End-to-end Music-mixed Speech Recognition

Jeongwoo Woo*, Masato Mimura*, Kazuyoshi Yoshii* and Tatsuya Kawahara*
* Kyoto University, Kyoto, Japan
{woo, mimura, yoshii, kawahara}@sap.ist.i.kyoto-u.ac.jp

***Abstract*—Automatic speech recognition (ASR) in multimedia content is one of the promising applications, but speech data in this kind of content are frequently mixed with background music, which is harmful for the performance of ASR. In this study, we propose a method for improving ASR with background music based on time-domain source separation. We utilize Conv-TasNet as a separation network, which has achieved state-of-the-art performance for multi-speaker source separation, to extract the speech signal from a speech-music mixture in the waveform domain. We also propose joint fine-tuning of a pre-trained Conv-TasNet front-end with an attention-based ASR back-end using both separation and ASR objectives. We evaluated our method through ASR experiments using speech data mixed with background music from a wide variety of Japanese animations. We show that time-domain speech-music separation drastically improves ASR performance of the back-end model trained with mixture data, and the joint optimization yielded a further significant WER reduction. The time-domain separation method outperformed a frequency-domain separation method, which reuses the phase information of the input mixture signal, both in simple cascading and joint training settings. We also demonstrate that our method works robustly for music interference from classical, jazz and popular genres.**

I. Introduction

Automatic speech recognition (ASR) for multimedia content such as movies, dramas, broadcast news and other online videos is useful for automatic subtitle generation. Speech signals in these kinds of real-world multimedia content are frequently mixed with background music in order to make listeners immersed in. However, the music in speech signal causes significant performance degradation in ASR [1][2].

Despite its importance, there is limited research on music-mixed ASR compared to noisy and reverberant ASR. In a small number of existing approaches, it was tackled by removing music interference with unsupervised separation methods such as NMF [3] and robust PCA [4] or denoising autoencoders [5][6]. In [3], NMF was employed for extracting speech signal from a speech-music mixture and ASR was performed with a GMM-HMM acoustic model trained on clean data. Zhao et al. [5] trained a denoising autoencoder using mixture data as input and clean speech as target. They used a hybrid DNN-HMM acoustic model for ASR. In general, music-mixed ASR is very challenging, since the music interference is highly non-stationary and at low signal-to-noise ratios.

Recently a fully-convolutional neural network called a convolutional time-domain audio separation network (Conv-TasNet) [7], which operates in the waveform domain, has shown to achieve an excellent performance for multi-speaker source separation. Due to its high performance and flexibility, it was also adopted for many other tasks. Kinoshita et al. [8] used a Conv-TasNet for speech denoising and dereverberation and achieved better ASR performance than a frequency-domain counterpart method. Défossez et al. [9] used a Conv-TasNet and its variant for singing voice separation.

In this paper, motivated by the recent progress in the time-domain source separation mentioned above, we investigate the use of a Conv-TasNet for the purpose of improving speech recognition with background music interference. Thus, we adopt a Conv-TasNet for the speech-music separation task and utilize an attention-based [10] ASR model for the speech recognition task, whereas the previous studies utilized an HMM-based ASR model.

This allows us to combine the Conv-TasNet front-end with the attention-based ASR back-end to form an end-to-end music-mixed ASR system that directly operates on raw waveform features. A similar network for the multi-speaker source separation task has been used in [11]. We retrain the pre-trained front-end and the back-end models jointly and evaluate our method through ASR experiments using speech data mixed with background music from a wide variety of Japanese animations.

The rest of the paper is organized as follows. We introduce speech-music separation with Conv-TasNet in Section II, describe joint training of Conv-TasNet and ASR model in Section III, describe the experimental details in Section IV, show the results of the experiments and discussions in Section V, and conclude the paper in Section VI.

II. Speech-Music Separation and ASR

A. Time-domain speech-music separation with Conv-TasNet

Conv-TasNet [7] is a single-channel source separation model which generates waveforms for a fixed number of speakers from a mixture waveform. For speech-music separation, it outputs an estimated speech audio stream and an estimated music audio stream in the time domain, given an input mixture signal $\mathbf{x}$:

$$[\mathbf{x}_s^{(enh)}, \mathbf{x}_m^{(enh)}] = ConvTasNet(\mathbf{x})$$

The model works directly on raw waveforms using an encoder and a decoder that can be learned in place of STFT and ISTFT. This makes it possible to obtain the phase information to reconstruct a waveform and to propagate the gradients from feature extraction to the waveform reconstruction part. In the training stage, the scale-invariant-source-to-distortion ratio (SI-SDR) [12] loss is optimized directly on the time domain.

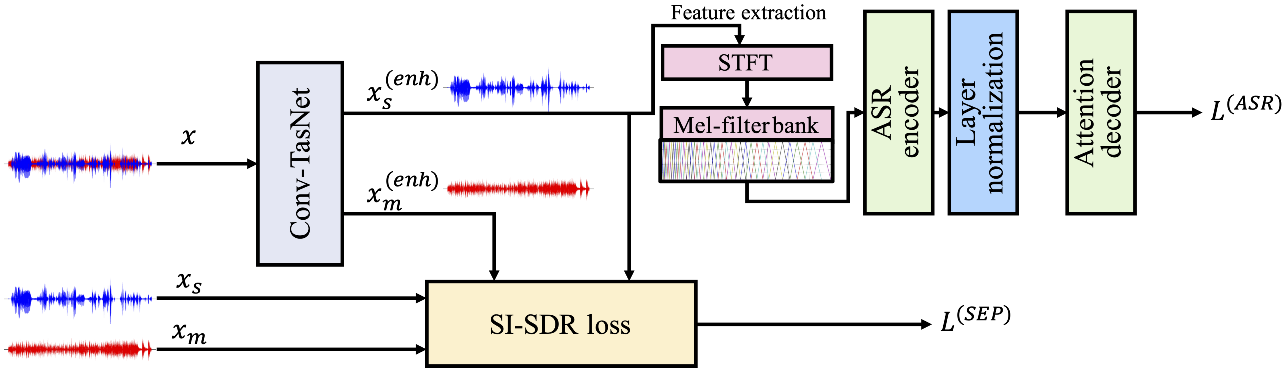

Fig. 1. The proposed network architecture

We refer to the loss for the Conv-TasNet as $L^{(SEP)}$,

$$L^{(SEP)} = -\frac{\text{SI-SDR}_s + \text{SI-SDR}_m}{2}$$

where SI-SDR$_s$ and SI-SDR$_m$ are the SI-SDR loss of estimated speech and music, respectively.

Note that Conv-TasNet was originally proposed for multi-speaker source separation, and usually requires permutation invariant training (PIT) [13] to solve permutation ambiguity of the output speech sources. Since the front-end outputs are speech and music in this work, we can fix the output order of the network unlike in [11], and do not need PIT.

B. Attention-based ASR model

For ASR, we use an attention-based model. We adopt an encoder-decoder architecture similar to [10]. This architecture consists of two distinct networks of encoder and decoder. The encoder transforms a sequence of acoustic features to a sequential representation, and the decoder predicts a label sequence from the encoded sequential representation using the attention mechanism. We refer to the loss for the ASR component as $L^{(ASR)}$ which is the cross-entropy loss function with label smoothing.

III. Joint training of Conv-TasNet and ASR

We propose to combine the Conv-TasNet speech-music separation front-end with the attention-based ASR back-end, as shown in Fig. 1. The input mixture $\mathbf{x}$ is separated by the front-end and the output speech audio stream is processed by the ASR back-end.

Though music-mixed ASR can be performed by simply cascading the independently pre-trained front-end and back-end models, the speech-music separator produces unknown artifacts that degrade the performance of ASR. According to [11], such mismatches can be mitigated by joint fine-tuning the entire model.

To propagate gradients from the ASR back-end to the front-end, we extract acoustic features for ASR directly from the speech waveform estimated by the front-end. Specifically, log Mel-scale filterbank (lmfb) features obtained by applying an lmfb to an amplitude spectrogram generated from the speech waveform using STFT are used as acoustic features for the ASR model. Since these processes are all differentiable, gradients can be propagated from the ASR back-end to the front-end. In this end-to-end model, however, the acoustic features for ASR cannot be normalized beforehand. Thus, we insert the layer normalization [14] behind the encoder of the ASR back-end. The losses for the front-end and the back-end are combined as

$$L = L^{(SEP)} + \alpha L^{(ASR)}$$

where α is an empirically chosen weight for the ASR loss.

IV. Experimental details

This section presents the experimental details. We describe the dataset and the ASR and separation models used for experiments. We also explain the detail of the joint training. We implemented our models in PyTorch.

A. Dataset

Experimental mixture data were generated by mixing utterances from speech database with background music. Both speech and music are sampled at 16 kHz. As the speech database, we used the Corpus of Spontaneous Japanese Academic Presentation Speech (CSJ-APS). The CSJ-APS has a duration of around 260 hours and consists of live recordings of academic presentations in nine different academic societies. The societies range from engineering, humanities, and social and behavioral sciences. For background music, we used around 30 hours of background music used in Japanese animations.

For the training dataset, we added background music to the speech with randomly sampled source-to-noise ratio (SNR) levels from a normal distribution with a mean of 0 dB and a standard deviation of 5 dB. For the test dataset, we added background music from animations not used for the training dataset to the speech of official CSJ-APS testset 1 with various SNR levels such as 5 dB, 0 dB and -5 dB. This test dataset is referred as the CSJ-anime.

We also evaluated on the dataset mixed with music of particular genres from the Real World Computing Music

TABLE I
WER and SDR on CSJ-anime for different variants of fine-tuning of joint model.

WER: Word error rate, SDR: Signal-to-distortion ratio

	fine-tune		WER(%)					SDR(dB)		
Model	SEP	ASR	Clean	5 dB	0 dB	-5 dB	avg	5 dB	0 dB	-5 dB
Clean	–	–	11.25	46.62	78.96	93.33	72.97	–	–	–
Mixture	–	–	12.29	15.26	19.63	31.57	22.15	–	–	–
Frequency Domain	–	–	11.24	16.03	22.76	37.84	25.54	11.95	9.69	7.23
separation Model	✓	–	11.73	14.91	18.85	30.15	21.30	11.07	8.76	6.21
+	–	✓	11.70	13.85	17.14	26.48	19.16	11.95	9.69	7.23
Clean ASR Model	✓	✓	11.61	13.47	16.56	23.90	17.98	11.55	9.29	6.79
Frequency Domain	–	–	12.28	15.16	19.66	31.20	22.01	11.95	9.69	7.23
separation Model	✓	–	12.49	15.17	19.56	30.66	21.80	7.01	2.69	-1.99
+	–	✓	12.22	13.96	16.92	25.91	18.93	11.95	9.69	7.23
Mixture ASR Model	✓	✓	12.41	14.77	18.83	28.99	20.86	7.93	3.88	-0.75
Time Domain	–	–	11.40	14.35	18.23	28.59	20.39	20.74	18.17	15.40
separation Model	✓	–	11.26	**12.90**	15.80	23.13	17.28	20.65	17.94	15.07
+	–	✓	12.88	14.23	15.86	21.30	17.13	20.74	18.17	15.40
Clean ASR Model	✓	✓	12.81	13.62	15.05	19.30	15.99	20.65	18.03	15.20
Time Domain	–	–	12.32	13.95	16.72	23.43	18.03	20.74	18.17	15.40
separation Model	✓	–	14.45	15.48	18.41	24.76	19.55	15.05	13.29	11.16
+	–	✓	12.84	13.87	15.84	19.53	16.41	20.74	18.17	15.40
Mixture ASR Model	✓	✓	12.96	13.30	**15.01**	**18.31**	**15.54**	20.52	17.85	14.92

Database (RWC-MDB) such as classical, jazz and popular. Among them, the popular music has lyrics. Music signal was added to the CSJ-APS testset speech with SNR levels of 5 dB, 0 dB and -5 dB. This test datasets are referred as the CSJ-genre.

B. Baseline model

We trained two types of baseline ASR models without the separation front-end, which differ in training data; One referred to as clean ASR is trained on clean speech data. The other referred to as mixture ASR is trained on speech-music mixture data.

C. Speech-music separation front-end

In this experiment, we compare the following two different speech-music separation networks, which operate in the time-domain and in the frequency-domain, respectively.

1) Conv-TasNet network: We investigated the performance of Conv-TasNet for speech-music separation that uses a similar configuration to the original Conv-TasNet [7]. In particular, following the hyper-parameter notations in the original paper [7], we set the hyper-parameters N=256, L=20, B=256, H=512, P=3, X=8, R=4. We used the Adam optimizer to train the network with a learning rate of 1e-3. This network is referred as the time-domain separation model.

2) Frequency-domain BiLSTM network: We compare the time-domain separation model with a frequency-domain BiLSTM network, which uses a mask estimation network consisting of five BiLSTM layers with 320 units followed by a linear layer with sigmoid activation. The input of the mask estimation network consists of an amplitude spectrogram computed with the STFT with a hanning window of 25 ms and a shift of 10 ms. We reconstructed the waveform of the predicted speech signal by applying the ISTFT to the masked spectrogram of the input mixture. We reused the phase information of the mixture. We set a learning rate of Adam to 1e-3. This network is referred as the frequency-domain separation model.

D. ASR back-end configuration

The attention-based ASR [10] back-end uses two CNN layers with a stride of 2 followed by five BiLSTM layers with 320 units for an encoder, layer normalization [14] and two LSTM layers with 320 units for a decoder. 40-channel lmfb features are used as acoustic features for ASR. The output of the decoder is a sequence of subwords defined by the byte-pair encoding (BPE) [15]. The number of the BPE units is 9,515.

E. Joint training

As explained in Section III, we design the entire end-to-end model in order that the gradients of the top-level ASR loss can be propagated down to the front-end. We convert the waveform output of the front-end to a sequence of 40-channel lmfb features before feeding it to the back-end. We used the STFT implemented in PyTorch through which we can propagate gradients. The STFT is set with a hanning window of 25ms and a shift of 10ms which is consistent with feature extraction for pre-training the ASR model.

We compare three different variants of joint fine-tuning: fine-tuning the front-end by propagating gradients through the ASR back-end but only updating the front-end parameters, fine-tuning only the ASR back-end on the enhanced signals, and jointly fine-tuning both components. For all variants of joint fine-tuning, the weight for the ASR loss α is set to 2 and a learning rate of Adam is set to 1e-4.

V. Results and Discussion

A. *Result of joint fine-tuning*

We evaluated the joint model with combinations of two kinds of front-ends of time-domain separation and frequency-domain separation model, and two kinds of back-ends of clean ASR and mixture ASR. Moreover, fine-tuning was conducted in three cases: the separation front-end only, the ASR back-end only, and both. We evaluate the performance in terms of source-to-distortion ratio (SDR) [16] and word error rate (WER). The results of these variants are listed for comparison in Table I. The comparison among the models is primarily based on the average WER over all SNR levels.

Without any front-end processing, the average WER of clean ASR was 72.97% and that of mixture ASR was 22.15%. It is notable that cascading the independently trained front-end and back-end models already give a better performance than the baseline models whether in the time-domain or in the frequency-domain. In this case, the back-end models based on mixture ASR achieve a relative WER reduction of at least 11% in both domains.

Fine-tuning the ASR back-end while freezing the front-end can further reduce the WER for all combinations of cascading models. Joint fine-tuning of the frequency-domain separation with clean ASR model achieves the average WER of 17.98%, which is the best among the frequency-domain methods. Joint fine-tuning of the frequency-domain separation with mixture ASR model did not improve from fine-tuning only the ASR back-end and significantly degrades the speech-music separation performance. In this case, fine-tuning of the front-end makes it unable to separate speech and music because the mixture ASR back-end can already deal with music-mixed speech input. On the other hand, joint fine-tuning of the time-domain separation with clean ASR and mixture ASR models were both effective, and achieved the average WER of 15.99% and 15.54% respectively. The time-domain separation performance was not degraded so much by combination with mixture ASR in terms of SDR. The joint fine-tuning with the mixture ASR back-end achieved a relative WER reduction of 29.8% from the baseline mixture ASR model alone. This joint model resulted in the best performance (average WER of 15.54%) among all settings.

Fine-tuning only the separation front-end is not so effective as fine-tuning the ASR back-end for all combinations of the joint models. It may be because ASR can adapt to the artifacts produced by the front-end.

B. *Result of particular music genres*

Table II shows the results of the baseline mixture ASR and the time-domain joint model for unseen genres in the training data. For all cases of the music genres and SNR levels, our best model improved the WER from the baseline. The results on the classical and jazz music datasets show the similar performance to that of the CSJ-anime. The popular music has lyrics which degrade the performance, but this result still shows that the proposed joint approach is also effective on the data with vocal music. In general, the results in Table II demonstrate the generalization capability of the proposed method for unseen interferences.

TABLE II
WER and SDR on CSJ-genre for Mixture ASR model and Joint model of Time-domain separation with Mixture ASR

WER: Word error rate, SDR: Signal-to-distortion ratio

		WER(%)			SDR(dB)		
Model	Genre	5 dB	0 dB	-5 dB	5 dB	0 dB	-5 dB
Mixture	Classical	14.43	18.20	26.46			
	Jazz	14.46	17.50	25.73			
	Popular	16.12	22.64	37.18			
Joint model (our best)	Classical	13.33	14.37	17.32	21.00	18.29	15.27
	Jazz	13.29	14.43	17.48	20.83	18.05	14.94
	Popular	13.98	16.19	22.60	19.67	16.83	13.56

VI. Conclusions

We have proposed to combine a time-domain speech-music separation model Conv-TasNet with an attention-based ASR model to form an end-to-end music-mixed speech recognizer. We show that time-domain separation is better than frequency-domain separation, and pre-training the ASR model on the mixture data is effective. Joint fine-tuning further significantly improved the performance. The effectiveness was confirmed with a variety of music genres.

VII. Acknowledgements

This research was supported by NII CRIS collaborative research program operated by NII CRIS and LINE Corporation.

References

[1] P. Vanroose and K. Arenberg, "Blind source separation of speech and background music for improved speech recognition," in *The 24th Symposium on Information Theory,* 2003, pp. 103-108.

[2] T. Hughes and T. Kristjansson, "Music models for music-speech separation," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Kyoto, 2012, pp. 4917-4920

[3] C. Demir, M. Saraçlar, and A. T. Cemgil, "Single-Channel Speech-Music Separation for Robust ASR with Mixture Models," in *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 21, no. 4, pp. 725-736, April 2013

[4] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Kyoto, 2012, pp. 57-60

[5] M. Zhao, D. Wang, Z. Zhang, and X. Zhang, "Music Removal by Convolutional Denoising Autoencoder in Speech Recognition," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA),* Hong Kong, 2015, pp. 338-341

[6] J. Malek, J. Zdansky, and P. Cerva, "Robust Automatic Recognition of Speech with background music," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* New Orleans, LA, 2017, pp. 5210-5214

[7] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 27, no. 8, pp. 1256-1266, Aug. 2019

[8] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, "Improving Noise Robust Automatic Speech Recognition with Single-Channel Time-Domain Enhancement Network," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Barcelona, Spain, 2020, pp. 7009-7013

[9] A. Défossez, N. Usunier, L. Bottou, and F. Bach, “Music Source Separation in the Waveform Domain,” *arXiv preprint arXiv:1911.13254,* 2019

[10] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” in *Advances in neural information processing systems,* 2015, pp. 577-585.

[11] T. von Neumann et al., “End-to-End Training of Time Domain Audio Separation and Recognition,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Barcelona, Spain, 2020, pp. 7004-7008

[12] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR – Half-baked or Well Done?,” in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Brighton, United Kingdom, 2019, pp. 626-630

[13] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, “Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks,” in *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, no. 10, pp. 1901-1913, Oct. 2017

[14] J. L. Ba, J. R. Kiros, and G. E. Hinton. “Layer Normalization,” *arXiv preprint arXiv:1607.06450,* 2016

[15] R. Sennrich, B. Haddow, and A. Birch, “Neural Machine Translation of Rare Words with Subword Units,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Berlin, Germany, vol 1, pp.1715-1725, Aug, 2016

[16] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” in *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, no. 4, pp. 1462-1469, July 2006