EasyChair Preprint

№ 10685

Single Channel Source Separation in the Wild - Conversational Speech in Realistic Environments

Emil Berger, Barbara Schuppler, Martin Hagmüller and Franz Pernkopf

August 7, 2023

Single Channel Source Separation in the Wild – Conversational Speech in Realistic Environments

Emil Berger, Barbara Schuppler, Martin Hagmüller, Franz Pernkopf

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Graz, Austria.
Email: berger@student.tugraz.at, {b.schuppler, hagmueller, pernkopf}@tugraz.at

Abstract

Recent progress in Single Channel Source Separation (SCSS) using deep neural networks led to impressive performance gains while also increasing the model sizes, requiring tremendous data resources. This demand is covered by artificially composed speech and noise mixtures that do not capture real-life characteristics of conversations taking place in noisy environments. This paper introduces a new dataset containing task-oriented dialogues spoken in a realistic environment and presents experimental results for two SCSS architectures - the Conv-TasNet and the transformer-based MossFormer. Overall, we observe a severe drop in performance of up to 4.3dB (SI-SDR improvement) for the 8kHz variant of the Conv-TasNet. For speaker pairs of homogeneous sex, the difference is even higher of up to 6dB. Only the model using 16kHz sample rate performs on a comparable level for speaker pairs of mixed sex. Our findings illustrate the need of using realistic data for both, training and evaluating.

1 Introduction

Robustness against noise, reverberation, and interfering audio signals has been identified as one of the grand challenges in speech recognition and understanding technology. Automatic speech recognition (ASR) for single-talker scenarios is performing reliably in acoustically clean environments. However, in harsh environments where the speech signal is distorted by interference with other acoustic sources or where simply the distance to the microphone is large, ASR performs far from satisfactory. In case of multiple interfering speakers, this is known as *cocktail party problem*. Driven by the success of deep learning, both speaker separation and speech enhancement have made major advances over the last years [1].

The focus of this paper is on single-channel speech separation (SCSS), for which different approaches have been developed. Frequency-domain algorithms such as Deep Clustering (DC) [2], Permutation Invariant Training (PIT) [3] and Deep Attractor Network (DAN) [4] rely solely on spectral features. Time-domain algorithms such as Wave-U-Net [5], TasNet [6] and Conv-TasNet [7] delivered promising results. Recently, some of these single-channel algorithms have been combined with a mask-based beamformer. In particular, a neural network estimates a gain mask of the desired signal, which is then used to construct a frequency-domain beamformer, i.e.: Beam-TasNet [8], SpeakerBeam [9], and Convolutional Beamforming [10]. More recently, end-to-end multi-channel speech separation has been done entirely in time domain [11, 12]. In [13] a sequential estimation of up to ten sources is performed in a single-channel setting using a *transformer-based* model. One of the most recent developments and thus representing the state-of-the-art performance is the MossFormer [14], which uses a joint local and global self-attention mechanism.

True for all recent developments is that the SCSS models are getting larger and thus they also require larger amounts of training data. For SCSS, a number of comprehensive data sets are the backbone of the current development: WSJ0 [2], WHAM! [15] or WHAMR! [16] and LibriMix [17]. Most of the available models were pretrained and/or benchmarked on one of these datasets. WHAM! and WHAMR! were created because of the need for more natural ambiences. WHAM! (WSJ0 Hipster Ambient Mixtures Dataset) uses the WSJ0 speech data and adds ambient noise of cafés and bars from the San Francisco Bay area. WHAMR! additionally includes reverberation. LibriMix uses speech data of the LibriSpeech corpus [18] and the ambience recordings of WHAM!. With its subsets *SparseLibri2Mix* and *SparseLibri3Mix* it provides evaluation sets for different overlapping speech situations. Approaches to face this problem were LibriCSS [19] and MMS-MSG [20].

Whereas these datasets provide more natural and realistic speech mixtures with environmental noise, one problem still remains: the mixtures are artificially composed conversations. The speech of two speakers are recorded separately in a silent environment and are then mixed in full overlap (i.e., the speech of both speakers is always present), as illustrated in Figure 1 (left). This is a rather unrealistic scenario, for several reasons. First, it is well known that speakers also adapt their speech tempo and spectral range when speaking in noise (e.g., for a survey cf. [21]). Second, when speakers are in dialogue, they mainly overlap at the beginning and end of chunks and only partly speak in full overlap (e.g., [22]). This challenges source separation even more so in recordings with speakers of the same sex, where the Euclidean distance in the feature space is small. Third, the speech in mentioned datasets does not contain characteristic properties of conversational speech such as hesitations, disfluencies, backchannels, laughter, nor other speaker noise [23, 24]. Last but not least, in natural conversations produced speech varies largely in loudness, not only from speaker to speaker, but also for one speaker within a larger conversation.

In this paper, we introduce different pre-processing methods and apply the Conv-TasNet and the MossFormer to perform SCSS for a newly collected real-life speech database, in which spontaneous dialogues were recorded between sex-homogenoeus and sex-heterogeneous speaker pairs while speaking in a large hall, having bubble noise from speakers of the same language in the background. We thus take up the challenges of 1) not having a speech reference from a silent scenario nor 2) having a continuously similar background noise, while 3) dealing with mentioned characteristics of spontaneous conversations.

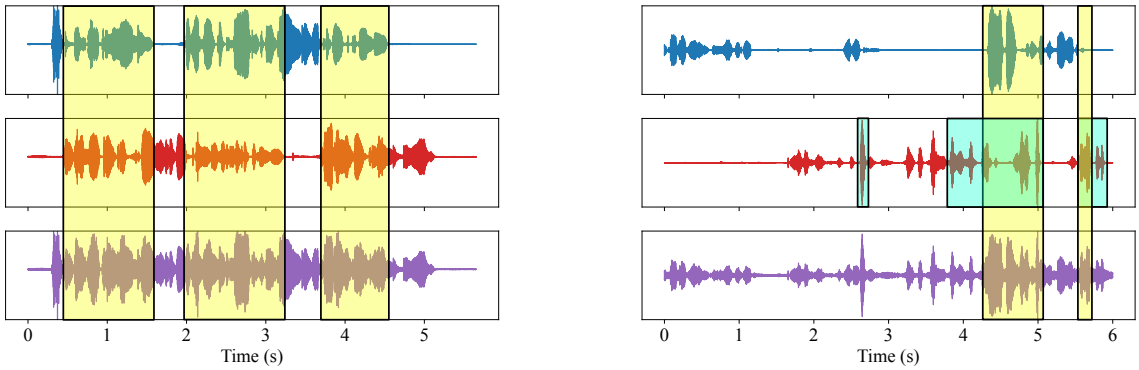

Figure 1: Waveforms of an example of LibriMix (left) and GRASS StudyFair (right): speaker 1 (upper, blue), speaker 2 (middle, red), mixture (lower, purple). The yellow boxes show segments of overlapping speech, the teal boxes show laughter and laughed speech of speaker 2.

2 Materials

2.1 GRASS StudyFair Corpus

For our experiments, we used the GRASS StudyFair Corpus (not published so far). It contains spontaneous, task-oriented conversations where Austrian speakers act in two roles: either as advisor at a study fair for a curriculum they studied themselves or as student who is visiting the fair and who is interested in pursuing a specific study. Each speaker is recorded in two conversations, once acting as advisor and once as student. These conversations lasted for approx. 15min each. In total, the corpus contains the speech from 20 speakers in sex-homogeneous and sex-heterogeneous couplings. The speakers did not know each other prior to the recordings.

The conversations took place in a large hall (76.4m x 3.2m x 10.75m, LxWxH) with T_{60} ≈1 s, in which - to enhance the ambience background- equally distributed 11 loudspeakers played bubble noise created from conversational speech from the GRASS Corpus [25]. Speakers wore a headset microphone (AKG HC577L), the background noise was recorded by four room microphones (AKG C480) and 13 booths separated by acoustically absorbent walls created an authentic atmosphere.

Figure 1 shows one example of LibriMix (left) which illustrates that the speech is rather static with a full overlap, and one example from the GRASS Study Fair (right), where the speech is more dynamic, contains less overlap but additional speaker noise (in this case: laughter). These characteristics of conversational speech are not covered by currently used datasets in the field of SCSS. For our experiments, we assigned the conversations of the GRASS Study Fair corpus to the training, validation and test datasets as shown in Table 1.

Sex	Speaker Pair	Time	Subset
F & M	030F & 005M	14m 28s	Train
		14m 39s	Train
	017M & 042F	13m 32s	Train
		15m 46s	Train
	027F & 003M	13m 12s	Val.
		13m 49s	Val.
	015M & 040F	08m 33s	Test
		06m 47s	Test
	Sum:	1h 40m 46s	
F & F	028F & 025F	09m 58s	Train
		12m 09s	Train
	023F & 022F	14m 52s	Train
		23m 43s	Train
	024F & 026F	07m 12s	Val.
		13m 10s	Val.
	038F & 041F	10m 09s	Test
		08m 30s	Test
	Sum:	1h 39m 43s	
M & M	008M & 001M	23m 34s	Train
		17m 56s	Train
	013M & 043M	13m 20s	Test
		15m 28s	Val.
	Sum:	1h 10m 18s	
Overall:		*4h 30m 47s*	

Table 1: List of all conversations of GRASS StudyFair. Each speaker has a unique speaker ID as defined in the GRASS Corpus [25].

2.2 Data Processing

Recording in harsh environments, even using high quality directional microphones, does not result in usable ground-truth signals, as there is always cross-talk and background noise. For this reason, we used the Audio Unit [26] "AUSoundIsolation" available in Logic Pro [27], which performs speech enhancement and denoises the speech signals from the headset. Subsequenly, we applied a limiter and normalised the signals to a maximal absolute amplitude of 0.5.

Cross-talk between speakers cannot be removed completely by this approach. We thus applied the following method to suppress the cross-talk: Let $\mathbf{X}_1$, $\mathbf{X}_2 \in \mathbb{C}^{N\times M}$ be the Short-Time Fourier Transform (STFT) of two audio signals $x_1, x_2 \in \mathbb{R}^L$ containing cross-talk, i.e., signal components from each other. We chose block length and hop size as 2048 and 1024 samples, respectively. For each block, we determine the spectral difference

$$\mathbf{X}_{diff} = \frac{\text{abs}(\mathbf{X}_1) - \text{abs}(\mathbf{X}_2)}{\text{abs}(\mathbf{X}_1) + \text{abs}(\mathbf{X}_2)}$$

from which two masks $\mathbf{H}_1, \mathbf{H}_2 \in \{0,1\}^{N\times M}$ are derived

as

$$\mathbf{H}_1(i,j) = \begin{cases} 0, & \text{if } \mathbf{X}_{diff}(i,j) < T \\ 1, & \text{otherwise} \end{cases},$$

$$\mathbf{H}_2(i,j) = \begin{cases} 0, & \text{if } (-\mathbf{X}_{diff}(i,j)) < T \\ 1, & \text{otherwise} \end{cases}$$

using a threshold T which we select as 0.5. Finally, the signals with reduced cross-talk are calculated by applying these masks as

$$s_1 = \text{STFT}^{-1}(\mathbf{X}_1 \odot \mathbf{H}_1),$$
$$s_2 = \text{STFT}^{-1}(\mathbf{X}_2 \odot \mathbf{H}_2),$$

where $\odot$ denotes the elementwise product. s_1 and s_2 are used as target signals for SCSS. We obtain the *restored mixture* by adding both signals with equal weight. Furthermore, we added the recorded noise with an SNR level drawn from a normal distribution, i.e.,

$$SNR_{noisy\_mix} \sim \mathcal{N}(\mu = 12\text{dB},\ \sigma = 5\text{dB})$$

to simulate different situations. Figure 2 shows these processing steps where h_1 and h_2 represent the recordings of the headset microphones and "AU_SI" is the Audio Unit [26]. Finally, the mixtures and targets were cut in samples of 10 seconds each and resampled to 8kHz and 16kHz using librosa [28].

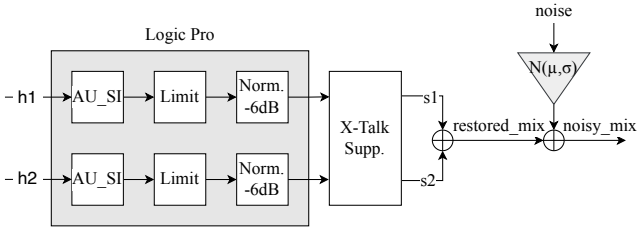

Figure 2: Processing steps of the recorded signals and mixture creation.

2.3 Speaker Activity

Speaker activity between a target $s \in \{s_1, s_2\}$ and the restored mixture m is defined as the normalised cross-correlation

$$\overline{R}_{sm,i} = \frac{\max(R_{sm,i})}{\max_{i=0,\ldots,N-1}(R_{sm,i})},$$

where i denotes the sample (i.e. 10s chunk). As voice activity detectors would also recognise the cross-talk, this method provides more accurate results – the higher the cross-correlation, the higher the speaker activity. Figure 3 shows that in most of the samples (i.e. 10s chunks) both speakers contribute and that only a few outliers exist where one speaks all the time while the second does not. For the (in large datasets over-represented) case of massive overlap (upper right quadrant in Figure 3 (a)), however, only a few samples could be identified. In our study fair setting, the most common mode was that one speaker delivering a pseudo-monologue while the other speaker was back-channeling or asking short questions. This is also shown in the histogram of Figure 3 (b), where higher correlation levels are mainly caused by samples of the 'advisor', while the first few bins (silence or almost silence) contained mainly speech produced by 'students'.

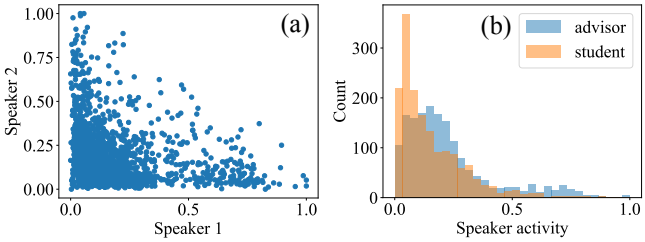

Figure 3: Normalised cross-correlation between speech signal and mixture: (a) Speaker one versus speaker two for each chunk, (b) Histogram for advisors against students.

3 Models

3.1 Conv-TasNet

The Conv-TasNet [7] uses an encoder-masking-decoder approach for SCSS in time-domain. In the encoder and decoder, a 1-D convolution models the waveform while the masking network consists of stacked 1-D dilated convolutional blocks with skip-connections. The mask is applied on the encoder output to separate the speech mixture. This model uses a relatively low number of parameters (5.1 millions) which makes it an ideal candidate for fine-tuning on limited data as in our case. Further reasons for choosing this model for our experiments were the detailed cross-dataset evaluation reported by Kadıoğlu et al. [29] and the good results achieved for different overlaps on SparseLibriMix [17].

3.2 MossFormer

Similar as the Conv-TasNet, the MossFormer [14] uses a convolutional encoder-decoder architecture and a masking network. In recent years, transformer models have been proposed for the masking network [30] to learn both short and long-term dependencies. The MossFormer introduces a gated single-head transformer model with joint local and global self-attention. This extension facilitates a more effective modelling of long-term dependencies and slightly improves the performance. For our experiments, we used the pre-trained MossFormer with a total of 42 millions of parameters. Here, we investigate if the MossFormer, which outperforms the Conv-TasNet by a large margin, can be effectively fine-tuned on our limited data set.

4 Experiments

4.1 Experimental Setup

Regarding the dataset split, we followed two rules: a) we wanted the sex combinations equally distributed in all subsets and b) no speaker pair should appear in more than one subset. As we had only two male & male speaker pairs, we decided to assign one pair for training and one for validating and testing. Overall, the dataset split resulted in a share of 59.32%, 23.21% and 17.47% for the training, validation and test set, respectively. Table 1 shows the detailed data composition.

We trained the Conv-TasNet with the train-360 subset of Libri2Mix [17] using the Adam [31] optimiser with a learning rate of 10^{-3} for 40 iterations and a PIT [3] loss using the SI-SDR as pairwise metric. For fine-tuning on our StudyFair corpus, we trained all parameters and used

the same setup with a lower learning rate of 10^{-4} and 10 iterations. For implementation, we used the Asteroid framework [32] and trained for 8kHz and 16kHz.

The MossFormer model was pre-trained on WSJ0-2mix [2] available at ModelScope [33] for 8kHz sampling rate. We fine-tuned this model using the Adam optimiser with a learning rate of $1.5 \cdot 10^{-4}$ for 30 iterations and a PIT loss using the SI-SNR as metric. For evaluation, we report the SI-SDR and SI-SDRi to make the results comparable with the results from the Conv-TasNet. We implemented the training script using ModelScope's API and SpeechBrain [34]. For evaluation we used Asteroid.

4.2 Results

Model	Sex	SI-SDR (dB)		SI-SDR$_i$ (dB)	
		8 kHz	16 kHz	8 kHz	16 kHz
Moss-Former	*all*	-2.83		-0.53	
	F&M	-2.03		0.55	
	F&F	-6.03		-3.93	
	M&M	0.77		3.03	
Conv-TasNet	*all*	5.41	6.37	7.71	8.67
	F&M	8.94	10.00	11.52	12.57
	F&F	3.70	5.16	5.81	7.26
	M&M	3.72	3.87	5.98	6.12

Table 2: Metrics of Conv-TasNet and MossFormer for different sex combinations and sampling frequencies, performed on the GRASS StudyFair test set.

Table 2 shows the Scale-invariant Signal to Distortion Ratio (SI-SDR) as well as the SI-SDR improvement (SI-SDRi) for all models, separately evaluated for all sex combinations. Both Conv-TasNet models, 8kHz and 16kHz, performed better for mixtures of mixed-sex in comparison to the mixtures of homogeneous sex. The MossFormer model performed worse than the Conv-TasNet in all conditions, independent of the speaker-pair constellation.

4.3 Discussion

For both, training and evaluating reasons, targets need to be free from background noise and cross-talk. As our recordings did not provide clean targets, we reduced the background noise using AUSoundIsolation and suppressed the cross-talk using a spectral masking technique. Every processing step, however, produces distorions. At 48kHz, a loss in clarity was audible after the processing with AUSoundIsolation. Regarding spectral masking, distortions were observed aurally at very loud moments of overlapping speech. These effects were clearly audible at 48kHz sampling rate and got majorly diminished after downsampling. Given that these processing steps did not work perfectly and leave artifacts of noise and cross-talk, we called the sum of the targets the *restored* instead of the *clean* mixture. We thus have to consider that for a system which separates perfectly, the evaluation score would be lower than for a system which produces the same artifacts (slight cross-talk and musical noise).

After this data processing procedure, we conducted experiments with the Conv-TasNet and with the MossFormer. In general, the MossFormer model performed worse than the Conv-TasNet in all cases. The low performance of the MossFormer is likely to be caused by the large parameter count of the available pre-trained model in comparison to the small-scale dataset used for fine-tuning, as well as by the fact that it was pre-trained on clean data using WSJ0. Our results thus support the findings by Kadıoğlu et al. [29], who reported that models pre-trained on WSJ0 may not generalise well to other (more realistic) datasets.

The SI-SDRi of Conv-TasNet trained on Libri2Mix for noisy data was reported as 12dB for 8kHz and 13.5dB for 16kHz in Cosentino et al. [17] for the standard Libri2Mix test set. Their cross-dataset evaluation for LibriMix showed that the performance of a model trained on LibriMix drops with approx. 1-2dB when evaluated on the WHAM! test set in comparison to a drop of 4dB for the other way around. For the standard Libri2Mix test set, Conv-TasNet obtained an SI-SDRi of 12dB for 8kHz and 13.5dB for 16kHz for the noisy separation task . Finally, they also reported an improvement for lower overlaps of up to 14.5dB for 8kHz sampling rate [17]. In our experiments, however, we observe a severe drop in performance even after fine-tuning, especially for the 8kHz model. We achieved a performance comparable to Cosentino et al. [17] only for speaker mixtures of heterogeneous sex and only when using a sample rate of 16kHz.

5 Conclusion

This paper presented SCSS experiments on the newly introduced GRASS StudyFair Corpus, containing realistic conversations in a natural ambience. We processed our recordings using AUSoundIsolation as well as spectral masking for cross-talk suppression and performed experiments on Conv-TasNet and MossFormer. The latter under-performed due to the small-scale dataset in comparison to the model size. Furthermore, the MossFormer was pre-trained on the clean WSJ0 data limiting its performance on noisy mixtures. Conv-TasNet was trained on LibriMix, the currently most sophisticated dataset. The Conv-TasNet fine-tuned on our StudyFair corpus also showed a substantial performance drop. We further showed that a separate evaluation of the speaker's sex combinations is important as this causes a varying difficulty level accompanied by immense differences in evaluation scores.

Current models can only obtain their reported scores under laboratory conditions, when used 'in the wild' the scores are dropping quickly. Not only the application, but also the evaluation gets unreliable when dealing with realistic data. To face this challenge, the SCSS community has to consider both, the design of new datasets for training which cover the effects of conversational speech and the development of methods to deal with noisy data.

6 Acknowledgements

This work was partly funded by grant P-32700-NB from FWF (Austrian Science Fund). We thank J. Balint and M. Stavric, with whom we collaboratively created the GRASS Study Fair Corpus.

References

[1] 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.

[2] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 31–35, 2016.

[3] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 241–245, 2017.

[4] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 246–250, 2017.

[5] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[6] Y. Luo and N. Mesgarani, "TaSNet: Time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 696–700, 2018.

[7] 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.

[8] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, "Beam-TasNet: Time-domain audio separation network meets frequency-domain beamformer," in *ICASSP 2020 - 2020 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 6384–6388, 2020.

[9] 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.

[10] T. Nakatani, R. Takahashi, T. Ochiai, K. Kinoshita, R. Ikeshita, M. Delcroix, and S. Araki, "DNN-supported mask-based convolutional beamforming for simultaneous denoising, dereverberation, and source separation," in *ICASSP 2020 - IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 6399–6403, 2020.

[11] 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.

[12] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020 - IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 46–50, 2020.

[13] S. Lutati, E. Nachmani, and L. Wolf, "SepIt: Approaching a single channel speech separation bound," *arXiv preprint arXiv:2205.11801*, 2022.

[14] S. Zhao and B. Ma, "MossFormer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions," in *ICASSP 2023-2023 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 1–5, IEEE, 2023.

[15] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "WHAM!: Extending speech separation to noisy environments," *arXiv preprint arXiv:1907.01160*, 2019.

[16] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *ICASSP 2020-2020 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 696–700, IEEE, 2020.

[17] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," pp. 5206–5210, 2015.

[19] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: Dataset and analysis," in *ICASSP 2020-2020 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 7284–7288, IEEE, 2020.

[20] T. Cord-Landwehr, T. Von Neumann, C. Boeddeker, and R. Haeb-Umbach, "Mms-msg: A multi-purpose multi-speaker mixture signal generator," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, IEEE, 2022.

[21] H. Brumm and S. A. Zollinger, "The evolution of the lombard effect: 100 years of psychoacoustic research," *Behaviour*, vol. 148, no. 11-13, pp. 1173 – 1198, 2011.

[22] E. Kurtic, G. J. Brown, and B. Wells, "Resources for turn competition in overlap in multi-party conversations: speech rate, pausing and duration.," in *INTERSPEECH*, pp. 2550–2553, 2010.

[23] B. Schuppler, M. Hagmüller, and A. Zahrer, "A corpus of read and conversational Austrian German," *Speech Communication*, vol. 94, pp. 62–74, 2017.

[24] B. Ludusan and B. Schuppler, "To laugh or not to laugh? the use of laughter to mark discourse structure," in *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 76–82, 2022.

[25] B. Schuppler, M. Hagmüller, J. A. Morales-Cordovilla, and H. Pessentheiner, "GRASS: The Graz corpus of read and spontaneous speech.," in *LREC*, pp. 1465–1470, 2014.

[26] Apple Inc., "Audio Unit." `https://developer.apple.com/documentation/audiounit`, 2023. Accessed: 2023-05-16.

[27] Apple Inc., "Logic Pro." `https://www.apple.com/logic-pro/`, Version 10.7.7, 2023. Accessed: 2023-05-16.

[28] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, pp. 18–25, 2015.

[29] B. Kadıoğlu, M. Horgan, X. Liu, J. Pons, D. Darcy, and V. Kumar, "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.

[30] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 21–25, IEEE, 2021.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, *et al.*, "Asteroid: The PyTorch-based audio source separation toolkit for researchers," *arXiv preprint arXiv:2005.04132*, 2020.

[33] ModelScope.cn Community, "ModelScope." `https://www.modelscope.cn/home`. Accessed: 2023-04-21.

[34] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, *et al.*, "SpeechBrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.