INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Universal Speaker Extraction in the Presence and Absence of Target Speakers for Speech of One and Two Talkers

Marvin Borsdorf [1], *Chenglin Xu* [2], *Haizhou Li* [2,1], *Tanja Schultz* [3]

[1]Machine Listening Lab (MLL), University of Bremen, Germany
[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[3]Cognitive Systems Lab (CSL), University of Bremen, Germany

marvin.borsdorf@uni-bremen.de

Abstract

Speaker extraction has been studied mostly for the scenarios where a target speaker is present in a two or more talkers mixture. Such scenarios do not adequately reflect everyday conversations. For example, a target speaker can be the only active talker, be quiet for a while, or leave the conversation, that means the target speaker is absent from the mixture. Traditional speaker extraction models fail in these scenarios. We propose a novel speaker extraction approach to handle speech mixtures with one or two talkers in which the target speaker can either be present or absent. First, we formulate four speaker extraction conditions to cover the typical scenarios of everyday conversations with one and two talkers. Second, we introduce a joint training scheme with one unified loss function that works for all four conditions. We show that only a small amount of data is required to adapt the model to work well in the four conditions.

Index Terms: Speaker extraction, selective auditory attention, cocktail party problem, everyday conversations, time domain

1. Introduction

Humans have the ability to select and follow a specific target speaker's speech in a multi-talker conversational scenario, that is selective auditory attention. Ever since the "cocktail party problem" was first formulated [1], the quest for equipping machines with such a human ability has never stopped. With the advent of deep learning, speech separation and speaker extraction represent the ways to solve the cocktail party problem through engineering approaches. These techniques have facilitated many speech processing tasks in multi-talker acoustic environments, including speech enhancement, automatic speech recognition (ASR), emotion recognition, speaker diarization, speaker verification, and speaker identification, to name a few.

The state-of-the-art speech separation techniques include deep clustering [2], PIT [3], DANet [4], TasNet [5], Conv-TasNet [6], DPRNN [7], Two Step Learning [8], Wavesplit [9], SepFormer [10], and SDNet [11]. Traditional speech separation techniques require prior knowledge about the number of speakers. In real-world conditions such as spontaneous meetings, the number of speakers is not known in advance and may vary from time to time. Some methods attempt to estimate the maximum number of speakers [12, 13, 14], while others apply recursive or iterative methods [15, 16, 17, 18] for multi-talker mixtures with an unknown number of speakers.

In practice, machines are built to focus on one of the speakers in a mixture just like humans do. The speaker extraction approach seeks to do just that. It uses a speaker's reference signal to extract the target speaker's voice in a multi-talker speech mixture without any prior knowledge about the number of speakers. The early success of neural speaker extraction includes beamforming approaches [19, 20], SpeakerBeam [21, 22, 23], SBF-MTSAL/SBF-MTSAL-Concat [24, 25], VoiceFilter [26], VoiceFilter-Light [27], and Atss-Net [28]. Recently, speaker extraction has benefited from the introduction of TasNet [5], and Conv-TasNet [6]. A cluster of time domain speaker extraction techniques has improved the performance, that include TseNet [29], SpEx [30], SpEx+ [31], SpEx++ [32], TD-SpeakerBeam [33], SCCM [34], IRA [35], a universal sound selector system [36], and a speaker conditioning mechanism [37].

All of these speaker extraction approaches assumed multi-talker scenarios with at least two talkers and always present target speaker. Apparently, these settings do not reflect real-world everyday conversational situations in which a target speaker can stop talking for a while or leave the conversation, i. e. the target speaker is absent from the mixture, or hold a monologue, i. e. resulting in a single-talker scenario. Recently, Wisdom et al. [38] proposed a Free Universal Sound Separation (FUSS) dataset and investigated a universal sound separation method which estimates silence for inactive sources. The idea is to combine different training losses for active and inactive sources, but to use a different metric for performance evaluation. We found a single prior study, X-TaSNet by Zhang et al. [39], that attempts a speaker extraction technique under multi-talker conditions with absent target speakers. However, there is no reported work on a single-talker scenario for the speaker extraction task yet.

We propose a novel universal speaker extraction approach which can handle multiple everyday conversational situations. For the first time, four speaker extraction conditions are formulated that cover all described situations, where in the presence of the target speaker, the model extracts the target speaker's voice, and in the absence of the target speaker, the model is expected to output silence. We introduce a joint training scheme with one unified loss function for all four conditions. The metric is similar to the evaluation metric in Wisdom et al. [38] but we utilize the same metric for both training and evaluation. Our approach considers single- and two-talker mixtures but it can easily be extended to any number of speakers in the mixture. All scripts and information will be made publicly available[1].

The paper is organized as follows. In Section 2, we reformulate the four speaker extraction conditions. In Section 3, we introduce the joint training scheme. In Section 4, we describe and perform the experiments, followed by the discussion in Section 5. Finally, Section 6 concludes the study.

2. Universal Speaker Extraction Conditions

For real-world applications, a speaker extraction system needs to produce reliable output when the target speaker is either

[1] https://github.com/mborsdorf/UniversalSpeakerExtraction

http://dx.doi.org/10.21437/Interspeech.2021-1939

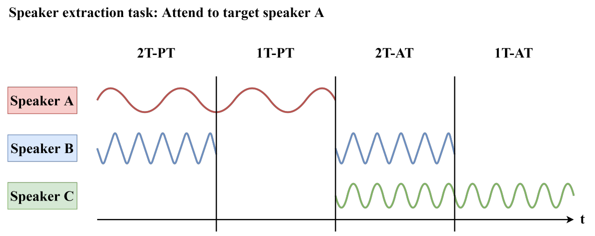

Figure 1: *Visualization of the four conditions in an example meeting. The speaker extraction task is to attend to target speaker A while speakers B and C represent interfering speakers. The meeting comprises two-talker (2T) or single-talker (1T) with present target (PT) or absent target (AT) conditions.*

present or absent in the speech, regardless of whether one talker or multiple talkers are actively speaking. Therefore, we consider that a universal speaker extraction system should perform under four conditions to cover all acoustic scenarios in everyday conversational situations. Without loss of generality, a two-talker scenario is studied, with straight-forward extension to multi-talker scenarios. We reformulate the speaker extraction task under four conditions as follows:

- **2T-PT:** The audio mixture comprises speech of two talkers with one being the target speaker (present target).
- **1T-PT:** The audio mixture comprises speech of one talker who is the target speaker (present target).
- **2T-AT:** The audio mixture comprises speech of two talkers with none being the target speaker (absent target).
- **1T-AT:** The audio mixture comprises speech of one talker who is not the target speaker (absent target).

Figure 1 shows the four different conditions modeled in an example conversation of three talkers. Speaker extraction models have been mostly trained on two- or multi-talker scenarios in the presence of the target speakers, denoted as 2T-PT. Those models usually fail when being applied to one of the other three scenarios, where they show poor performance, extract artifacts, or recover a random speaker's voice. This failure is not desired. Rather the model shall either extract the target speaker's voice (present target speaker) or silence (absent target speaker).

3. Joint Training Scheme

The formulation of universal speaker extraction covers the scenarios of single-talker and absent target speaker. We propose a joint training scheme with a unified loss function as we are building one system for all four conditions.

3.1. Data adaptation

To train a speaker extraction model in a supervised fashion usually a triple of audio files is required. The first component is given by the mixed speech audio signal which contains different speakers including the target speaker. The second part is a reference signal of the target speaker. The clean audio data of the target speaker (ground truth), which is also spoken in the mixed speech audio, represents the last component of the triple. The speaker extraction model uses the reference signal to find the target's speech in the input mixture and extracts the speech. The recovered speech signal is compared to the ground truth speech signal. The difference denotes the reconstruction error and has to be minimized. This procedure is also applicable to each of the four conditions. For 2T-PT and 1T-PT the model extracts the target's speech which is given as ground truth. For 2T-AT and 1T-AT the model extracts silence and therefore requires silent audio files (i. e. the audio file contains zeros) as ground truths with the duration of the mixture signal.

3.2. Metric stabilization

Since the supervised training strategy of the conditions includes silent ground truth data which solely consists of zeros, we have to ensure that the used traditional metrics for both the training loss calculation and the evaluation do not break. This counts for any kind of metric but in this work we concentrate on the scale-invariant signal-to-distortion ratio (SI-SDR) proposed by Le Roux et al. [40]. The SI-SDR is defined as

$$\text{SI-SDR} = 20\log_{10}\left(\frac{\left\|\frac{\hat{s}^T s}{\|s\|^2}s\right\|}{\left\|\frac{\hat{s}^T s}{\|s\|^2}s-\hat{s}\right\|}\right) \tag{1}$$

where s represents the ground truth and $\hat{s}$ the reconstructed target signal. Commonly a little stabilization value ε is introduced to prevent the equation from breaking. This leads to

$$\text{SI-SDR} \approx 20\log_{10}\left(\frac{\left\|\frac{\hat{s}^T s}{\|s\|^2+\varepsilon}s\right\|}{\left\|\frac{\hat{s}^T s}{\|s\|^2+\varepsilon}s-\hat{s}\right\|+\varepsilon}+\varepsilon\right) \tag{2}$$

with $\varepsilon = 1e^{-8}$ in our experiments. With $s = 0$ for absent target speaker conditions, the SI-SDR based on Eq. 2 would lead to a constant value regardless of the quality of the reconstructed signal $\hat{s}$. Hence, we move the second term of the logarithm to the numerator of the first term, similar to the evaluation metric in Wisdom et al. [38]. This little modification still preserves the stability for $s = 0$ but reflects the reconstructed signal $\hat{s}$ in the calculation. That allows us to evaluate the quality of the reconstructed silence. In the following we will refer this as silence-evaluating SI-SDR (SE-SI-SDR), defined as:

$$\text{SE-SI-SDR} = 20\log_{10}\left(\frac{\left\|\frac{\hat{s}^T s}{\|s\|^2+\varepsilon}s\right\|+\varepsilon}{\left\|\frac{\hat{s}^T s}{\|s\|^2+\varepsilon}s-\hat{s}\right\|+\varepsilon}\right) \tag{3}$$

The term SI-SDR will refer to Eq. 2. The differences between SI-SDR and SE-SI-SDR are negligible for our experiments. This is also confirmed by a practical test of both equations on 2T-PT and 1T-PT which shows differences in the order of $1e^{-5}$. The higher the SE-SI-SDR the better the achieved performance of the system. This is valid for both present and absent target speaker conditions. If the reconstructed signal $\hat{s}$ solely contains zeros, the SE-SI-SDR for the absent target speaker condition will be zero, which represents the theoretically ideal value. Otherwise, the SE-SI-SDR is always lower than zero for the absent target speaker condition. To optimize the speaker extraction network, the negative SE-SI-SDR is employed as the loss, which is the negative value of Eq. 3.

4. Experimental Setup

We conducted experiments on different training schemes and database compositions, while maintaining the same network architecture for fair comparisons. The experiments were implemented in Python using PyTorch.

4.1. Network architecture

We applied the SpEx+ architecture[2] which was recently proposed by Ge et al. [31] and showed state-of-the-art performance

[2] Original implementation: https://github.com/gemengtju/SpEx_Plus

on the speaker extraction task. SpEx+ works in time domain and thus can directly process the input signals as waveforms. Two speech encoders share their weights and project the mixture signal and the reference signal respectively to the same latent space. The main part of the model fulfills two tasks. First, it performs a speaker classification task based on the reference signal to provide a target speaker embedding for the extraction part. Second, it extracts the target speaker's voice out of the input mixture. The decoder transforms the latent projection of the separated signal back to its waveform and thus allows to incorporate a training objective which directly measures the reconstructed speech quality in time domain. We kept the same SpEx+ baseline architecture as in Ge et al. [31].

4.2. Data

We investigated our proposed joint training scheme with initial experiments on the WSJ0-2mix-extr[3] database which was recently introduced by Xu et al. [24] and Rao et al. [25] and used e. g. in Xu et al. [30], Xu et al. [29], Ge et al. [31] and Deng et al. [35]. This database is an adaptation of the popular WSJ0-2mix database which has been widely used for blind source separation experiments and was introduced by Hershey et al. [2]. WSJ0-2mix-extr contains training, development and test sets with a sampling frequency of 8 kHz. We used the 8 kHz sampled version to follow the recent works and to reduce the computational costs. The WSJ0-2mix-extr database can also be created as 16 kHz sampled version.

The training and development sets share the same 101 speakers (closed condition), whereas the 18 speakers in the test set are different (open condition). In total 20000, 5000 and 3000 utterances are provided for the training, development and test sets respectively. Since each speaker in a two-talker mixture can be considered as a target, each mixture can be utilized twice. All sets comprise audio data triples as described in Section 3.1.

To match our experimental requirements, we modified the WSJ0-2mix-extr database according to the description in Section 3.1 and created one specific database for each of the four conditions. The following steps were applied on the training, development and test sets. For 2T-PT we adopted the original WSJ0-2mix-extr database since it completely matched this condition. 1T-PT was created by replacing the input mixtures with the ground truth data in the respective data triple. For half of this database the target's reference signal was given by a different utterance and for the rest the reference signal was the same spoken utterance as for the mixture and the ground truth signals. Those triples were randomly selected. We did this to include data which simulates a speaker verification task as the enrolment speech has to pass through both speaker extraction and speaker encoder parts.

To create the databases with absent target speakers, i. e. 2T-AT and 1T-AT, we first changed the reference signal of each triple to be different from the speakers in the two- or single-talker input mixture. Second, we replaced the ground truth data with silent audio files with the same length of the respective input mixture. We also created smaller versions of 2T-AT and 1T-AT by randomly sampling 50 % of the data. Furthermore, we suspected that having exactly the same input mixtures one time with present and one time with absent target speaker could confuse the model in the training phase. To investigate this, we randomly removed 15 % of the data of 2T-PT and 1T-PT and used those portions to create small 2T-AT and 1T-AT databases.

4.3. Training and evaluation

We have trained seven models on different training methods. Each training method was defined by a pre-trained model (except for the training of the baseline model) and a database composition. During training, we used the negative SE-SI-SDR as the training loss. In parallel, each model performed a speaker classification task to create a speaker embedding out of the target speaker's reference signal. For the training of model 1 and model 2 we applied the SI-SDR because only data with present target speaker was considered. For the remaining models we applied SE-SI-SDR. Both losses are directly comparable because the differences are small (see description in Section 3.2).

We utilized Adam as the optimizer to update the networks' weights during training. For the sequential training of the models 1, 2, 3 and 4, we set the initial learning rates to $1e^{-3}$, $1e^{-4}$, $5e^{-5}$, and $2.441e^{-8}$ respectively. Each initial learning rate was defined by the last learning rate of the previously trained model (except for model 1 which was the baseline).

Since we always kept the training data of the previous conditions when data of a new condition was added, we did not expect a strong degradation in performance for previous conditions. To investigate this assumption, we trained models 5 and 6 with a larger initial learning rate of $1e^{-4}$. To extend this even more, model 7 was trained with an initial learning rate of $1e^{-3}$. We used a learning rate scheduler to halve the learning rate after two subsequent epochs with no improvements. We trained all models for at most 150 epochs independently of the training method. This was controlled by an early stopping mechanism which stopped the training if no decrease in the joint loss was observed for 6 consecutive epochs. The data were fed to the network as chunks of four seconds. We initially trained the model 1 on 2T-PT and started from this baseline with different training methods to fine-tune/adapt the model to new conditions.

Each model was evaluated on all four different conditions. This was done by separately calculating the mean SE-SI-SDR in dB on each test set of 2T-PT, 1T-PT, 2T-AT and 1T-AT.

5. Results and Discussion

We evaluated the performances of all seven trained models, as introduced in Section 4.3, on four different test sets which simulate the presence and absence of the target speaker in speech mixtures of one and two talkers. We report the results as mean SE-SI-SDR (dB) for each test set, illustrated in Table 2.

With model 1 we present our baseline speaker extraction model trained on 2T-PT. On 1T-PT the model either extracts parts of the target speaker signal only or a noisy unintelligible signal. This results in a poor performance. Applied to the absent target speaker conditions, the model fails entirely. Due to the mismatch of the reference signal, the model randomly selects one out of two speakers on 2T-AT. The wrongly recovered speech shows a good intelligibility but occasionally contains artifacts from the other speaker or noise. On the 1T-AT test set, the model mostly extracts noisy unintelligible signals. On both test sets, the model show similar results in terms of SE-SI-SDR. This experimentally proofs that the SE-SI-SDR penalizes extracted speech or noise in the same manner when the model is actually meant to extract silence.

Fine-tuning model 1 by adding single-talker data with present target speakers shows a strong raise in performance on 1T-PT but at a cost of a slight performance degradation on 2T-PT. The results on the absent target speaker conditions are barely affected. We highlight that small changes in performance

[3] https://github.com/xuchenglin28/speaker_extraction

Table 1: *Speaker extraction results for different training methods and database compositions. The models are separately evaluated on the four test sets 2T-PT, 1T-PT, 2T-AT and 1T-AT. The results are reported as mean SE-SI-SDR (dB) over the entire respective test database. The first row specifies the mean SE-SI-SDR over the input mixtures before applying the speaker extraction. (n %) refers to the amount of adaptation data which is used in the respective training method.*

Model	Training method	Training databases				Results on test sets in SE-SI-SDR (dB)			
		2T-PT	1T-PT	2T-AT	1T-AT	2T-PT	1T-PT	2T-AT	1T-AT
-	Input mixture	-	-	-	-	0.0	181.8	-185.8	-182.5
Model 1	From scratch (baseline)	✓	✗	✗	✗	**16.9**	10.9	-184.2	-176.0
Model 2	Fine-tuning model 1	✓	✓	✗	✗	16.8	58.0	-184.2	-179.3
Model 3	Fine-tuning model 2	✓	✓	✓	✗	10.6	**82.7**	-6.6	-42.9
Model 4	Fine-tuning model 3	✓	✓	✓	✓	9.5	79.5	**-4.3**	-16.2
Model 5	Fine-tuning model 2	✓	✓	✓(50 %)	✗	12.4	79.5	-10.8	-54.6
Model 6	Fine-tuning model 1	✓	✓	✓(50 %)	✓(50 %)	12.8	61.2	-9.1	**-5.1**
Model 7	Exclusive fine-tuning model 2	✓(85 %)	✓(85 %)	✓(15 %)	✓(15 %)	14.7	78.0	-17.5	-11.2

on 2T-PT matter more than on the other conditions since it is important to preserve the performance on the initial condition.

Adding 2T-AT data to the training leverages a performance increase on both 2T-AT and 1T-AT as shown by model 3. This indicates that telling the model what to do if a target speaker is not present in a two-talker mixture, also helps the model to better handle single-talker mixtures with absent target speakers. However, we encounter a huge drop in performance on the 2T-PT test data. Our observations show that the model sometimes starts to extract silence even if the target speaker is present in the mixture. Surprisingly, adding 2T-AT data to the training helps to strongly improve the performance on 1T-PT even though the condition is completely different.

The training method for model 4 adds the fourth database to the training that includes single-talker mixtures with absent target speakers. This further improves the test performance on 2T-AT and on 1T-AT. However, the results on the present target speaker conditions are worse than for model 3. Compared to the baseline model we observe high performance improvements on the 1T-PT, 2T-AT and 1T-AT conditions whereas the performance on the initial condition 2T-PT drops of about 43.8 %. The steps from model 1 to 4 can be considered as an iterative fine-tuning of the initial model in which each iteration exposes the model to the already seen and one new unseen condition.

The results for model 5 can be compared to those of model 3. We see that a reduction of 50 % of 2T-AT training data is sufficient to allow an adaptation to both absent target speaker conditions. Even if the results for 1T-PT, 2T-AT and 1T-AT are slightly worse than the results of model 3, the performance on the source domain condition 2T-PT remains more stable.

With model 6 we introduce a fine-tuning from the baseline condition to three different new conditions in one step. In this method the model is facing both the single-talker mixtures and the absent target speaker conditions all together for the first time. Based on the results of model 5, that only a little amount of data is required to adapt to the absent target speaker conditions, 2T-AT and 1T-AT are reduced to 50 % of the amount of data. 1T-PT is used with full amount. This leads in total to 25 % less training data compared to the training of model 4. This training method outperforms model 4 on 2T-PT and 1T-AT. On the other two conditions it is vice versa. However, this also shows that a little amount of absent target speaker data is sufficient to adapt to the new conditions and also helps to reduce the performance degradation on the initial condition.

The results for model 7 show that just 15 % of the original data are sufficient to adapt to the absent target speaker conditions even though the performance is worse than for model 6. We also observe a more stable performance on the initial condition 2T-PT and a better performance on 1T-PT compared to model 6. We believe this is not only due to the further reduction of the absent target speakers data but also because of the disjoint input mixtures for present and absent target speaker conditions as described in Section 4.2. The latter may reduce potential confusion of the model. Additionally, this method reduces the total amount of utilized data to 50 % compared to model 4 and hence also the training time as well as storage consumption.

Across all models which have been trained with at least one absent target speaker dataset, model 7 shows the best performance stability on 2T-PT. This is our main goal when adapting to other conditions because of two reasons. First, speaker extraction has been widely studied on two-talker mixtures with present target speakers and we intend to preserve high performance on this condition. Second, if the model would forget previously learnt conditions, we would end up with a cluster of models, each specialized for a certain condition. This does not reflect our motivation of having one model that can work for all conditions. Model 7 also denotes a high performance on 1T-PT as well as decent performances on 2T-AT and 1T-AT.

6. Conclusion and Future Work

In this work, we proposed a universal speaker extraction approach which works in the presence and absence of target speakers for speech mixtures comprising one and two talkers. We introduced a joint training scheme with one unified loss for all speaker conditions. The training metric is also applicable for the model evaluation. The experiments show that the baseline model adapts well to the new conditions, in particular, to the conditions in which the target speaker is absent with only a small amount of adaptation data. However, we found that the fine-tuned model occasionally extracts silence although the target speaker is present in the mixture.

In our future work, we will analyze this behavior and possible solutions. We will study additional classification approaches to assist the model's decision about whether a target speaker is present or absent. Moreover, we will examine other more tangible metrics to incorporate the absent target speaker conditions. In conclusion, the experimental results of our study confirm that the proposed universal speaker extraction approach works.

7. Acknowledgements

This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen).

8. References

[1] E. C. Cherry, “Some Experiments on the Recognition of Speech, with One and with Two Ears,” *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep Clustering: Discriminative Embeddings for Segmentation and Separation,” in *ICASSP*, 2016.

[3] D. Yu, M. Kolbæk, Z. Tan, and J. Jensen, “Permutation Invariant Training of Deep Models for Speaker-Independent Multi-talker Speech Separation,” in *ICASSP*, 2017.

[4] Z. Chen, Y. Luo, and N. Mesgarani, “Deep Attractor Network for Single-microphone Speaker Separation,” in *ICASSP*, 2017.

[5] Y. Luo and N. Mesgarani, “TasNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation,” in *ICASSP*, 2018.

[6] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 1256–1266, 2019.

[7] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation,” in *ICASSP*, 2020.

[8] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan, and P. Smaragdis, “Two-Step Sound Source Separation: Training on Learned Latent Targets,” in *ICASSP*, 2020.

[9] N. Zeghidour and D. Grangier, “Wavesplit: End-to-End Speech Separation by Speaker Clustering,” *arXiv preprint arXiv:2002.08933v2*, 2020.

[10] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, “Attention is All You Need in Speech Separation,” in *ICASSP*, 2021.

[11] C. Li, J. Xu, N. Mesgarani, and B. Xu, “Speaker and Direction Inferred Dual-Channel Speech Separation,” in *ICASSP*, 2021.

[12] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Multi-talker Speech Separation with Utterance-level Permutation Invariant Training of Deep Recurrent Neural Networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[13] Y. Luo, Z. Chen, and N. Mesgarani, “Speaker-independent Speech Separation with Deep Attractor Network,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.

[14] Y. Luo and N. Mesgarani, “Separating Varying Numbers of Sources with Auxiliary Autoencoding Loss,” in *INTERSPEECH*, 2020.

[15] K. Kinoshita, L. Drude, M. Delcroix, and T. Nakatani, “Listening to Each Speaker One by One with Recurrent Selective Hearing Networks,” in *ICASSP*, 2018.

[16] J. Shi, J. Xu, G. Liu, and B. Xu, “Listen, Think and Listen Again: Capturing Top-down Auditory Attention for Speaker-independent Speech Separation,” in *IJCAI*, 2018.

[17] N. Takahashi, S. Parthasaarathy, N. Goswami, and Y. Mitsufuji, “Recursive Speech Separation for Unknown Number of Speakers,” in *INTERSPEECH*, 2019.

[18] T. von Neumann, C. Boeddeker, L. Drude, K. Kinoshita, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, “Multi-talker ASR for an Unknown Number of Sources: Joint Training of Source Counting, Separation and ASR,” in *INTERSPEECH*, 2020.

[19] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, “Speaker-aware Neural Network Based Beamformer for Speaker Extraction in Speech Mixtures,” in *INTERSPEECH*, 2017.

[20] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, “Learning Speaker Representation for Neural Network Based Multichannel Speaker Extraction,” in *ASRU*, 2017.

[21] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, “Single Channel Target Speaker Extraction and Recognition with Speaker Beam,” in *ICASSP*, 2018.

[22] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, S. Araki, and T. Nakatani, “Compact Network for SpeakerBeam Target Speaker Extraction,” in *ICASSP*, 2019.

[23] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černocký, “SpeakerBeam: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[24] C. Xu, W. Rao, E. S. Chng, and H. Li, “Optimization of Speaker Extraction Neural Network with Magnitude and Temporal Spectrum Approximation Loss,” in *ICASSP*, 2019.

[25] W. Rao, C. Xu, E. S. Chng, and H. Li, “Target Speaker Extraction for Multi-Talker Speaker Verification,” in *INTERSPEECH*, 2019.

[26] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, “VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking,” in *INTERSPEECH*, 2019.

[27] Q. Wang, I. L. Moreno, M. Saglam, K. Wilson, A. Chiao, R. Liu, Y. He, W. Li, J. Pelecanos, M. Nika, and A. Gruenstein, “VoiceFilter-Lite: Streaming Targeted Voice Separation for On-Device Speech Recognition,” in *INTERSPEECH*, 2020.

[28] T. Li, Q. Lin, Y. Bao, and M. Li, “Atss-Net: Target Speaker Separation via Attention-based Neural Network,” in *INTERSPEECH*, 2020.

[29] C. Xu, W. Rao, E. S. Chng, and H. Li, “Time-Domain Speaker Extraction Network,” in *ASRU*, 2019.

[30] C. Xu, W. Rao, E. S. Chng, and H. Li, “SpEx: Multi-Scale Time Domain Speaker Extraction Network,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1370–1384, 2020.

[31] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, “SpEx+: A Complete Time Domain Speaker Extraction Network,” in *INTERSPEECH*, 2020.

[32] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, “Multi-Stage Speaker Extraction with Utterance and Frame-Level Reference Signals,” in *ICASSP*, 2021.

[33] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, “Improving Speaker Discrimination of Target Speech Extraction With Time-Domain SpeakerBeam,” in *ICASSP*, 2020.

[34] J. Shi, J. Xu, Y. Fujita, S. Watanabe, and B. Xu, “Speaker-Conditional Chain Model for Speech Separation and Extraction,” in *INTERSPEECH*, 2020.

[35] C. Deng, S. Ma, Y. Zhang, Y. Sha, H. Zhang, H. Song, and X. Li, “Robust Speaker Extraction Network Based on Iterative Refined Adaptation,” *arXiv preprint arXiv:2011.02102v1*, 2020.

[36] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, “Listen to What You Want: Neural Network-based Universal Sound Selector,” in *INTERSPEECH*, 2020.

[37] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, “Time-Domain Speech Extraction with Spatial Information and Multi Speaker Conditioning Mechanism,” in *ICASSP*, 2021.

[38] S. Wisdom, H. Erdogan, D. P. W. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, “What’s All the FUSS About Free Universal Sound Separation Data?” in *ICASSP*, 2021.

[39] Z. Zhang, B. He, and Z. Zhang, “X-TaSNet: Robust and Accurate Time-Domain Speaker Extraction Network,” in *INTERSPEECH*, 2020.

[40] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR – Half-Baked or Well Done?” in *ICASSP*, 2019.