SPECTRON: TARGET SPEAKER EXTRACTION USING CONDITIONAL TRANSFORMER WITH ADVERSARIAL REFINEMENT

Tathagata Bandyopadhyay

Visual Computing Lab, Technical University of Munich, Germany
tathagata.bandyopadhyay@tum.de

ABSTRACT

Recently, attention-based transformers have become a de facto standard in many deep learning applications including natural language processing, computer vision, signal processing, etc.. In this paper, we propose a transformer-based end-to-end model to extract a target speaker's speech from a monaural multi-speaker mixed audio signal. Unlike existing speaker extraction methods, we introduce two additional objectives to impose speaker embedding consistency and waveform encoder invertibility and jointly train both speaker encoder and speech separator to better capture the speaker conditional embedding. Furthermore, we leverage a multi-scale discriminator to refine the perceptual quality of the extracted speech. Our experiments show that the use of a dual path transformer in the separator backbone along with proposed training paradigm improves the CNN baseline by 3.12 dB points. Finally, we compare our approach with recent state-of-the-arts and show that our model outperforms existing methods by 4.1 dB points on an average without creating additional data dependency.

Index Terms— Target Speaker Extraction, Speech Separation, Transformers, DPTNet, Adversarial Refinement

1. INTRODUCTION

Blind source separation [1, 2] or more specifically speech separation [3, 4], in general refers to splitting a mixed signal to all of its constituting component audios; i.e., separating out all the speakers from a mixed speech. With the advent of deep learning, many different approaches and corresponding architectures that use convolutional [5], recurrent [6], or transformer [7, 8] models have been proposed to tackle single channel speech separation. Inspired from two breakthrough papers "TasNet" [9] and "ConvTasNet" [5], the majority of supervised speech separation approaches follow a common high-level structure including a *waveform encoder* to transform the audio input to a spectrogram-like latent representation, a *separator backbone* to separate the speech sources in latent space, and then a *waveform decoder* to generate the wave forms for individual speakers.

General speech separation for known and small number of speakers has seen great success in recent years with deep neural networks; e.g., "SepFormer" [8], "DPTNet" [7], "Sandglasset" [10], "Conv-Tasnet" [5]. However, knowing the number of speakers a priory in real-world mixtures is an impractical assumption. This is further exacerbated as there is no easy way to resolve the ambiguity between the separated channels due to lack of specific ordering amongst the speakers.

On the contrary, often many downstream applications such as a voice activated virtual assistant need to extract the speech of a pre-enrolled specific speaker from mixed speech audio inputs, where it is quite practical to assume, without loss of generality, to have the clean reference speech sample pre-recorded during speaker enrollment. In the literature this task is referred to as target speaker extraction.

In this work, we propose a target speaker extraction system which takes a monaural multi-talker speech mixture and a clean reference speech sample of the target speaker as inputs and extracts out the target speaker's speech from the mixture. A state-of-the-art speaker verification module [11] is used to obtain the representative speaker embedding which is fed into the separator backbone as a condition alongside the spectrogram-like representation of the input speech mixture. The separator backbone conditionally estimates a mask on the transformed representation to suppress the interfering speeches. Finally, a waveform decoder is used to generate clean speech waveform[1] from this masked representation.

Like most of the existing methods for speech separation or speaker extraction [9, 5, 6, 7, 12], we learn an internal representation based on a masking approach and further leverage a speaker extraction framework consisting of a speaker encoder and a separator module. However, instead of defining custom speaker encoder as in [13] or custom separation module as in [14], we use a current SOTA speaker verification module [11] as speaker encoder and adapt one of the transformer based general speech separation SOTA models known as "DPTNet" [7] for separation module, to combine the best of both worlds. Like many speaker extraction systems [5, 12, 7], we too use SI-SNR [15] loss as our main

[1] Output samples: `https://tatban.github.io/spec-res/`

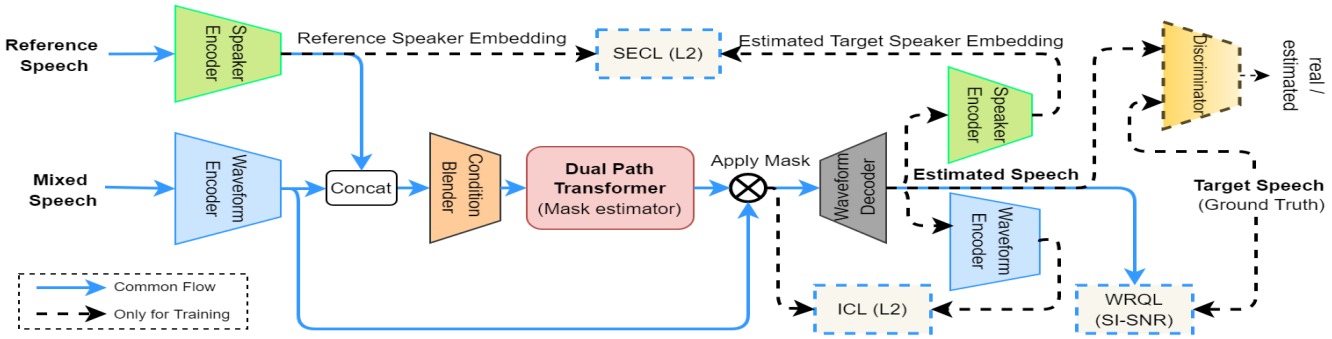

Fig. 1. Spectron framework: same color blocks refers to shared weights; dashed boxes refers to objective functions. Speaker Embedding Consistency Loss (SECL) and waveform encoder decoder Inverse Consistency Loss (ICL) are realized as MSE Loss, where SI-SNR [15] is used for Waveform Reconstruction Quality Loss (WRQL) and "MSD" [16] for discriminator loss.

Waveform Reconstruction Quality Loss (WRQL). However, unlike the existing approaches, we introduce *Inverse Consistency Loss* (ICL) to impose invertibility of learned waveform encoder and decoder and *Speaker Embedding Consistency Loss* (SECL) to impose similarity between reference speaker sample and extracted speech embeddings. Finally, we use a *Multi-Scale Discriminator* (MSD) from [16] to further improve the perceptual quality of the extracted speech.

2. RELATED WORK

Speech Non-Speech separation. This is also known as speech enhancement or speech denoising [17, 18, 19, 20] and works with single speaker assumption. In contrast, our method works with multi-speaker assumption and extracts the clean speech of the speaker of interest.

Multi-Speech Separation. It assumes there are n $(n > 1)$ speakers speaking simultaneously and the goal is to separate out all the speakers into n different channels at once [8, 7, 6, 5, 9]. Our approach is different as we aim to conditionally extract out only the target speaker's speech.

Target Speaker Extraction. This is the general category of our approach and there are quite a few existing works with common high level structure consisting of a speaker encoder module and a speech separation module. Voice Filter [14] uses a pre-trained speaker verification model from [11] as a fixed speaker encoder along with a fixed time frequency representation (STFT and inverse STFT) based CNN LSTM architecture which involves complex phase estimation and long term sequential dependency modeling, making it inefficient and limited in performance. Atss-Net [13] improves over it by using multi-head attention based separator jointly trained with ResNet-18 based speaker encoder. However, this approach still suffers from complex phase estimation related shortcomings due to STFT and inverse STFT. X-TasNet [12] significantly pushed further previous two SOTAs by using learnable time frequency like representation as first proposed in [5]. However, CNN based separator architecture is not efficient enough to simultaneously capture utterance level and long term dependencies due to fixed receptive field of the convolution kernels. Our approach differs from existing approaches as we utilize five things together namely: i) *learned spectrogram-like representation*, ii) *transformer based separator backbone*, iii) *joint training of speaker encoder and speech separator*, iv) *introduction of two additional objectives to impose speaker embedding consistency and waveform encoder invertibility* and v) *use of multi-scale discriminator ("MSD") to refine the extracted speeches*.

3. METHOD

A schematic diagram of our proposed pipeline is shown in Fig. 1. We first discuss the Spectron architecture and finally elaborate on design of different objective (loss) functions involved in training.

3.1. Model Architecture

Spectron consists of two high level components namely *speaker encoder* and *speech separator*.

3.1.1. Speaker Encoder

This block is responsible for computing a speaker representative embedding $e \in \mathbb{R}^d$ from a clean reference speech $r \in \mathbb{R}^{1\times t}$. Speaker representation learning is a crucial part of any speaker verification system. Hence, following [14] and [12], we too adopt pre-trained GE2E [11] speaker verification model as our speaker encoder. However, unlike previous approaches, we jointly train it with the separator module. We also keep the speaker embedding dimension to 256 not to alter the original GE2E architecture.

3.1.2. Speech Separator

This module takes input speech mixture $i \in \mathbb{R}^{1\times t}$ along with reference speech embedding $e \in \mathbb{R}^d$ and conditionally esti-

mates $\hat{s} \in \mathbb{R}^{1\times t}$, the speech of the target speaker selected by the reference speech embedding e. Speech Separator module of Spectron is adapted from "DPTNet" [7] and has following sub-modules.

Waveform Encoder is 1D CNN with N filters of kernel size k and stride st. It is a learnable mapping from $x \in \mathbb{R}^{1\times t}$ to $X \in \mathbb{R}^{N\times T}$ where x is the input waveform and X is a spectrogram-like internal representation. Following "DPTNet" [7], we set $N = 64$, $k = 16$ and $st = 8$ and pass the output of this mapping through ReLU activation.

Condition Blender is a 1D CNN with both kernel size and stride equal to 1 and it takes the Waveform Encoder output concatenated with the speaker embedding to reduce its dimension to match it with the required input dimension of the Separator core. This helps us to keep the separator core architecture unaltered from "DPTNet" [7]. Number of filters used in it is equal to the input channel dimension of the separator core which is 64 in our case.

Separator Core is the same dual path transformer architecture from "DPTNet" [7] paper with only difference in number of attention heads. We use 8 attention heads instead of originally proposed 4, keeping other configuration unaltered. This block estimates a target speaker specific conditional mask of same shape of waveform encoder output. This mask is multiplied element-wise with the waveform encoder output to separate out the target speaker in this transformed spectrogram-like space.

Waveform Decoder is intuitively an inverse mapping of the waveform encoder to produce the target speech waveform from the masked internal representation. However, instead of strictly enforcing this inverse property, we keep it as learnable just like "DPTNet" [7] but introduce an additional loss function to softly impose invertibility.

3.2. Design of Training Objectives

We use four objective (loss) functions as shown with dashed boxes in Fig. 1. Descriptions of those are as follows:

3.2.1. Waveform Reconstruction Quality

As the main goal of speaker extraction is to extract out the clean speech of the target speaker from the input mixture, we need to improve the waveform reconstruction quality (WRQ) by maximizing signal-to-noise ratio (SNR) of the estimated speech. To avoid the scale dependency we use SI-SNR (scale invariant SNR also known as SI-SDR) as proposed by [15] and use negative value of it to formulate as minimization objective. Let us consider $s, \hat{s} \in \mathbb{R}^{1\times t}$ represent ground truth and estimated speech signals respectively and both of them are normalized to zero mean. Then, SI-SNR and WRQL can be formulated as:

$$s_{\text{target}} = \frac{\langle \hat{s}, s\rangle s}{\|s\|^2} \tag{1}$$

$$e_{\text{noise}} = \hat{s} - s_{\text{target}} \tag{2}$$

$$SI\text{-}SNR := 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2} \tag{3}$$

$$WRQL := -SI\text{-}SNR \tag{4}$$

3.2.2. Speaker Embedding Consistency

As per our problem formulation, the input reference speech and the estimated target speech, even though content wise different, are spoken by the same speaker. Therefore, both of them should have similar voice textures which means they should produce similar embedding vectors when passed through speaker encoder. With this intuition we formulate the speaker embedding consistency loss (SECL) as follows:

$$SECL := \|SE_\theta(r) - SE_\theta(\hat{s})\|^2 \tag{5}$$

where $r, \hat{s} \in \mathbb{R}^{1\times t}$ represent reference and estimated speeches respectively and $SE_\theta : \mathbb{R}^{1\times t} \to \mathbb{R}^d$ is the Speaker Encoder parameterized by θ, which produces fixed length embedding from variable length speech samples.

3.2.3. Inverse Consistency

Waveform encoder (WE) and waveform decoder (WD) should be inverse operation of each other with an intuitional analogy of STFT and inverse STFT. We introduce inverse consistency loss (ICL) to implicitly impose this constraint. Let $m \in \mathbb{R}^{N\times T}$ denotes the masked representation of the separated speech in spectrogram-like transformed space. Then ICL can be formulated as:

$$ICL := \|m - WE_\gamma(WD_\delta(m))\|^2 \tag{6}$$

where $WE_\gamma : \mathbb{R}^{1\times t} \to \mathbb{R}^{N\times T}$ and $WD_\delta : \mathbb{R}^{N\times T} \to \mathbb{R}^{1\times t}$ are waveform encoder parameterized by γ and waveform decoder parameterized by δ respectively.

3.2.4. Adversarial Refinement

Finally, we use a multi-scale discriminator (MSD) [16, 21] in an adversarial setting with a goal to make the estimated target speech indistinguishable from the ground truths. We train MSD to classify the ground truth samples to class 1 and estimated samples to class 0, where as the generator i.e the Speech Separator (See 3.1.2) is trained to fool the discriminator. The adversarial losses can be formulated as:

$$\mathcal{L}_{\text{d}}(D;G) := \mathbb{E}_{(s,i,e)}\left[(D(s)-1)^2 + (D(G(i,e)))^2\right] \tag{7}$$

$$\mathcal{L}_{\text{g}}(G;D) := \mathbb{E}_{(i,e)}\left[(D(G(i,e))-1)^2\right] \tag{8}$$

where $G : \mathbb{R}^{1\times t} \times \mathbb{R}^d \to \mathbb{R}^{1\times t}$ denotes generator i.e speech separator, $D : \mathbb{R}^{1\times t} \to \{0,1\}$ denotes discriminator and s, i, e carry previously defined meanings. $\mathcal{L}_{\text{g}}(G;D)$ is added with the other three losses and $\mathcal{L}_{\text{d}}(D;G)$ is optimized with a second optimizer (AdamW) with same learning rate.

Model Variant	SDRi (dB)	SI-SNRi (dB)
Baseline	11.13	10.42
Baseline+ICL	10.92	10.07
Baseline+ICL+SECL	10.95	10.15
Baseline+ICL+SECL+JointTraining	12.41	11.72
DPTNet+ICL+SECL+JointTraining	13.94	13.23
Spectron (with "DPTNet" and "MSD")	**14.25**	**13.44**

Table 1. Ablation study of Spectron on 2 speaker mixtures.

Model	SDRi (dB)	SI-SNRi (dB)
VoiceFilter [14]	7.8	-
AtssNet [13]	9.3	-
X-Tasnet [12]	13.8	12.7
Spectron without MSD (ours)	13.9	12.8
Spectron (ours)	**14.4**	**13.3**

Table 2. Spectron performance vs recent state-of-the-arts.

4. RESULTS

4.1. Dataset

We base our experiments on different mixture subsets generated from LibriSpeech[2] data [22]. In particular, for ablation we use LibriMix [23] script[3] on *"train-clean-100"*, *"dev-clean"* and *"test-clean"* for creating training, validation and test mixtures respectively. We collectively refer this data as "LibriMix Data". On the other hand, for SOTA comparison we use the same mixture subsets released by google[4] and already used in [14, 13, 12]. We refer this dataset as "VoiceFilter Data".

4.2. Experimental Setup

For all the experiments we have used $batchsize = 4$, $learningrate = 1e^{-4}$, $weightdecay = 1e^{-7}$ and train with Adam (and AdamW for discriminator) optimizer(s) for 201 epochs and use the best validation weights to compute the performance measures on the test set. Our speech separator module operates at 8 KHz sampling frequency, where as speaker encoder uses 16 KHz. Therefore, re-sampling is taken care on the fly with torchaudio defaults.

It is also noteworthy, that for "LibriMix Data" we split the clean speeches in non-overlapping 3 second and 2 second segments to use them as ground truth and reference speech respectively as we don't have separate reference speech. However, for "VoiceFilter Data", as we have mix, ground truth and reference speeches we use 3 second segment for each of them.

4.3. Ablation Study

We show ablation study of Spectron in Table 1. We start with a naive CNN baseline with fixed pre-trained Speaker Encoder[5] and "ConvTasnet" Separator module trained with only Waveform Reconstruction Quality Loss (WRQL). Then, we introduce Inverse Consistency Loss (ICL) followed by Speaker Embedding Consistency Loss (SECL). Training the speaker encoder jointly instead of keeping it fixed improves the baseline performance by roughly 1.3 dB points. This is probably because joint training allows to capture better relationship between the speaker embedding computation and its use in conditional speech separation. Finally, introduction of "DPTNet" separator backbone and multi-scale discriminator (MSD) loss, along with previous losses and joint training, improves the baseline by 3.12 dB points.

4.4. Comparison with Existing Works

We compare quantitative performance of Spectron with existing state-of-the-art target speaker extraction frameworks in Table 2. To make a fair comparison across models, here we use same train and same test data sets as used in [14, 13, 12]. As we see, Spectron undoubtedly performs better than "VoiceFilter" [14], "AtssNet" [13] and naive "X-Tasnet" [12]. However, a variant of "X-Tasnet" which uses *Loss on Distortion* (LoD), performs slightly (+0.3 dB points) better than Spectron, but at the cost of significant amount of additional data and more complicated training procedure as LoD needs ground truth speeches of all the speakers in the mixture.

5. CONCLUSION

We have presented Spectron, which uses dual path transformer conditioned on speaker embedding produced by a speaker encoder to extract the speaker of interest. Attention mechanism in the transformer, introduction of two additional objective functions followed by adversarial refinement and joint training of speaker encoder and speech separator - these four ideas all together improves the baseline performance and thus push the current SOTA further in target speaker extraction. Spectron can be directly used in different down stream speech based applications like automatic speech recognition, conditional speaker diarization, voice command activated personal assistants and so on. Additionally, it enables interactive audio manipulation, where the clean portions of an audio can be used as reference to de-noise the cluttered portions of the same audio. In future, Spectron can be explored as a general framework for any kind of target audio extraction with suitable reference audio encoder.

[2] http://www.openslr.org/12/

[3] https://github.com/JorisCos/LibriMix

[4] https://tinyurl.com/2vc795dw

[5] pre-trained encoder from: https://tinyurl.com/4r4a63np

6. ACKNOWLEDGEMENT

This project was done at the Visual Computing and Artificial Intelligence lab of Technical University of Munich under the kind supervision of Prof. Dr. Matthias Niessner. Author cordially thanks all lab members for useful discussion and timely co-operation.

7. REFERENCES

[1] Adel Belouchrani and Moeness G Amin, "Blind source separation based on time-frequency signal representations," *IEEE Transactions on Signal Processing*, vol. 46, no. 11, pp. 2888–2897, 1998.

[2] Ganesh R Naik, Wenwu Wang, et al., "Blind source separation," *Berlin: Springer*, vol. 10, pp. 978–3, 2014.

[3] Shoji Makino, Te-Won Lee, and Hiroshi Sawada, *Blind speech separation*, vol. 615, Springer, 2007.

[4] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[5] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[6] Yi Luo, Zhuo Chen, and Takuya Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[7] Jingjing Chen, Qirong Mao, and Dong Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.

[8] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.

[9] Yi Luo and Nima Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[10] Max WY Lam, Jun Wang, Dan Su, and Dong Yu, "Sandglasset: A light multi-granularity self-attentive network for time-domain speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5759–5763.

[11] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[12] Zining Zhang, Bingsheng He, and Zhenjie Zhang, "X-tasnet: Robust and accurate time-domain speaker extraction network," *arXiv preprint arXiv:2010.12766*, 2020.

[13] Tingle Li, Qingjian Lin, Yuanyuan Bao, and Ming Li, "Atss-net: Target speaker separation via attention-based neural network," *arXiv preprint arXiv:2005.09200*, 2020.

[14] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John Hershey, Rif A Saurous, Ron J Weiss, Ye Jia, and Ignacio Lopez Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," *arXiv preprint arXiv:1810.04826*, 2018.

[15] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "Sdr–half-baked or well done?," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[16] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.

[17] Qinglong Li, Fei Gao, Haixin Guan, and Kaichi Ma, "Real-time monaural speech enhancement with short-time discrete cosine transform," *arXiv preprint arXiv:2102.04629*, 2021.

[18] Efthymios Tzinis, Yossi Adi, Vamsi Krishna Ithapu, Buye Xu, Paris Smaragdis, and Anurag Kumar, "Remixit: Continual self-training of speech enhancement models via bootstrapped remixing," *arXiv preprint arXiv:2202.08862*, 2022.

[19] Umut Isik, Ritwik Giri, Neerad Phansalkar, Jean-Marc Valin, Karim Helwani, and Arvindh Krishnaswamy, "Poconet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss," *arXiv preprint arXiv:2008.04470*, 2020.

[20] Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li, "Fullsubnet: a full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6633–6637.

[21] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[22] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[23] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.