Directed Speech Separation for Automatic Speech Recognition of Long-form Conversational Speech

Rohit Paturi, Sundararajan Srinivasan, Katrin Kirchhoff, Daniel Garcia Romero

Amazon AWS AI

Abstract

Many of the recent advances in speech separation are primarily aimed at synthetic mixtures of short audio utterances with high degrees of overlap. Most of these approaches need an additional stitching step to stitch the separated speech chunks for long form audio. Since most of the approaches involve Permutation Invariant training (PIT), the order of separated speech chunks is nondeterministic and leads to difficulty in accurately stitching homogenous speaker chunks for downstream tasks like Automatic Speech Recognition (ASR). Also, most of these models are trained with synthetic mixtures and do not generalize to real conversational data. In this paper, we propose a speaker conditioned separator trained on speaker embeddings extracted directly from the mixed signal using an over-clustering based approach. This model naturally regulates the order of the separated chunks without the need for an additional stitching step. We also introduce a data sampling strategy with real and synthetic mixtures which generalizes well to real conversation speech. With this model and data sampling technique, we show significant improvements in speaker-attributed word error rate (SA-WER) on Hub5 data.
Index Terms: Speech Separation, Speaker embeddings, Spectral clustering, ASR, deep learning

1. Introduction

Despite the recent advances in Automatic speech recognition (ASR), multi-speaker scenarios still pose a significant challenge to ASR systems [1–3] because of the difficulty of attending to the target speech signal from other interfering speech signals. One approach to recognize multi-speaker overlapped speech is by end-to-end speaker-attributed automatic speech recognition (SA-ASR) systems [4–8] which jointly model speaker identification and speech recognition for monaural overlapped speech. Though these systems have shown promise in recognizing multi-talker speech, when more downstream tasks (like emotion recognition, speech diarization, etc.) from overlapped speech conversations are needed, every task needs to be re-trained in this framework, reducing modularity. Also, these are shown to not generalize well to long form audio [8].

The other approach is to perform robust speech separation, which can then be a common frontend for all tasks and this is the approach we consider in this paper. Monaural speech separation has recently witnessed a rapid progress with the advent of supervised neural networks [9–13] in the time-frequency domain and end-to-end time domain approaches [14–17]. One set of approaches leverage speaker information to improve the separation performance. These use either pre-enrolled speaker utterances to perform target speaker separation [18–20] or extract speaker from preliminary separation [21–24] to improve the speaker agnostic separation performance.

Most of these approaches above are trained with a PIT loss [11] leading to a nondeterministic ordering of separated channels. In order to apply these to long audio recordings, an explicit stitching step is needed to stitch the separated chunks to form the long homogenous speaker channels. A common stitching mechanism compares similarity between overlapping regions of adjacent chunks [25, 26] to determine the correct chunk permutation to be stitched. But, this can be error-prone as one wrongly stitched chunk can lead to error propagation throughout the subsequent chunks and is sensitive to the separation quality of every chunk. Also, studies [27–29] report that, even though these separation models have consistently advanced the state of the art on some of the popular synthetic datasets in the field like wsj0-mix [10] and LibriMix [27], the ability to generalize to speech coming from real conversation settings in terms of ASR performance has not been achieved.

In this paper, we propose a speaker conditioned 2-speaker speech separation model for conversational telephone speech (CTS) without the need for pre-enrolled utterances and doesn't require PIT loss as the separated channels are directed by the order of the speakers fed into to the model. An over-clustering based approach is used to find speaker embeddings robust to speech overlaps which serve as inputs for the Directed Speech Separation (DSS) module. To train our system, we propose a data sampling strategy leveraging both synthetic read speech and in-domain real conversational datasets. We demonstrate that applying this DSS as a frontend to ASR on long-form audio is superior to stitching separation outputs with a PIT loss trained model, with an SI-SDR improvement of 10dB on CALLHOME American English [30] and SA-WER [4] improvement of 30% on Hub5 dataset [31].

2. Related Work

Two related approaches that don't rely on pre-enrolled speaker utterances are Wavesplit [21] and Continuous speech separation using speaker inventory (CSSUSI) for long recording [25].

Wavesplit performs preliminary separation and speaker embedding extraction followed by clustering to extract speaker centroids. These are then used to condition the separation stack. Wavesplit has been explored for only short fully overlapping utterances and is complex to train due to multiple stages of separation and speaker stack involved. Our approach differs from Wavesplit as it doesn't need any preliminary separation and can make use of a strong pre-trained speaker embedding network reducing the complexity of the system substantially.

CSSUSI directly extracts embeddings from mixed speech and forms a speaker inventory to condition the separation network. The separation network operates in time-frequency domain and is trained with PIT loss and hence, the order of the separated chunks is nondeterministic. A stitching mechanism using overlap similarity with adjacent chunks is used to stitch back the separated chunks. Our approach differs from CSSUSI mainly by conditioning a more robust end-to-end time domain separator network without the need for an additional stitching mechanism.

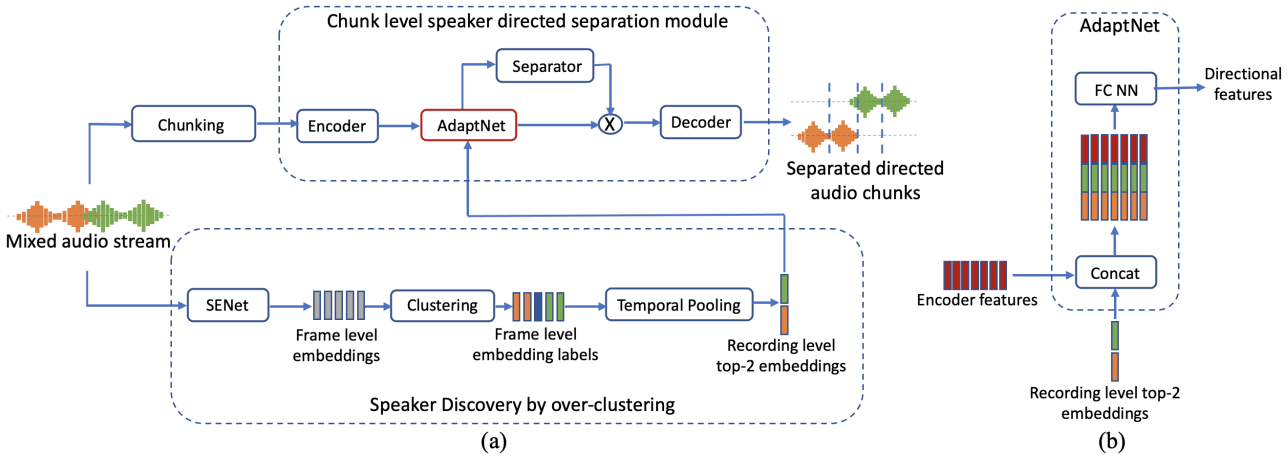

Figure 1: *(a) Overall architecture of the Directed Speech Separation for 2 speaker use-case, (b) Structure of the AdaptNet block.*

3. Directed Speech Separation

In this section, we introduce the components of the directed speech separation (DSS) system. It mainly comprises of two modules: a speaker discovery module robust to speaker overlaps and a speaker conditioned directed separation module.

3.1. Robust Speaker Discovery by Over-Clustering

The speaker discovery module is used to discover and extract embeddings of the constituent speakers of an audio mixture by taking advantage of the large number of non-overlapping speech regions in multi-talker conversational scenarios. A pretrained speaker embedding network (SENet) extracts frame level speaker embeddings $\{\boldsymbol{f_i}\}_{i=1}^{F}, \boldsymbol{f_i} \in \mathbb{R}^{1\times K}$, where F is the number of frames in the recording. These embeddings are clustered using spectral clustering with maximum eigen gap [32] for detecting the number of clusters C with additional constraints such that $N \leq C \leq M$ where N is the number of speakers to be separated and M is the maximum detected clusters, where $M > N$. Thus, we over-cluster the embeddings by setting these constraints and show in §4.4 that separation performance is insensitive to the value of M as long as $M > N$ and that over clustering the embeddings produces cleaner speaker clusters by attributing overlapped or noisy/background speech to additional clusters. Once the frame level embeddings $\{\boldsymbol{f_i}\}_{i=1}^{F}$ are partitioned into clusters $I_1, ..., I_C$ of sorted cardinalities $n_1, ..., n_C$, such that $n_1 > ... > n_C$ and $n_1 + ... + n_C = F$, the top-N utterance level embeddings $\{\boldsymbol{z_j}\}_{j=1}^{N}, \boldsymbol{z_j} \in \mathbb{R}^{1\times K}$ are computed by temporally pooling the frame level embeddings of the corresponding N clusters. In this paper, we use a simple mean pooling to produce recording level embeddings

$$\boldsymbol{z_j} = \frac{1}{n_j} \sum_{\boldsymbol{f_i} \in I_j} \boldsymbol{f_i}$$

These embeddings are used by the separation network to condition and direct the network as outlined in §3.2. This speaker discovery module is pretrained and is frozen during training and inference.

3.2. Speaker Directed Separation Network

The speaker directed separation module separates audio at the chunk level without the need for restitching the separated chunks in order to separate a long utterance. The encoder, separator and decoder architectures are based on ConvTasNet [14] architecture in this paper but can be based on any of the more recent transformer architectures [16, 17]. The inputs to this module are the audio chunk x, the recording level embeddings $\{\boldsymbol{z_j}\}_{j=1}^{N}$ and the outputs are the separated chunk level waveforms $\{\hat{\boldsymbol{s_j}}\}_{j=1}^{N}$ .The audio chunk x is divided into overlapping segments of length L, represented by $\{\boldsymbol{x_k}\}_{i=1}^{T}, \boldsymbol{x_k} \in \mathbb{R}^{1\times L}$, where T denotes the total number of encoder frames in the input chunk. $\boldsymbol{x_k}$ is transformed into a E dimensional encoder representation, $\{\boldsymbol{e_k}\}_{k=1}^{T}, e_k \in \mathbb{R}^{1\times E}$ by a 1-D convolution operation:

$$\boldsymbol{e_k} = \mathcal{H}(\boldsymbol{x_k U})$$

where $\boldsymbol{U} \in \mathbb{R}^{L\times E}$ contains E encoder basis functions with length L each and $\mathcal{H}(\cdot)$ is the ReLU non-linear function. We introduce an adaptation network (AdaptNet) shown in Fig. 1(b), which concatenates the N recording level speaker embeddings $\{\boldsymbol{z_j}\}_{j=1}^{N}$ to each frame of the encoder features $\boldsymbol{e_k}$ to form the intermediate directional features $\{\boldsymbol{a_k}\}_{k=1}^{T}, \boldsymbol{a_k} \in \mathbb{R}^{1\times A}$, such that

$$\boldsymbol{a_k} = concat(\boldsymbol{e_k}, concat(\boldsymbol{z_1}, ..., \boldsymbol{z_N})),$$

where $A = E + N \times K$. The intermediate directional features are transformed by a fully connected neural network to form the D-dimensional directional features $\{\boldsymbol{d_k}\}_{k=1}^{T}, \boldsymbol{d_k} \in \mathbb{R}^{1\times D}$

$$\boldsymbol{d_k} = \mathcal{H}(\boldsymbol{a_k W})$$

where $\boldsymbol{W} \in \mathbb{R}^{A\times D}$ is the AdaptNet weight matrix, and $\mathcal{H}(\cdot)$ is the ReLU non-linear function. The separator consists of stacked dilated temporal convolutional networks [14] and predicts a representation for each of the N sources by learning N masks $\{\boldsymbol{m_j}\}_{j=1}^{N}, \boldsymbol{m_j} \in \mathbb{R}^{1\times D}$ such that $\boldsymbol{m_j} \in [0, 1]$. The representation of each separated source $\{\boldsymbol{t_{k,j}}\}_{j=1}^{N}, \boldsymbol{t_{k,j}} \in \mathbb{R}^{1\times D}$ is calculated by applying the corresponding mask $\boldsymbol{m_j}$ to the directional features $\boldsymbol{d}$:

$$\boldsymbol{t_{k,j}} = \boldsymbol{d_k} \odot \boldsymbol{m_j}$$

where $\odot$ denotes element-wise multiplication. The waveform of each separated overlapping segment $\hat{\boldsymbol{s}}_{\boldsymbol{k,j}} \in \mathbb{R}^{1\times L}$ is reconstructed by the decoder:

$$\hat{\boldsymbol{s}}_{\boldsymbol{k,j}} = \boldsymbol{t_{k,j} V}$$

where $\boldsymbol{V} \in \mathbb{R}^{L\times E}$ contains E decoder basis functions. The overlapping reconstructed segments are summed together to generate the separated chunk level sources $\hat{\boldsymbol{s}}_{\boldsymbol{j}}$. Since we condition the separator using speaker embeddings, we train the network to minimize the negative scale invariant signal to distortion ratio (SI-SDR) [33] between separated sources $\{\hat{\boldsymbol{s_j}}\}_{j=1}^{N}$ ordered consistently with the input speaker embeddings $\{\boldsymbol{z_j}\}_{j=1}^{N}$ and the

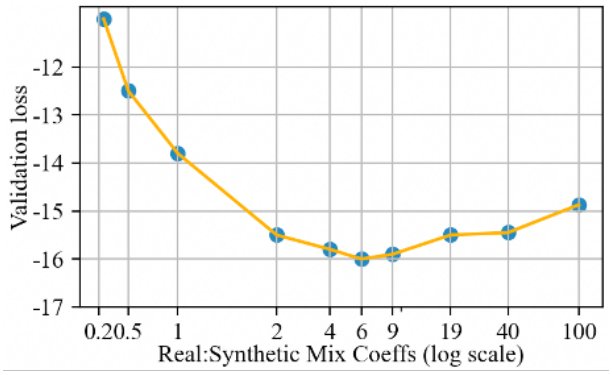

Figure 2: *Sampling coefficient parameter search on dev set*

ground truth sources $\{s_j\}_{j=1}^{N}$ and thus, avoid the permutation problem. As we need to align $\{\hat{z}_j\}_{j=1}^{N}$ to the corresponding $\{s_j\}_{j=1}^{N}$ for training the separator network, we leverage Hungarian algorithm of md-eval tool[34] to provide the best alignment.

4. Experiments

4.1. Datasets

This work aims at evaluating speech separation as a front-end for ASR for long form CTS data. In order to train and evaluate the separation model, we use the two channel Fisher [35, 36] and CALLHOME American English (CHAE) [30] datasets respectively. Fisher contains 2000 hours and CHAE contains $\sim$ 60 hours of largely two speaker conversations, available in separate channels. Two speaker conversations were filtered from Fisher using the call and speaker metadata. We generate mixture data by mixing both the channels into a single channel and labeling the individual channels as the ground truths for speech separation. Along with these CTS datasets, we also use synthetic fully overlapping Libri2Mix [27] and wsj0-mix [10] to train the models using the data selection strategy outlined in §4.3. We use the default CHAE dev and test sets [30] for finding the best sampling coefficient (§4.3) and separation quality evaluation respectively. For ASR evaluations, we use HUB5 2000 English data [31] which is a $\sim$ 11 hours CTS dataset used for evaluating CTS ASR systems. Popular benchmarks [37, 38] on HUB5 consider the telephone channels separately. In this work, we mix them to create single-channel HUB5 and report ASR performance before and after speech separation. The average duration of a conversation in CHAE is $\sim$ 30 minutes and is $\sim$10 minutes for Fisher and HUB5 datasets. All CTS datasets in this work have a sampling rate of 8KHz and the synthetic mixes were downsampled to 8KHz.

4.2. Implementation Details

We solve for two-speaker separation use-case in this paper but this can be extended to more speakers by training the separator for multi speaker use-case similar to the base separation architectures in [14–17]. The SENet for this work is modelled using the ResNet34 architecture and is pretrained with a combination of classification and metric loss [39] with 12k speakers and 4k hours of CTS data. The frame duration for embedding extraction is 0.5 seconds and the embedding dimension is 512. The extracted embeddings were augmented with Gaussian noise [21] in addition to the implicit noise due to overlapping speech and clustering errors for training the separator. At train time, we also randomly flip the pooled embeddings along with the target separation signals. This is dynamically applied to 50% of the training samples and helps improve the generalization of the system. The separator in this work follows the best ConvTasNet architecture in [14] and is trained with 8s chunks. We train the separator with Adam optimizer with a batch size of 32 and learning rate of 1e-3 for 100 epochs. We set M to a large value of 6 (analysis on test set in §4.4) to account for any noisier recordings in training and N is 2 for the two-speaker separation use-case. We call the ConvTasNet trained with PIT loss as undirected speech separator (USS) as it produces outputs in a nondeterministic order. For the conversational English ASR system, we use the pretrained Aspire model from Kaldi [40].

Table 1: *SI-SDR (dB) of a ConvTasNet (USS) model trained and evaluated on different simulated and real datasets. CHAE is evaluated at the chunk level*

Train Data	wsj0mix	LibriMix	SparseLibriMix	CHAE
wsj0mix	15.8	8.3	8.1	-2.2
LibriMix	14.2	14.5	22	4.2
Fisher	9.4	12	20	14.1
RealSynMix	14.2	14.4	21.8	**15.5**

4.3. Data Sampling Strategy

Previous works [27–29] have reported subpar separation performance on realistic datasets when trained with the fully overlapping synthetic datasets. In addition, we also observe that relying only on real conversational data is not optimal as the amount of single speaker regions outweighs the amount of overlapping speaker regions by a large margin (approximately 10:1), causing skewed data for training the separation models. So, we propose a data sampling strategy (RealSynMix) which leverages both synthetic mixes (LibriMix) along with CTS data (Fisher). During training, we sample the fully overlapping synthetic data and real conversational data parameterized by a sampling coefficient which defines the ratio of real to synthetic utterances to be sampled in each batch and is treated as a hyperparameter, learnt by optimizing the separation performance on the CHAE dev-set. We choose a sampling coefficient of 6 (6 parts of the Fisher sampled with 1 part of LibriMix) for our experiments, as it has the lowest negative SI-SDR from Figure 2.

We also compare the performance of this sampling strategy on commonly used separation datasets wsj0mix, LibriMix and SparseLibriMix [27] in Table 1. For these experiments, SparseLibriMix has been generated with 9% overlap to simulate the overlap in CHAE and the SI-SDRs are evaluated at the chunk level, where the chunk size was 8s. From Table 1, we can see that the performance of the model trained with RealSynMix significantly outperforms the performance of wsj0mix and LibriMix trained models on the real CHAE dataset while also performing well on the synthetic mixes. It also outperforms the Fisher only trained model on the simulated datasets as well as real CHAE. Though SparseLibriMix was also generated with the same overlap ratio as CHAE, the SDRs on SparseLibriMix with LibriMix trained models being much better than CHAE shows that the simulated sparse datasets derived from audiobooks don't fully capture the conversational structure and dynamics of CTS data well enough. Also, the synthetic mixes are derived from read speech whereas conversation speech is the typical use-case for speech separation and the ASR system that follows.

Table 2: *Chunk and Recording level SI-SDR (dB) on CHAE dataset at different recording durations to highlight the efficiency of the DSS system over USS system on long-form audio.*

Model	Max Clusters M	Chunk Level	Recording level at durations			
			20s	100s	300s	600s
USS	-	15.5	15.1	12.8	10.5	6.7
DSS	2	14.5	14.4	14.3	14.3	14.4
	3	16.5	16.5	16.4	16.5	16.5
	4	**16.6**	**16.6**	**16.4**	**16.5**	**16.6**
	5	16.6	16.6	16.4	16.5	16.6

Table 3: *SA-WER (%) on HUB5 (CH and Switchboard subsets). Full, Non-ovl are SA-WERs of full utterance and non-ovl regions*

Model	Hub5 Subsets			
	Callhome		Switchboard	
	Full	Non-ovl	Full	Non-ovl
None (unseparated)	26.3	20.7	25.5	13.3
Oracle channels	18.4	17.9	10.6	9.7
USS	52	48.3	46.2	42.8
DSS	**23.0**	**19.2**	**14.6**	**12.8**

4.4. Directed speech separation

To compare the performance of the DSS system with the USS system, we evaluate the chunk level SI-SDR on the held-out test subset of CHAE. To evaluate the directedness of the system, we evaluate the recording level SI-SDR for different durations of audio on the CHAE test set. For the recording level evaluations, the outputs of USS are stitched with adjacent overlapping chunk similarity as in [26]. The chunk size during inference is 8s with no overlap for DSS and has an overlap of 4s for USS with stitching.

From Table 2, we see that not only does the DSS system improve the chunk level separation quality, it also remains consistent across different durations of the recordings. On the other hand, the USS system performance degrades as the duration of recording increases. This is mainly due to error propagation following an erroneous stitched chunk as the stitching relies only on the adjacent chunks. These erroneous stitches can happen frequently based on the separation quality and as the number of chunks increase with the recording duration.

We also analyze the effect of number of clusters on the separation quality in Table 2 and show over clustering ($M > N$) improves the separation quality due to cleaner speaker clusters. It can be observed that the separation quality significantly improves for $M = 3$ compared to $M = 2$. This is due to some of the noisy and overlapping speech being attributed to the 3rd cluster for $M = 3$, producing cleaner and more robust top-2 speaker embeddings. The separation quality is almost identical once the number of clusters is not fewer than the number of speakers, i.e. $M > 2$. The separation quality slightly improves for $M = 4$ compared to $M = 3$ as few noisier utterances are assigned an extra cluster for the noisy/overlap regions. The maximum number of detected clusters using max eigen gap across all CHAE utterances was 4 and hence the results for $M > 4$ are exactly identical.

Finally, we evaluate the ASR performance of both the systems on the HUB5 dataset in Table 3. We pass single channel HUB5 through the separators followed by the ASR system to get the WERs of the separated audio. We also pass the single channel HUB5 directly through ASR without any speech separation frontend to get WERs for unseparated audio. The oracle SA-WER is obtained by passing the oracle speaker channels of the original multi-channel HUB5 independently through the ASR system. We also report the SA-WER on the non-overlap (non-ovl) regions, i.e. single speaker regions to compare the separation performance in areas of no speech overlap. ASCLite [41] which can align multiple hypotheses against multiple reference transcriptions, is used to calculate the SA-WERs.

The DSS model improves the SA-WER on HUB5 by 24% relative (13% and 43% on the CallHome (CH) and Switchboard (SWBD) subsets respectively) compared to unseparated HUB5 data which shows the clear advantage of the DSS frontend in conversational ASR. We see that the USS system fails heavily on both subsets in terms of SA-WER as well, similar to the SI-SDR numbers on long recordings. Another important observation is that the SA-WER of non-overlapping regions with the DSS frontend is also better than unseparated non-overlap (non-ovl) SA-WER though these regions comprise of just single speaker speech. This can be attributed to the ASR (mainly language models) having better context by separating the adjacent overlapping speech regions. Finally, there is still a good difference between the oracle SA-WER and the DSS SA-WER, suggesting that there is still room for improvement for the long form directed speech separation model.

5. Conclusion

In this work, we introduced a speaker conditioned directed speech separation (DSS) model for long form real conversational telephone speech (CTS). This uses an over-clustering based approach to extract robust speaker embeddings without the need for pre-enrolled utterances. This not only naturally directs and stitches the separated short chunks in the order of the extracted speaker embeddings, but also improves the separation quality of the short chunks. In addition, we highlighted drawbacks of using some of the popular simulated datasets for training a CTS separation model. We solved this by proposing a data sampling strategy that combines the benefits of both real and synthetic datasets which shows significant improvements on the speech separation quality for CTS data when compared to the synthetic datasets or real datasets alone. With the DSS model, we achieved with an SI-SDR improvement of 1dB on short form and 10dB on long form CALLHOME American English and a SA-WER improvement of $\sim 30\%$ on Hub5 dataset compared to the PIT based undirected speech separation (USS) model.

Future work will focus on scaling the system to a variable number of speakers, designing a block-online system instead of an offline system and improving the separation performance with stronger conditioning techniques and base separator architectures using Transformer networks [16, 17].

6. References

[1] T. Yoshioka, H. Erdogan, *et al.*, "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks," *Proc. Interspeech 2018*, pp. 3038–3042, 2018.

[2] J. Barker, S. Watanabe, *et al.*, "The fifth'chime'speech separation and recognition challenge: Dataset, task and baselines," *Proc. Interspeech 2018*, pp. 1561–1565, 2018.

[3] N. Kanda, R. Ikeshita, *et al.*, "The hitachi/jhu chime-5 system," in *Proc. CHiME 2018 Workshop on Speech Processing in Everyday Environments*, pp. 6–10, 2018.

[4] N. Kanda, Y. Gaur, *et al.*, "Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers," *Proc. Interspeech 2020*, pp. 36–40, 2020.

[5] N. Kanda, X. Chang, *et al.*, "Investigation of end-to-end speaker-attributed asr for continuous multi-talker recordings," 2020.

[6] N. Kanda, G. Ye, *et al.*, "End-to-end speaker-attributed asr with transformer," *arXiv e-prints*, pp. arXiv–2104, 2021.

[7] D. Raj, L. Lu, *et al.*, "Continuous streaming multi-talker asr with dual-path transducers," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7317–7321, IEEE, 2022.

[8] X. Chang, N. Kanda, *et al.*, "Hypothesis stitcher for end-to-end speaker-attributed asr on long-form multi-talker recordings," *arXiv preprint arXiv:2101.01853*, 2021.

[9] D. Wang and J. 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.

[10] J. R. Hershey, Z. Chen, *et al.*, "Deep clustering: Discriminative embeddings for segmentation and separation,"

[11] M. Kolbæk, D. Yu, *et al.*, "Multitalker 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.

[12] Y. Luo, Z. Chen, *et al.*, "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.

[13] Z.-Q. Wang, J. Le Roux, *et al.*, "Alternative objective functions for deep clustering," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 686–690, IEEE, 2018.

[14] 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, no. 8, pp. 1256–1266, 2019.

[15] Y. Luo, Z. Chen, *et al.*, "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)*, pp. 46–50, IEEE, 2020.

[16] J. Chen, Q. Mao, *et al.*, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.

[17] C. Subakan, M. Ravanelli, *et al.*, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21–25, IEEE, 2021.

[18] K. Zmolikova, M. Delcroix, *et al.*, "Auxiliary loss function for target speech extraction and recognition with weak supervision based on speaker characteristics," *Proc. Interspeech 2021*, pp. 1464–1468, 2021.

[19] Q. Wang, H. Muckenhirn, *et al.*, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," *Proc. Interspeech 2019*, pp. 2728–2732, 2019.

[20] K. Žmolíková, M. Delcroix, *et al.*, "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.

[21] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2840–2849, 2021.

[22] E. Nachmani, Y. Adi, *et al.*, "Voice separation with an unknown number of multiple speakers," in *International Conference on Machine Learning*, pp. 7164–7175, PMLR, 2020.

[23] F.-L. Wang, Y.-H. Peng, *et al.*, "Dual-path filter network: Speaker-aware modeling for speech separation," *arXiv preprint arXiv:2106.07579*, 2021.

[24] J. Byun and J. W. Shin, "Monaural speech separation using speaker embedding from preliminary separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2753–2763, 2021.

[25] C. Han, Y. Luo, *et al.*, "Continuous speech separation using speaker inventory for long multi-talker recording," *arXiv e-prints*, pp. arXiv–2012, 2020.

[26] Z. Chen, T. Yoshioka, *et al.*, "Continuous speech separation: Dataset and analysis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7284–7288, IEEE, 2020.

[27] J. Cosentino, M. Pariente, *et al.*, "Librimix: An open-source dataset for generalizable speech separation," 2020.

[28] B. Kadıoğlu, M. Horgan, *et al.*, "An empirical study of conv-tasnet," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7264–7268, IEEE, 2020.

[29] T. Menne, I. Sklyar, *et al.*, "Analysis of deep clustering as preprocessing for automatic speech recognition of sparsely overlapping speech," 2019.

[30] D. G. Canavan, Alexandra and G. Zipperlen, "Callhome american english speech ldc97s42," *Web Download. Philadelphia: Linguistic Data Consortium*, 1997.

[31] e. a. Cieri, Christopher, "2000 hub5 english evaluation speech ldc2002s09," *Web Download. Philadelphia: Linguistic Data Consortium*, 2002.

[32] Q. Wang, C. Downey, *et al.*, "Speaker diarization with lstm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5239–5243, IEEE, 2018.

[33] J. Le Roux, S. Wisdom, *et al.*, "Sdr–half-baked or well done?," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, IEEE, 2019.

[34] "Available as part of the speech recognition scoring toolkit (sctk): https://github.com/usnistgov/sctk,"

[35] e. a. Cieri, Christopher, "Fisher english training speech part 1 speech ldc2004s13," *Web Download. Philadelphia: Linguistic Data Consortium*, 2004.

[36] e. a. Cieri, Christopher, "Fisher english training part 2, speech ldc2005s13," *Web Download. Philadelphia: Linguistic Data Consortium*, 2005.

[37] W. Xiong, J. Droppo, *et al.*, "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.

[38] G. Saon, G. Kurata, *et al.*, "English conversational telephone speech recognition by humans and machines," *Proc. Interspeech 2017*, pp. 132–136, 2017.

[39] J. S. Chung, J. Huh, *et al.*, "In defence of metric learning for speaker recognition," *Proc. Interspeech 2020*, pp. 2977–2981, 2020.

[40] D. Povey, A. Ghoshal, *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF, IEEE Signal Processing Society, 2011.

[41] J. G. Fiscus, J. Ajot, *et al.*, "Multiple dimension levenshtein edit distance calculations for evaluating automatic speech recognition systems during simultaneous speech," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 2006.