mathbf{X}_s^{\text{AUX}}$ is not given), we can extract both speakers by forcing the attention weights to be $a_{s,t,i=1} = 1, a_{s,t,i=2} = 0$ for the first speaker ($s = 1$) and $a_{s,t,i=1} = 0, a_{s,t,i=2} = 1$ for the second speaker ($s = 2$). This forced attention corresponds to the uninformed behavior of the proposed framework, described in Section 2.1. Consequently, the proposed attention mechanism can switch between informed and uninformed modes using estimated or forced attention, depending on the availability of the auxiliary information.

2.3. Details of speaker attention mechanism

Figure 2-(b) shows the proposed speaker attention module in detail. We use the additive attention mechanism proposed in [14] to compute the attention weight vectors, but extend it by incorporating sequence feature extraction networks. The attention mechanism could be time-invariant (constant over a whole utterance) or time-variant. In this paper, we adopted a time-variant attention mechanism because it offers a more general formulation and could potentially mitigate remaining permutation ambiguities.

The time-variant attention weight vector $\{a_{sti}\}_{i=1}^{I}$ for the target speaker s are derived from the separate internal embedding vectors $\{\mathbf{Z}_i\}_{i=1}^{I}$ and the auxiliary input sequence for the target speaker $\mathbf{X}_s^{\text{AUX}}$ as follows:

$$e_{sti} = \mathbf{w}\tanh(\mathbf{W}^{\text{V}}\mathbf{s}_{it}^{\text{V}} + \mathbf{W}^{\text{IV}}\mathbf{s}_{i}^{\text{IV}} + \mathbf{W}^{\text{AUX}}\mathbf{s}_{s}^{\text{AUX}} + \mathbf{b}), \quad (6)$$

$$a_{sti} = \frac{\exp(\alpha e_{sti})}{\sum_{i=1}^{I}\exp(\alpha e_{sti})}, \quad (7)$$

where $\mathbf{w}, \mathbf{b}, \mathbf{W}^{\text{V}}, \mathbf{W}^{\text{IV}}, \mathbf{W}^{\text{AUX}}$ are trainable weight and bias parameters, and α is a sharpening factor [14]. $\mathbf{s}_{it}^{\text{V}}$ is a time-varying (local) embedding vector of $\mathbf{z}_{it}$. $\mathbf{s}_{i}^{\text{IV}}$ and $\mathbf{s}_{s}^{\text{AUX}}$ are time-invariant (global) embedding vectors of $\mathbf{Z}_i$ and $\mathbf{X}_s^{\text{AUX}}$. These embedding vectors are computed as follows:

$$\{\mathbf{s}_{it}^{\text{V}}\}_{t=1}^{T} = \text{MLP}^{\text{V}}(\mathbf{Z}_i), \quad (8)$$

$$\mathbf{s}_{i}^{\text{IV}} = \text{MEAN}(\text{MLP}^{\text{IV}}(\mathbf{Z}_i)), \quad (9)$$

$$\mathbf{s}_{s}^{\text{AUX}} = \text{MEAN}(\text{MLP}^{\text{AUX}}(\mathbf{X}_s^{\text{AUX}})), \quad (10)$$

where $\text{MLP}(\cdot)$ are simple MLP-based networks, and $\text{MEAN}(\cdot)$ represents the mean operation over the time axis. $\mathbf{s}_{s}^{\text{AUX}}$ is thus similar to the sequence summary network-based approach that has been used for speaker adaptation and target speech extraction [15, 16].

2.4. Multitask learning-based training procedure

We assume that a set of speech features $\{\mathbf{Y}, \{\mathbf{X}_s^{\text{TGT}}\}_{s=1}^{S}, \{\mathbf{X}_s^{\text{AUX}}\}_{s=1}^{S}\}$ is available to train the model, where $\mathbf{X}_s^{\text{TGT}}$ is a sequence of STFT-based amplitude spectrum features of the s-th target speaker signal, $\mathbf{X}_s^{\text{AUX}}$ is the auxiliary input sequence corresponding to each target speaker, and the number of mixed speakers S is assumed to be equal to I. To enable both separation and extraction behaviors in the proposed framework, we adopted a multitask learning-based objective function L_{MTL} incorporating uninformed separation and informed extraction losses, as follows:

$$L_{\text{MTL}} = \epsilon L_{\text{PIT}} + (1-\epsilon)L_{\text{ATT}}, \quad (11)$$

$$L_{\text{PIT}} = \min_{P\in\text{perm}(S)}\frac{1}{S}\sum_{s=1}^{S} l(\mathbf{M}_{p_s}\odot\mathbf{Y}, \mathbf{X}_s^{\text{TGT}}), \quad (12)$$

$$L_{\text{ATT}} = \frac{1}{S}\sum_{s=1}^{S} l(\mathbf{M}_{s}^{\text{ATT}}\odot\mathbf{Y}, \mathbf{X}_s^{\text{TGT}}), \quad (13)$$

where $\text{perm}(S)$ produces all possible permutations, $P = \{p_1, ..., p_S\}$ is the selected permutation, $\epsilon \in [0, 1]$ is an interpolation weight, and $l(\mathbf{A}, \mathbf{B}) = \frac{1}{TF}\|\mathbf{A}-\mathbf{B}\|^2$ is the mean squared error (MSE) criterion. Here, $\mathbf{M}_{p_s}$ indicates TF masks generated in the uninformed separation mode described in Eqs. (1)-(2), while $\mathbf{M}_{s}^{\text{ATT}}$ represents TF masks generated in the informed extraction mode described in Eqs. (3)-(5).

L_{PIT} corresponds to uninformed separation loss based on the PIT loss [7]. Because the correspondence between estimated TF masks and speakers is unknown in the uninformed mode, the PIT-based loss uses the permutation that minimizes the utterance-level separation error. On the other hand, L_{ATT} corresponds to informed extraction loss. Because the correspondence between estimated TF masks and speakers is known thanks to the auxiliary information, we can define the speaker-aware loss without solving the permutation.

Multitask-leaning for the PIT-based approach was investigated in [17] using two uninformed separation losses: PIT-based and deep clustering-based losses. Our use of multitask learning is different since we use an informed (speaker-aware) extraction loss as an auxiliary task. Since our proposed framework retains the fundamental properties of the conventional PIT-based approach, techniques proposed for PIT-based separation such as the deep clustering-based multitask loss could also be applied to our proposed framework.

3. EXPERIMENTS

We compared our proposed ASENet with two uninformed separation methods, i.e., conventional PIT-based speech separation (PIT) and our introduced PIT-based internal separation (Internal-PIT). In addition, we also compared our method with SpeakerBeam [10, 18], which is an informed target speech extraction method that adapts the network behavior based on the auxiliary information to extract the target speaker's speech. We employed the scaling adaptation layer-based SpeakerBeam described in [18], which adapts the intermediate layer's output by multiplying it element by element with the output of the sequence summary network (similar to Eq. (10)).

3.1. Experimental conditions

3.1.1. Data

We evaluated our proposed method on the WSJ0-2mix dataset [4], which consists of two-speaker mixtures generated by mixing utterances from the WSJ0 corpus [19] at signal-to-noise ratio (SNR) between 0 dB and 5 dB. The sampling frequency was 8 kHz.

The training set consists of 20000 mixtures (30 hours) from 101 speakers. The development set consists of 5000 mixtures (10 hours) from the same 101 speakers. The evaluation set consists of 3000 mixtures (5 hours) from 18 different speakers. For the auxiliary information in informed extraction, we used a randomly selected utterance (different from those in the mixture) from each speaker in the mixture for both training and testing.

3.1.2. Settings

We used magnitude spectrograms as an input for the mask estimation networks, which are computed using STFT with 64 ms window length and 16 ms window shift.

For all the experiments, we used a 3-layer BLSTM network with 512 units as shown in Figs. 1 and 2. Each BLSTM layer is followed by a linear projection layer with 512 units to combine the forward and backward LSTM outputs, which is based on the ESPNet implementation [20]. We employed a tanh activation after the last projection layer and 1 fully connected layer to output an amplitude mask estimated with a sigmoid activation function. The main difference between the evaluated methods is the position of the separate linear layers (see Figs. 1 and 2).

For the sequence feature extraction network (i.e., $\text{MLP}(\cdot)$ in Eqs. (8)-(10)), we used a network with 2 fully connected layers with 200 units and ReLU activations, followed by 1 linear output layer with 200 units for ASENet and 512 units for SpeakerBeam. We set the dimension of the attention inner product (i.e., the dimension of $\mathbf{w}$ in Eq. (6)) at 200, and the sharpening factor α at 2.

Table 1. SDR (dB) for different separation/extraction methods.

Method	Position	SDR (uninfo)	SDR (info)
Mixture	–	0.2	
PIT	3	8.6	–
Internal-PIT	1	8.8	–
	2	8.7	–
ASENet	1	9.2	9.6
	2	8.7	9.1
SpeakerBeam	1	–	9.6

We adopted the Adam algorithm [21] for optimization with an initial learning rate of 0.0001 and used gradient clipping [22]. The training procedure was stopped after 200 epochs. For the proposed ASENet, we set the multitask interpolation weight ϵ at 0.5. Note that Internal-PIT corresponds to the ASENet ($\epsilon = 0.0$).

As an evaluation metric, we used the signal-to-distortion ratio (SDR) of BSSeval [23]. To evaluate the performance of the uninformed separators, we used the oracle permutation (permutation achieving best scores) to compute the SDR. This would thus correspond to the upper bound performance of a system undertaking target speaker identification on top of speech separation [24].

3.2. Experimental Results

Table 1 shows the SDR obtained with the four methods and an unprocessed mixture. "Position" denotes the number of BLSTM layers of the separation network (or the position of the adaptation layer for SpeakerBeam). "(info)" and "(uninfo)" denote whether or not auxiliary information $\mathbf{X}_s^{\mathrm{AUX}}$ is provided. In other words, it denotes whether each method works as an informed extractor or an uninformed separator.

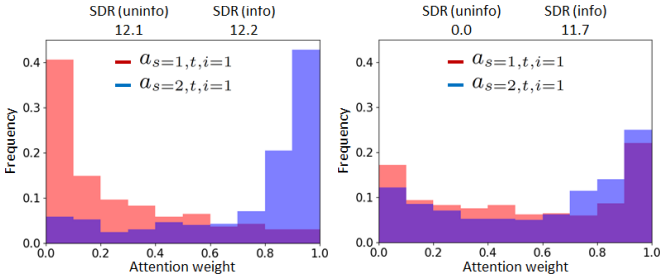
From Table 1, we confirmed that Internal-PIT achieved comparable or slightly better performance compared with PIT. ASENet in the uninformed separator mode achieved better performance than PIT and Internal-PIT. This result demonstrated that the use of auxiliary information in the training stage based on the multitask learning scheme could improve the uninformed separation performance.

ASENet in the informed extractor mode successfully improved the performance compared to the uninformed separation methods and achieved comparable performance to SpeakerBeam, which is specially designed for the target speaker extraction. Note that this result demonstrated that our proposed ASENet performed better than uninformed separation with oracle permutation (oracle target speaker assignment, i.e. speaker identification error of 0%), which confirms the potential of performing the speaker selection process internally. These results proved the effectiveness of our proposed framework, which realized both uninformed separation and informed extraction behaviors using a unified attention-based architecture.

Finally, focusing on "Position" for ASENet, we confirmed that the allocation of the separate linear layers and attention module at a lower layer is important as regards achieving better performance.

3.3. Analysis of Attention Module Behaviors

To observe the behavior of our proposed speaker attention module, we analyzed the histograms of the attention weights. Figure 3 shows histograms of the time-variant attention weights for two typical mixtures. The attention weights for the first speaker $\{a_{s=1,t,i}\}_{t=1}^{T}$ and second speaker $\{a_{s=2,t,i}\}_{t=1}^{T}$ are shown in red and blue, respectively. The histograms indicate which separate internal embedding vectors $\{\mathbf{z}_{it}\}_{i=1}^{I=2}$ the speaker attention module attended over the time frames. For the purpose of discussion, we focused on the case where $i = 1$, and therefore $a_{s,t,i=1} = 1$ means that the speaker attention module attends the first internal embedding vector $\mathbf{z}_{i=1,t}$ to extract the speech of the s-th target speaker at time step t. Moreover, $a_{s,t,i=1} = 0$ means that it attends the second internal embedding vector $\mathbf{z}_{i=2,t}$ (see Fig. 2-(a) and Eqs. (4) and (7)).



(a) Mixture 1 (b) Mixture 2

Fig. 3. Histograms of time-variant attention weights for two typical mixtures.

Figure 3-(a) shows a mixture for which the separation performance was similar with and without auxiliary information. As seen in the figure, the attention module mainly attended to one of the two internal embedding vectors $\{\mathbf{z}_{it}\}_{i=1}^{I=2}$. This distribution resembles that of the uninformed separation mode in the proposed method, where we forced the use of one of the two internal embedding vectors. Therefore, for this mixture, the separation performance did not improve significantly when using the auxiliary information. This result suggests that the developed attention module had the potential to automatically adopt behavior close to the uninformed separation mode, when the uninformed separation had already achieved sufficient separation quality.

Figure 3-(b) shows a mixture for which the separation performance greatly improved when using auxiliary information. As seen in the figure, in this case the attention module attended to both the internal embedding vectors $\{\mathbf{z}_{it}\}_{i=1}^{I=2}$ and interpolated both of them more frequently. This result suggested that the developed attention module had the potential to recover the extracted speech quality by utilizing the auxiliary information, even when the uninformed separation completely failed to separate the mixture.

4. CONCLUSION

In this paper, we proposed a unified framework for neural network-based uninformed speech separation and informed speech extraction, and a training procedure based on multitask learning scheme. Our experimental results obtained with the WSJ0-2mix dataset showed that we could successfully perform both separation and extraction using a single model. In particular, our proposed ASENet achieved better performance than a conventional blind speech separation method, i.e., PIT, even with oracle permutation and performed comparably to a conventional speaker-aware speech extraction method, i.e., SpeakerBeam. Moreover, thanks to the multitask learning scheme, we were able to improve the uninformed separation performance compared with a conventional PIT-based approach.

In this paper, we evaluated our proposed method only for two-speaker mixtures without noise. However, because our proposed method follows the conventional PIT-based separation framework, it could be applied to more than three-speaker mixtures with noise [7]. Evaluating the effectiveness of our proposed method for such challenging setups will form part of our future research work.

5. REFERENCES

[1] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, and Brian Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] Wayne Xiong, Lingfeng Wu, Fil Alleva, Jasha Droppo, Xuedong Huang, and Andreas Stolcke, "The Microsoft 2017 conversational speech recognition system," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5934–5938.

[3] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal, "The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Interspeech*, 2018, pp. 1561–1565.

[4] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31–35.

[5] Zhuo Chen, Yi Luo, and Nima Mesgarani, "Deep attractor network for single-microphone speaker separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 246–250.

[6] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245.

[7] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1901–1913, 2017.

[8] Keisuke Kinoshita, Lukas Drude, Marc Delcroix, and Tomohiro Nakatani, "Listening to each speaker one by one with recurrent selective hearing networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5064–5068.

[9] Katerina Zmolikova, Marc Delcroix, Keisuke Kinoshita, Takuya Higuchi, Atsunori Ogawa, and Tomohiro Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in *Interspeech*, 2017, pp. 2655–2659.

[10] Marc Delcroix, Katerina Zmolikova, Keisuke Kinoshita, Atsunori Ogawa, and Tomohiro Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5554–5558.

[11] Jun Wang, Jie Chen, Dan Su, Lianwu Chen, Meng Yu, Yanmin Qian, and Dong Yu, "Deep extractor network for target speaker recovery from single channel speech mixtures," in *Interspeech*, 2018, pp. 307–311.

[12] Yang Shao and DeLiang Wang, "Model-based sequential organization in cochannel speech," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 14, no. 1, pp. 289–298, 2006.

[13] Rich Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations (ICLR)*, 2015.

[15] Karel Veselỳ, Shinji Watanabe, Kateřina Žmolíková, Martin Karafiát, Lukáš Burget, and Jan Honza Černockỳ, "Sequence summarizing neural network for speaker adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5315–5319.

[16] Kateřina Žmolíková, Marc Delcroix, Keisuke Kinoshita, Takuya Higuchi, Atsunori Ogawa, and Tomohiro Nakatani, "Learning speaker representation for neural network based multichannel speaker extraction," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 8–15.

[17] Yi Luo, Zhuo Chen, John R Hershey, Jonathan Le Roux, and Nima Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 61–65.

[18] Marc Delcroix, Katerina Zmolikova, Tsubasa Ochiai, Keisuke Kinoshita, Araki Shoko, and Tomohiro Nakatani, "Compact network for SpeakerBeam target speaker extraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, (to be submitted).

[19] John Garofolo, David Graff, Doug Paul, and David Pallett, "CSR-I (WSJ0) Complete LDC93S6A," *Philadelphia: Linguistic Data Consortium*, 1993.

[20] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Yalta, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Interspeech*, 2018, pp. 2207–2211.

[21] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[22] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning (ICML)*, 2013, pp. 1310–1318.

[23] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing (TASLP)*, vol. 14, no. 4, pp. 1462–1469, 2006.

[24] Lukas Drude, Thilo von Neumann, and Reinhold Haeb-Umbach, "Deep attractor networks for speaker re-identification and blind source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 11–15.