LibriMix: An Open-Source Dataset for Generalizable Speech Separation

Joris Cosentino[1], *Manuel Pariente*[1], *Samuele Cornell*[2], *Antoine Deleforge*[1], *Emmanuel Vincent*[1]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[2]Department of Information Engineering, Università Politecnica delle Marche, Italy

joris.cosentino@inria.fr, manuel.pariente@inria.fr, s.cornell@pm.univpm.it,
antoine.deleforge@inria.fr, emmanuel.vincent@inria.fr

Abstract

In recent years, wsj0-2mix has become the reference dataset for single-channel speech separation. Most deep learning-based speech separation models today are benchmarked on it. However, recent studies have shown important performance drops when models trained on wsj0-2mix are evaluated on other, similar datasets. To address this generalization issue, we created LibriMix, an open-source alternative to wsj0-2mix, and to its noisy extension, WHAM!. Based on LibriSpeech, LibriMix consists of two- or three-speaker mixtures combined with ambient noise samples from WHAM!. Using Conv-TasNet, we achieve competitive performance on all LibriMix versions. In order to fairly evaluate across datasets, we introduce a third test set based on VCTK for speech and WHAM! for noise. Our experiments show that the generalization error is smaller for models trained with LibriMix than with WHAM!, in both clean and noisy conditions. Aiming towards evaluation in more realistic, conversation-like scenarios, we also release a sparsely overlapping version of LibriMix's test set.

Index Terms: Speech separation, generalization, corpora.

1. Introduction

A fundamental problem towards robust speech processing in real-world acoustic environments is to be able to automatically extract or separate target source signals present in an input mixture recording [1]. To date, state-of-the-art performance on the single-channel speech separation task is achieved by deep learning based models [2–6]. In particular, end-to-end models which directly process the time-domain samples seem to obtain the best performance [7, 8]. Such systems (e.g. Conv-TasNet [4], Dual-path RNN [5] or Wavesplit [6]) perform so well in separating fully overlapping speech mixtures from the wsj0-2mix dataset [2] that the separated speech estimates are almost indistinguishable from the reference signals. This led to the development of WHAM! [9] and WHAMR! [10], respectively the noisy and reverberant extensions of wsj0-2mix.

While these datasets have moved the field towards more realistic and challenging scenarios, there are still steps to be made. In fact, a recent study reports important drops of performance when Conv-TasNet is trained on wsj0-2mix and tested on other comparable datasets [11]. This suggests that, even though Conv-TasNet's separation quality is close to perfect on wsj0-2mix, the ability to generalize to speech coming from a wider range of speakers and recorded in slightly different conditions has not yet been achieved. Additionally, fully overlapping speech mixtures such as the ones from wsj0-2mix are unnatural. Real-world overlap ratios are typically in the order of 20% or less in natural meetings [12] and casual dinner parties [13]. A few studies have shown that speech separation algorithms trained on fully overlapping speech mixtures do not generalize well to such sparsely overlapping mixtures [6, 14]. Finally, models relying on some kind of speaker identity representation [2, 6, 15] cannot easily detect overfitting, since wsj0-2mix's speakers are shared between the training and validation sets.

There have been few initiatives to address these issues. A sparsely overlapping version of wsj0-2mix proposed in [14] has shown the limitation of Deep Clustering [2] on such mixtures. As the original utterances are the same as the ones from wsj0-2mix, we expect the generalization issue to remain the same. In [11], a new speech separation dataset based on LibriTTS [16] has been designed. The results show that generalizability is improved thanks to the variability of recording conditions and the larger number of unique speakers in the dataset. Sadly, the dataset is limited to two-speaker mixtures without noise, and has not been open-sourced. LibriCSS [17], an open-source dataset for sparsely overlapping continuous speech separation, has recently been released. While it addresses most of the shortcomings of wsj0-2mix, its short 10-hour duration restricts its usage to evaluation rather than training purposes. Real dinner-party recordings [13, 18] as well as meeting recordings [19, 20] are also available. While these are natural recordings, the clean speech signals for individual sources are not available[1] and thus, speech separation algorithms cannot be directly evaluated in terms of usual speech separation metrics [21, 22].

In this work, we introduce LibriMix, an open-source dataset for generalizable noisy speech separation composed of two- or three-speaker mixtures, with or without noise. The speech utterances are taken from LibriSpeech [23] and the noise samples from WHAM! [9]. An additional test set based on VCTK [24] is designed for fair cross-dataset evaluation. We evaluate the generalization ability of Conv-TasNet when trained on LibriMix or WHAM! and show that LibriMix leads to better generalization in both clean and noisy conditions. Stepping further towards real-world scenarios, we introduce a sparsely overlapping version of LibriMix's test set with varying amount of overlap. The scripts used to generate these datasets are publicly released[2,3,4].

The paper is organised as follows. We explain LibriMix's design and give some insights about its characteristics in Section 2. In Section 3, we report experimental results on LibriMix as well as across datasets. We conclude in Section 4.

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`).

[1] Close-talk signals are available as references, but these are too corrupted for the evaluation of modern separation algorithms.

[2] `https://github.com/JorisCos/LibriMix`

[3] `https://github.com/JorisCos/VCTK-2Mix`

[4] `https://github.com/popcornell/SparseLibriMix`

2. Datasets

In the following, we present existing speech separation datasets derived from Wall Street Journal (WSJ0), and introduce our new datasets derived from LibriSpeech. Statistics about the original speech datasets and the speech separation datasets derived from them can be found in Tables 1 and 2, respectively.

Table 1: *Statistics of original speech datasets.*

Dataset	Split	Hours	per-spk minutes	# Speakers
WSJ0	si_tr_s	25	15	101
	si_dt_05	1.5	11	8
	si_et_05	2.3	14	10
LibriSpeech clean	train-360	364	25	921
	train-100	101	25	251
	dev	5.4	8	40
	test	5.4	8	40
VCTK	test	44	24	109

Table 2: *Statistics of derived speech separation datasets.*

Dataset	Split	# Utterances	Hours
wsj0-{2,3}mix	train	20,000	30
	dev	5,000	8
	test	3,000	5
Libri2Mix	train-360	50,800	212
	train-100	13,900	58
	dev	3,000	11
	test	3,000	11
Libri3Mix	train-360	33,900	146
	train-100	9,300	40
	dev	3,000	11
	test	3,000	11
SparseLibri2Mix	test	3,000	6
SparseLibri3Mix	test	3,000	6
VCTK-2mix	test	3,000	9

2.1. WSJ0, wsj0-2mix and WHAM!

The WSJ0 dataset was designed in 1992 as a new corpus for automatic speech recognition (ASR) [25]. It consists of read speech from the Wall Street Journal. It was recorded at 16 kHz using a close-talk Sennheiser HMD414 microphone. The wsj0-2mix dataset [2] uses three subsets of WSJ0: `si_tr_s`, `si_dt_05` and `si_et_05` which all come from the 5k vocabulary part of WSJ0. This represents around 30 h of speech from 119 speakers. Table 1 reports details on speaker and hour distributions within the subsets.

The wsj0-2mix datatet is made of a training set, a validation set and a test set. The training and validation sets share common speakers from the `si_tr_s` subset and the test set is made from a combination of `si_dt_05` and `si_et_05`. Speech mixtures are generated by mixing pairs of utterances from different speakers at random signal-to-noise ratios (SNRs). The SNR is drawn uniformly between 0 and 5 dB. Four variations of the dataset are available, which correspond to two different sampling rates (16 kHz and 8 kHz) and two modes (*min* and *max*). In the *min* mode, the mixture stops with the shortest utterance. In the *max* mode, the shortest utterance is padded to the longest one. The wsj0-2mix equivalent for three-speaker mixtures is called wsj0-3mix and was generated in a similar way [2]. Note that, in order to generate more mixtures, utterances from WSJ0 were used multiple times in the three subsets. Each utterance is repeated up to fifteen times, with an average of four times.

In the WHAM! dataset, wsj0-2mix was extended to include noisy speech mixtures. Noise samples recorded in coffee shops, restaurants, and bars were added to the mixtures so that the SNR between the loudest speaker and the noise varies from -6 to +3 dB. The dataset follows the same structure as wsj0-2mix, with the same four variations and the three same subsets. In addition to separation in clean (*sep_clean*) and noisy conditions (*sep_noisy*), other enhancement tasks can be considered. Statistics on noise durations can be seen in Table 3.

WHAM! noises have been released under the CC BY-NC 4.0 License, but WSJ0 and derived data are proprietary (LDC). Note that no noisy version of wsj0-3mix has been released.

Table 3: *Statistics of WHAM!'s noises.*

Datasets	Split	Hours	Number of utterances
WHAM! noise	train	58	20,000
	dev	14.7	5,000
	test	9	3,000

2.2. LibriSpeech, LibriMix and sparse LibriMix

LibriSpeech [23] is a read ASR corpus based on LibriVox audiobooks[5]. To avoid background noise in the reference signals, we only use the `train-clean-100`, `train-clean-360`, `dev-clean`, and `test-clean` subsets of LibriSpeech. This represents around 470 h of speech from 1,252 speakers, with a 60k vocabulary. More statistics are given in Table 2.

We propose a new collection of datasets derived from LibriSpeech and WHAM!'s noises which we call LibriMix. These datasets are entirely open source.

The two main datasets, Libri2Mix and Libri3Mix, consist of clean and noisy, two- and three-speaker mixtures. Libri2Mix follows the exact same structure as WHAM! and allows for the same tasks. Mirroring the organization of LibriSpeech, they have two training sets (`train-100`, `train-360`), one validation set (`dev`) and one test set (`test`). In order to cover the `train-360` subset of LibriSpeech without repetition, training noise samples were speed-perturbed with factors of 0.8 and 1.2 as described in [26]. Instead of relying on signal power to scale individual utterances as in wsj0-2mix, we rely on loudness units relative to full scale (LUFS) [27][6], expressed in dB. Based on the ITU-R BS.1770-4 recommendation [27], LUFS measure the perceived loudness of an audio signal. Compared to classical SNRs, LUFS better correlate with human perception, are silence-invariant, and are little sensitive to downsampling.

Speech mixtures are generated by randomly selecting utterances for different speakers. The loudness of each utterance is uniformly sampled between -25 and -33 LUFS. Random noise samples with uniformly distributed loudness between -38 and -30 LUFS are then added to the speech mixtures. The noisy mixtures are then clipped to 0.9, if need be. The resulting SNRs are normally distributed with a mean of 0 dB and a standard deviation of 4.1 dB in the clean condition and a mean of -2 dB and a standard deviation of 3.6 dB in the noisy condition.

[5] https://librivox.org/

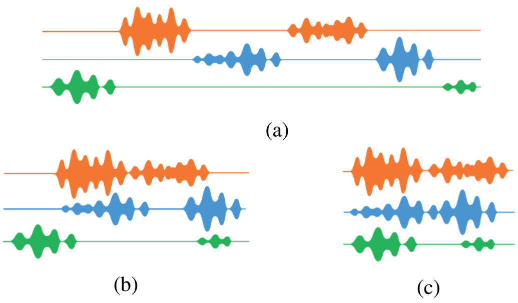
[6] Available at https://github.com/csteinmetz1/pyloudnorm

Note that in `train-100` and `train-360` each utterance is only used once. For `dev` and `test`, the same procedure is repeated enough times to reach 3,000 mixtures. This results in around 280 h of noisy speech mixtures, against 45 h for WHAM!. The variety of speakers is much wider in LibriMix's training set with around 1,000 distinct speakers against 100 in WHAM!. The total number of unique words is also much larger, with 60k unique words in LibriMix against 5k in wsj0-2mix.

Stepping towards more realistic, conversation-like scenarios, we also release sparsely overlapping versions of LibriMix's two- and three-speaker test sets. We refer to these datasets as SparseLibri2Mix and SparseLibri3Mix. For each mixture, we first sample speaker identities, then, for each speaker, we select an utterance from `test-clean`. Cycling through the selected utterances, we keep adding sub-utterances whose boundaries were obtained with the Montreal Forced Aligner (MFA) [28], until a maximum length of 15 s has been reached. This mixing process ensures that each speaker utters semantically meaningful speech, which is important for future ASR experiments. We used the same loudness distribution as the non-sparse version but we sampled it for each sub-utterance. This allows for alternating dominant speakers in the mixtures [6].

For both two- and three-speaker versions, we produced 500 mixtures for six different amounts of speech overlap: 0%, 20%, 40%, 60% 80%, and 100%. For three-speaker mixtures we count the amount of three-speaker overlap and not the total overlap, which is higher because two-speaker overlap also occurs. Note that these overlap ratios reflect the amount of overlap of each sub-utterance with the preceding ones. Because sub-utterances don't have the same length, the real overlap ratios of the mixtures are lower, as it happens with *max* versions of LibriMix and WHAM!.

Because WHAM! noise samples are short on average, the maximum mixture length was restricted to 15 s in order to obtain a reasonable number of samples for testing. Examples of such sparsely overlapping utterances can be visualized in Fig. 1.



(a)

(b)

(c)

Figure 1: *SparseLibri3Mix example with different 3-speaker overlap percentages: (a) 0% overlap, (b) 20% overlap, (c) 100% overlap.*

2.3. VCTK and VCTK-2mix

We also release VCTK-2mix, an unmatched open-source test set derived from VCTK [24]. VCTK comprises 109 native English speakers reading newspapers. As VCTK utterances contain a significant amount of silence, we use energy-based voice activity detection to remove silent portions with a 20 dB threshold.

The mixing procedure for VCTK-2mix is identical to that for LibriMix. The noise samples are also taken from WHAM!'s test set. The resulting dataset contains around 9 h of speech with 3,000 utterances from 108 speakers.

3. Results

In order to assess the results achievable using our newly released LibriMix datasets, we use the optimal configuration of Conv-TasNet reported in [4] for the separation tasks, as implemented in Asteroid [29] [7]. Training is done by maximizing the permutation-invariant, scale-invariant signal-to-distortion ratio (SI-SDR) [21, 30] on 3 s segments with a batch size of 24 and Adam [31] as the optimizer. All the experiments are performed with the exact same parameters. Since the SI-SDR is undefined for silent sources, results reported on all *max* versions correspond to models trained on the corresponding *min* version.

3.1. Results on LibriMix

The results achieved by Conv-TasNet on the clean and noisy versions of Libri2mix and Libri3Mix are reported in Table 4 and compared with the Ideal Binary Mask (IBM) and the Ideal Ratio Mask (IRM) for a short time Fourier transform (STFT) window size of 32 ms. Conv-TasNet was trained on `train-360`, which leads to better performance than `train-100`. Results are reported in terms of SI-SDR improvement compared to the input mixture (SI-SDR$_i$). We refer to the clean two-speaker separation task as *2spk-C*, to the noisy one as *2spk-N*, etc. We see that for two-speaker mixtures, Conv-TasNet outperforms ideal masks in clean conditions and is on par with them in noisy conditions, as in [4, 32]. However, oracle performance is still out of reach for three-speaker mixtures, with and without noise.

Table 4: *SI-SDR$_i$ (dB) achieved on LibriMix (SI-SDR for the "Input" column).*

	mode	Input	IRM	IBM	Conv-TasNet
2spk-C	8k min	0.0	12.9	13.7	14.7
	16k max	0.0	14.1	14.5	16
2spk-N	8k min	-2.0	12	12.6	12
	16k max	-2.8	13.4	13.7	13.5
3spk-C	8k min	-3.4	13.1	13.9	12.1
	16k max	-3.7	14.5	14.9	13
3spk-N	8k min	-4.4	12.6	13.3	10.4
	16k max	-5.2	14.1	14.4	10.9

3.2. Results on SparseLibriMix

We report the results obtained on the 8 kHz test sets of SparseLibri2Mix and SparseLibri3Mix in Table 5, in clean and noisy conditions. We used the same 8 kHz models as in Table 4, which were trained on non-sparse LibriMix. It can be seen that, for both two- and three-speaker mixtures, the higher the overlap, the lower the SI-SDR$_i$, as was also shown in [6]. In the 100% overlap case we obtain results similar to the ones in Table 4 for the non-sparse, 8kHz *min* version. The values are slightly higher here because mixtures are not truncated to the shortest utterance. Interestingly, we see that Conv-TasNet performs *worse* than IRM for smaller overlaps. This suggests that there is still room for improvement for source separation of sparsely-overlapping mixtures.

[7] github.com/mpariente/asteroid

Table 5: *SI-SDR$_i$ (dB) achieved on SparseLibriMix (8kHz). Conv-TasNet is abreviated TCN.*

	2spk-C		2spk-N		3spk-C		3spk-N	
Overlap	IRM	TCN	IRM	TCN	IRM	TCN	IRM	TCN
0%	43.7	31.9	16.1	14.5	44.2	24.8	18.7	13.0
20%	19.6	20.0	14.7	13.9	18.1	15.8	15.6	12.1
40%	16.2	17.6	13.8	13.2	16.4	14.4	14.9	11.7
60%	14.9	16.3	13.3	12.7	15.5	13.8	14.4	11.5
80%	14.2	15.7	13	12.5	14.6	13.1	13.9	11
100%	13.8	15.3	12.7	12.2	14.3	12.5	13.6	10.7

3.3. Dataset comparisons

The experiments in [11] have shown that models trained on wsj0-2mix do not generalize well to other datasets. Similarly to [11], we investigate the generalization ability of Conv-TasNet when trained on different datasets. We train six different Conv-TasNet models on WHAM! `train`, LibriMix `train-100` and `train-360` in both clean and noisy condition. We evaluate each model on the corresponding (clean or noisy) test sets of Libri2Mix, WHAM!, and VCTK2Mix. The results in clean and noisy conditions are shown in Figs. 2 and 3, respectively. Note that noise samples are matched across the three noisy test sets. For both clean and noisy separation, we can see that WHAM!-trained models poorly generalize to LibriMix, with a 4 dB SI-SDR drop compared to LibriMix-trained models, while LibriMix-trained models obtain closer performance to WHAM!-trained models on the WHAM! test set with a 0.8 dB SI-SDR drop only. On the clean and noisy versions of VCTK-2mix, WHAM!-trained models perform around 3–4 dB less well than models trained on Librimix's `train-360`. The general performance drop from models trained on LibriMix's `train-100` compared to LibriMix's `train-360` confirms again that the amount of data is key to better generalization and that the amount of data available in WHAM! is insufficient. Altogether, these results indicate that the clean and noisy versions of LibriMix allow better generalization than the wsj0-mix and WHAM! datasets.

Several factors can influence generalization. While VCTK-2mix was generated with statistics matching the ones in LibriMix, we argue that this is not the reason, as results reported in [11] go in the same direction. Instead, we believe that the number of speakers (100 against 900), the size of the vocabulary (5k against 60k), the recording conditions (same room same recording material against varying rooms and material) and the total amount of training data (30 h against 212 h) add up to explain that models trained with LibriMix's `train-360` offer better generalization.

Results reported in [11] are somewhat different than the ones we report here, which can be explained by several factors. First, the VCTK-based two-speaker test set in [11] was designed using the Matlab scripts from [2]. These scripts do not remove silences and compute SNRs based on signal power instead of LUFS. As utterances from VCTK can be filled with silence, this greatly increases the effective SNR range of mixtures. For example, a short utterance in a long silence mixed at 0 dB with a long utterance without silence can produce a mixture where the second source is almost inaudible. This explains the low performance obtained on VCTK in [11]. Second, the alternative training and test sets are based on LibriTTS [16] which is itself derived from LibriSpeech [23]. LibriTTS has shorter and cleaner utterances, which could explain the higher performance reported on its test set in [11], and the larger drop in performance when tested on wsj0-2mix's test set.

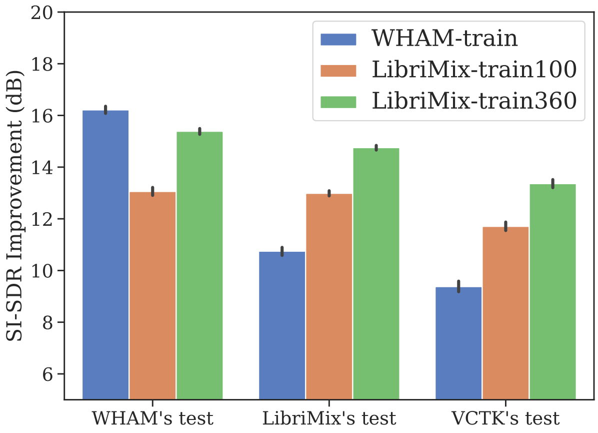


Figure 2: *Cross-dataset evaluation on the clean separation task. Errors bars indicate 95% confidence intervals.*

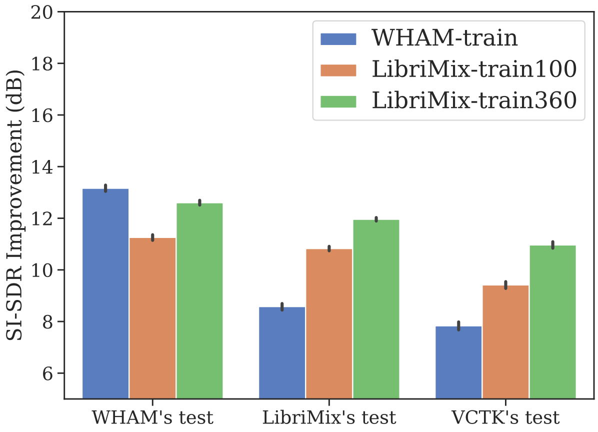


Figure 3: *Cross dataset evaluation on the noisy separation task. Errors bars indicate 95 % confidence interval*

4. Conclusions

In this work, we introduced LibriMix, a new family of datasets for generalizable single-channel speech separation. Libri2Mix and Libri3Mix enable two- and three-speaker separation in clean and noisy conditions. We report competitive results in all conditions using Asteroid's implementation of Conv-TasNet. A new independent test set, VCTK-2mix, is also released to enable reproducible cross-dataset evaluation. Experiments show that models trained on Libri2Mix generalize better to VCTK-2mix than models trained with WHAM!. Additionaly, Libri3Mix is the first open-source dataset to enable three-speaker noisy separation. Stepping towards more realistic scenarios, we release SparseLibri2Mix and SparseLibri3Mix, two- and three-speaker test sets consisting of sparsely overlapping speech mixtures with a varying amount of overlap. Initial results reported on it suggest that there still is room for improvement on this scenario. Future work includes the design of a training set of sparsely overlapping speech mixtures, as well as a more diverse set of noise samples.

5. References

[1] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*, 1st ed. Wiley, 2018.

[2] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: discriminative embeddings for segmentation and separation," in *ICASSP*, 2016, pp. 31–35.

[3] Y. Luo and N. Mesgarani, "TasNet: Time-domain audio separation network for real-time, single-channel speech separation," in *ICASSP*, 2018, pp. 696–700.

[4] ——, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[5] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP*, 2020, pp. 46–50.

[6] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *arXiv preprint arXiv:2002.08933*, 2020.

[7] F. Bahmaninezhad, J. Wu, R. Gu, S.-X. Zhang, Y. Xu, M. Yu, and D. Yu, "A comprehensive study of speech separation: Spectrogram vs waveform separation," in *Interspeech*, 2019, pp. 4574–4578.

[8] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying TasNet: A dissecting approach," in *ICASSP*, 2020, pp. 6359–6363.

[9] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "WHAM!: extending speech separation to noisy environments," in *Interspeech*, 2019, pp. 1368–1372.

[10] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *ICASSP*, 2020, pp. 696–700.

[11] B. Kadıoğlu, M. Horgan, X. Liu, J. Pons, D. Darcy, and V. Kumar, "An empirical study of Conv-TasNet," in *ICASSP*, 2020, pp. 7264–7268.

[12] Ö. Çetin and E. Shriberg, "Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: insights for automatic speech recognition," in *Interspeech*, 2006, pp. 293–296.

[13] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," *Interspeech*, pp. 1561–1565, 2018.

[14] T. Menne, I. Sklyar, R. Schlüter, and H. Ney, "Analysis of deep clustering as preprocessing for automatic speech recognition of sparsely overlapping speech," in *Interspeech*, 2019, pp. 2638–2642.

[15] Y. Liu and D. Wang, "Divide and conquer: A deep casa approach to talker-independent monaural speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, p. 2092–2102, 2019.

[16] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from libriSpeech for text-to-speech," in *Interspeech*, 2019, pp. 1526–1530.

[17] 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, pp. 7284–7288.

[18] M. V. Segbroeck, A. Zaid, K. Kutsenko, C. Huerta, T. Nguyen, X. Luo, B. Hoffmeister, J. Trmal, M. Omologo, and R. Maas, "DiPCo – dinner party corpus," *arXiv preprint arXiv:1909.13447*, 2019.

[19] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *ICASSP*, 2003.

[20] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction*, 2006, pp. 28–39.

[21] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[22] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR — half-baked or well done?" in *ICASSP*, 2019, pp. 626–630.

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

[24] J. M. K. Veaux, Christophe Yamagishi, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[25] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Workshop on Speech and Natural Language*, 1992, p. 357–362.

[26] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Interspeech*, 2015, pp. 3586–3589.

[27] ITU-R, "Recommendation ITU-R BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level," 2015.

[28] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Interspeech*, 2017, pp. 498–502.

[29] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," *arXiv preprint arXiv:2005.04132*, 2020.

[30] 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, pp. 241–245.

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

[32] M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Filterbank design for end-to-end speech separation," in *ICASSP*, 2020, pp. 6364–6368.