

Error Analysis in a Modular Meeting Transcription System

Peter Vieting*, Simon Berger*[‡], Thilo von Neumann[§], Christoph Boeddeker[§], Ralf Schlüter*[‡] and Reinhold Haeb-Umbach[§]

*Machine Learning and Human Language Technology Group, RWTH Aachen University, Germany

Email: {vieting,berger,schluter}@hltpr.rwth-aachen.de

[‡]AppTek GmbH, Germany

[§]Paderborn University, Germany

Email: {vonneumann,boeddeker,haeb}@nt.upb.de

Abstract

Meeting transcription is a field of high relevance and remarkable progress in recent years. Still, challenges remain that limit its performance. In this work, we extend a previously proposed framework for analyzing leakage in speech separation with proper sensitivity to temporal locality. We show that there is significant leakage to the cross channel in areas where only the primary speaker is active. At the same time, the results demonstrate that this does not affect the final performance much as these leaked parts are largely ignored by the voice activity detection (VAD). Furthermore, different segmentations are compared showing that advanced diarization approaches are able to reduce the gap to oracle segmentation by a third compared to a simple energy-based VAD. We additionally reveal what factors contribute to the remaining difference. The results represent state-of-the-art performance on LibriCSS among systems that train the recognition module on LibriSpeech data only.

Index Terms: speech separation, speech recognition, meeting transcription, LibriCSS.

1 Introduction

Meeting transcription is a task of increasing importance as it is key to enable a diverse set of applications. Different approaches have been proposed for meeting transcription that can be grouped into modular (e.g. [1, 2]) and end-to-end systems (e.g. [3, 4]). While modular systems consist of cascaded submodules and are more interpretable, end-to-end approaches have the advantage of taking only one single global decision. Despite impressive progress in recent years [4–7], meeting transcription is still challenging [8]. This is due to different factors such as a setting with far-field recordings and noisy acoustic conditions, the requirement of accurate speaker attribution depending on the application and the fact that separation of overlapping speech is still facing challenges such as leakage. In this work, we investigate components and factors that contribute to errors of a modular system.

The impact of speech enhancement errors on automatic speech recognition (ASR) has been studied in [9, 10]. Similarly, separation artifacts can affect the downstream performance [11]. However, the thorough analysis of leakage in speech separation systems and its impact on subsequent ASR has received little attention. With leakage, we refer to situations where a separation output channel should be silent but instead contains audio that was either partially moved or copied there from the other channel. Common metrics that are computed purely on the transcription using Levenshtein distances (i.e., word error rate (WER) and its variants) may blur leakage effects as they do not guarantee temporal locality. On the other hand, signal-level metrics may struggle to handle silence appropriately and may penalize properties that do not degrade the final WER [9]. These challenges can be addressed using Hamming distances of frame-wise word-level alignments [12, 13].

We recently proposed a framework to detect leakage in a modular meeting transcription system using frame-wise word-level alignments [14]. This framework is applied and extended here to have a more detailed look at different types of leakage that may occur. In particular, we enable the measurement of leakage in situations that were not covered in our previous analysis. Furthermore, we investigate which specific error types contribute

to the significant degradation caused by an imperfect segmentation that was observed in [14].

The extended framework is applied to analyze a strong modular meeting transcription system that follows the continuous speech separation (CSS) idea [5, 15] and utilizes TF-GridNet [6, 16] for speech separation. We extend our system with a diarization module that utilizes ASR transcriptions to refine the segmentation [17] and demonstrate that this improves performance even for measures that do not penalize speaker attribution errors. In addition to the original diarization result, we replace the Whisper ASR [18] with our own system in the diarization pipeline to remove external dependencies that apply large models and require extensive amounts of data and compute. The results demonstrate a competitive performance on LibriCSS and outperform other systems that only train on LibriSpeech data.

The main contributions of this work are

- the extension of a leakage analysis framework that allows measuring significant leakage to the cross channel in areas where only the primary speaker is active,
- a detailed breakdown of segmentation error types and their contribution to the gap to the oracle segmentation performance,
- state-of-the-art results for single-microphone meeting transcription on the LibriCSS task among systems that train the ASR module on LibriSpeech data only.

2 Meeting Transcription

Speech separation, recognition and diarization modules can be composed in different orders to form a meeting transcription pipeline [19]. This section describes the pipeline used in this work, which is similar to [17]. This pipeline can shortly be described as CSS followed by segmentation, ASR, and diarization and is thus referred to as CSS-AD. Figure 1 depicts an overview.

2.1 Continuous Speech Separation

The CSS idea [5, 15] allows speaker separation for an arbitrary number of speakers. A separation module separates the observed signal into two overlap-free signals within a small window. This is possible under the assumption that the window size is small enough to ensure that at maximum two speakers are active within a single window. A given speaker might be assigned to different output channels by the separator when shifting the window. This is referred to as the permutation problem. It is tackled by enabling a small overlap between neighboring windows and choosing the permutation of adjacent segments that results in a minimal mean squared error (MSE) on the overlapping parts.

2.2 Segmentation

The separated audio channels are then segmented with a voice activity detection (VAD) module that extracts the speech regions. We use an energy-based VAD that uses the ratio of the energies across both separated channels to identify active speech. Using both channels is more robust against leakage than the classical approach of only looking at one channel to find the speech activity.

Note that in principle, the two separation output channels are equal in the sense that there is no notion of primary and cross channel. However, as soon as we consider a given segmentation,



Figure 1: Meeting transcription pipeline. A continuous speech separation (CSS) system separates overlapping speech of multiple speakers into two overlap-free channels. A voice activity detection (VAD) identifies regions with speech activity. Automatic speech recognition (ASR) transcribes each detected segment individually. The diarization assigns speaker labels and refines the segmentation using ASR information. ASR can again transcribe the resulting segments. Different colors represent different ground-truth speakers.

the speech segments are not positioned at the same times in both channels. For a given segment, we refer to the channel that contains the separated speech of that segment as the primary channel (e.g. the channel with index 1). The cross channel (in that case the channel with index 2) may contain silence or another speaker that overlapped in the original mixture observation. This is important to keep in mind for the analysis in Section 3.1.

2.3 Recognition

The segments containing separated single-speaker speech can be recognized using ASR. It outputs the transcription of what has been spoken in the given segment. While different architectures are conceivable, we use a hybrid hidden Markov model as in [14]. It has the advantage of providing accurate word-level timestamps.

2.4 Diarization

We use the ASR-supported diarization approach from [17] with slight modifications. It utilizes the word-level timestamps from the ASR to further detect speaker changes that were undetected by VAD alone. The segmentation returned by the diarization module can thus be different from the segmentation of the VAD. Note that it is possible to rerun the ASR on the refined segmentation of the CSS-AD output and this is the approach we adopt in this work.

3 Methods

3.1 Leakage Analysis

In [14], we measured the effect of leakage in order to understand how separation errors influence the ASR performance. For this analysis, we used coincidence rates (CRs), which measure the fraction of frames where the word-level alignments of both channels match. These matches are an indication of a possible spill-over of one channel into the other. We distinguish between word CR (WCR) on single-best hypotheses and graph CR (GCR) on lattices in analogy to word and graph error rates. Note that the GCR is optimistic because it checks per time frame whether the word labels match for any arc in the lattice irrespective if the considered arcs compose a valid path. An example for the GCR computation is depicted in Figure 2.

As explained in Section 2.2, the notion of a primary and a cross channel arises when considering a segmentation of the data. For a given segment, the primary channel is the channel containing the speech (e.g. channel 1) while the cross channel (in that example channel 2) may contain parts of another speech segment or silence. We still consider all segments from all channels. As a consequence, channel 2 in this example is considered as the primary channel for segments located on that channel.

We showed that the cross channel is well suppressed and that words leaked from cross-talkers into the primary channel hardly play a role in the primary channel’s search space [14]. This means that for a given segment, the ASR hypotheses rarely contain words from the cross channel’s forced aligned ground-truth transcription. However, the motivation was to study only the direct impact on the recognition performance to assess the feasibility of utilizing cross-speaker transcriptions in sequence discriminative training. Thus, only leakage from the cross-talker onto the primary channel was analyzed within the boundaries of the oracle segments. An example is depicted in Figure 3 on the left for the green segment, where possibly leaked words from the red speaker into the green segment ("dolor", "sit" or "amet") would have been measured.

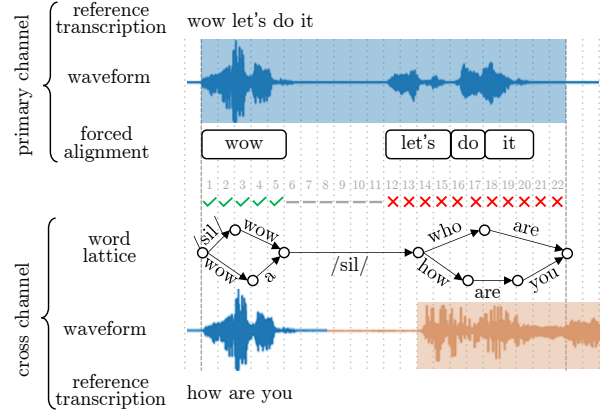


Figure 2: Visualization of GCR computation measured between the primary channel forced alignment and the cross channel word lattice. We indicate whether the word labels match for any arc (✓), mismatch (✗) or both channels’ forced alignments contain silence (—). The example results in $GCR = 5/16 \approx 31\%$. The vertical dotted lines represent frame boundaries, the dashed lines indicate boundaries of the blue segment on the primary channel. Frames are not drawn to scale. Modified from [14].

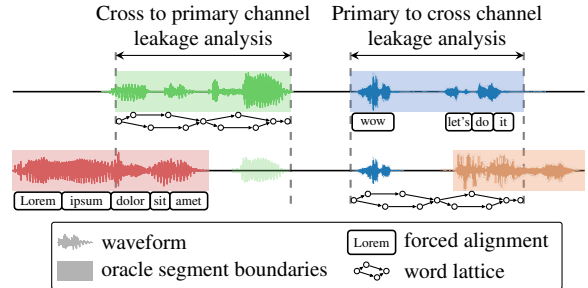


Figure 3: Illustration of the leakage analysis from the cross channel to the primary channel (left) and vice versa (right) on the CSS output streams. The analysis is illustrated for two (green and blue) of the four segments, where in both cases the upper stream is the primary channel and the bottom one the cross channel. For the red and orange segment, the roles would be swapped.

In this work, we extend the analysis by measuring leakage from the primary channel to the cross channel, which includes regions where no speech is active on the cross channel. These regions are typically outside the segments and were therefore previously ignored. Furthermore, leakage from the primary channel to the cross channel could not be measured before because only forced alignments on the cross channel were considered that could not contain leaked words which are not in the transcription. An example is given by the red transcription that does not contain the green leak in Figure 3. Now, we also compare forced alignments of the primary speaker to hypotheses on the cross channel as shown in Figure 3 on the right for the blue segment. When recognizing the cross channel signal, the blue leak is then likely to be present in the resulting hypothesis. The GCR computation for this example is shown by Figure 2.

Beyond the above extension, a few general improvements are applied. Previously, we used forced alignments obtained on the separated LibriCSS signals. However, this may result in problematic alignments where the separation creates artifacts, which are

Diarization	ASR	Data			cpWER [%]	
		Diarization	Separation	ASR		
CSS-AD (Whisper) [17]	WavLM AED	VoxCeleb	LS960	LS960 + Mix94k	6.2	
TS-SEP [20]		LS960 + VoxCeleb			6.4	
DCF-DS [21]	Transformer AED	VoxCeleb + LS960 + NSF + NF			LS960	6.3
	WavLM AED				LS960 + Mix94k	4.4
Oracle	Ours	n/a	LS960	LS960	4.3	
CSS-AD (Whisper)		VoxCeleb			5.8	
CSS-AD (ours)					5.8	

Table 1: Comparison of cpWER on LibriCSS test with different diarizations in our setup vs. results in the literature. All results use a single microphone. Our results are with the Trafo LM. The other works use attention-based encoder decoder (AED) ASR models. LS960 denotes the LibriSpeech training data, Mix94k refers to the 94k hours WavLM pre-training data. NSF and NF denote the NOTSOFAR and Near Field datasets used for training the front-end in [21]. VoxCeleb is exclusively used to train the speaker embedding extractors across all works.

the most interesting positions, in fact. To mitigate potential biases, we now use forced alignments obtained on the clean LibriSpeech signals and synchronize them to LibriCSS. Furthermore, the state-level minimum Bayes risk (sMBR) fine-tuned model with the best final performance is used to generate the hypotheses instead of the frame-wise cross-entropy (f-CE) model. Finally, we now apply the same regular search settings for lattices as for the 1-best case instead of settings targeted for sMBR training.

3.2 Segmentation Analysis

Motivated by the large performance gap of our system with VAD segmentation to the oracle segmentation in [14], we aim to study the errors that occur during segmentation. We used the visualization tool from the MeetEval toolkit [22] to find typical problems. For each error type, we define a heuristic that exploits oracle information to eliminate these errors from the segmentation. The WER obtained on the refined segmentation is then compared to the WER on the original segmentation to assess the impact of each error type.

4 Experimental Setup

4.1 Data

We evaluate our models on the LibriCSS dataset [5]. It features re-recordings of utterances from the LibriSpeech [23] *test-clean* set in meeting rooms with varying degrees of overlap. LibriCSS is well suited as an evaluation task for this work because of its meeting-like structure, real-world acoustic conditions and because of its widespread use in the research field [4, 5, 7].

In this work, we address single-microphone meeting transcription. We therefore only use the first microphone of the LibriCSS data. Furthermore, *Session0* is used as a dev set to tune hyperparameters as suggested in [19], and we report the results on the remaining sessions.

Since LibriCSS only provides data for evaluation, we use the LibriSpeech signals to simulate spatialized and mixed training data similar to [24]. The speech separators are trained on this data. The ASR model is first trained on clean LibriSpeech and then fine-tuned on the signals that are obtained by applying the separator to the simulated data. For more details, see [14]. The LibriSpeech text corpus is used to train the language models (LMs) [25].

4.2 Meeting Transcription

In this work, we use the meeting transcription system from [14] and extend it with a diarization module. The pipeline is outlined in Section 2.

TF-GridNet [6, 16] is used to separate the observed signal into two overlap-free signals within a sliding window with a size of 4 s and a shift of 3 s as described in Section 2.1. For segmentation, the baseline is a simple energy-based VAD [14]. In addition, we evaluate the refined segmentation obtained by the different diarization systems.

ASR is performed using a hybrid hidden Markov model. The neural encoder consists of 12 conformer blocks [26] and has a

total of 87M parameters. We use the sMBR fine-tuned model with baseline encoder from [14]. Further details are outlined in [14]. During recognition, we use the official LibriSpeech 4gram LM as well as a neural transformer-based LM (Trafo LM).

The ASR-supported diarization pipeline builds on top of the word-level timestamps from ASR [17]. For every word boundary, one speaker embedding vector is computed each for the left and right context of 3 s around the boundary. If the similarity between the two is below a threshold and the lowest among the context of 4 s, it is considered as a speaker change, and the segment is split. Afterwards, one speaker embedding vector is extracted for every segment and a k-means clustering is applied to obtain speaker labels. Note that the initial VAD hyperparameters are selected differently here to create shorter segments and avoid segments that contain speaker changes.

Finally, we pass the refined segments again to ASR to obtain more accurate transcriptions. Here, we additionally merge subsequent segments if the diarization assigns them to the same speaker and the pause between the segments is not longer than 3 s to have more context within the segments if possible.

5 Results

Table 1 presents the concatenated minimum-permutation WER (cpWER) of our meeting transcription system on the LibriCSS task. Our results are competitive with existing works and outperform the systems in [17, 20] that both use a large pre-trained WavLM¹ [27] model for ASR. The cpWER for both of our diarizations is identical and constitutes a new state-of-the-art performance for systems that only use LibriSpeech data for ASR training. Solely the DCF-DS system in combination with WavLM obtains a better single-microphone cpWER on LibriCSS. Notably, our results are achieved without external dependencies that apply large models and require extensive amounts of data or compute. Only the speaker embedding extractor was trained on data other than LibriSpeech, namely VoxCeleb.

To compare our systems to the baseline VAD segmentation in [14], Table 2 reports the optimal reference combination WER (ORC WER) [3]. Unlike the cpWER, it does not account for speaker attribution errors and is therefore generally lower. The oracle segmentation is obtained by using the boundaries of the original LibriSpeech utterance provided by the LibriCSS annotation and selecting the separated channel with minimum signal-to-distortion ratio (SDR) [28] to the clean audio. Table 2 compares the oracle and VAD results to the refined segmentations from diarization based on either Whisper’s or our transcription are tested. In specific, Whisper [18] was deployed in the “large-v2” configuration. Note that the hyperparameters for the preceding VAD are tuned individually for the latter two cases. We can observe clear improvements compared to [14], closing around a third of the previous gap to the oracle performance.

¹https://huggingface.co/espnet/simpleoier_librispeech_asr_train_asr_conformer7_wavlm_large_raw_en_bpe5000_sp

Segmentation	ORC WER [%]	
	4gram LM	Trafo LM
Oracle	5.8	4.3
VAD [14]	7.0	5.6
CSS-AD (Whisper)	6.7	5.4
CSS-AD (ours)	6.5	5.2

Table 2: ORC WER on LibriCSS test with different segmentations.

5.1 Leakage Analysis

Table 3 presents the results of the leakage analysis. We report CRs which describe the fraction of frames in which both sequences contain the same word. The first line (spoken vs. spoken) represents the natural coincidence, i.e., how often both speakers really uttered the same word at the same time. Leakage from the cross channel to the primary channel is considered in the next part of the table (1-best/lattice vs. spoken), similarly to [14]. The results do not deviate much despite our updates in the analysis. The CRs for the primary channel hypotheses (1-best and lattice) with the cross channel ground truth (spoken) are higher than the natural coincidence, but this is mainly caused by silence. We observe higher CRs for words in regions with two active speakers and for silence generally compared to [14]. However, the overall conclusion that the cross speaker is well suppressed and does not have a major impact on the primary channel’s search space still holds.

Finally, leakage from the primary channel to the cross channel is addressed (spoken vs. 1-best/lattice). Significant leakage can be observed for this direction in areas where only one speaker is active. This can likely be explained by the nature of the segmentation. In a given segment, there are few positions where the cross speaker is active and the primary speaker is not because the segment is targeted to the primary speaker. In contrast, there are many frames where only the primary speaker is active and therefore more chances to create leakage from the primary channel to the cross channel.

This could be a hint why there is such a significant performance gap to the oracle segmentation. If leakage of this type occurs, the VAD could create a segment for the leaked speech which would be transcribed and cause edits in the WER. The oracle segmentation automatically discards these leaks. If this is really a major cause of recognition errors, will be investigated in the next section.

5.2 Segmentation Analysis

By manual inspection of the MeetEval visualizations, we identified the following segmentation error types:

1. The cross channel should be silent but audio from the primary channel leaks through, either by partially moving or by copying the audio to the cross channel ("leakage"). This is the error type analyzed in the previous section.
2. Some parts of segments are missing, removing relevant speech contents ("missing").
3. The segmentation creates long segments that merge several oracle segments into one ("merges").
4. Even in the case of clear correspondence of VAD and oracle segments, the boundary times typically deviate slightly ("boundaries").

Table 4 shows the impact of running ASR on a segmentation that eliminates these different error types using oracle information. Unlike the results in Section 5.1 might suggest, removing leaked segments does not result in a major improvement. Our best ORC WER even remains unchanged. Similarly, splitting of segments that consist of multiple oracle segments and adjusting the boundary times does not have a significant effect on the performance either. In this case, this can be considered as expected. Splitting segments according to the oracle segmentation mainly leads to a smaller context size for the LM which might hurt or not depending on how related the individual segments are. Note that in some cases, the subsequent segments are subsequent segments in LibriSpeech *test-clean* suggesting that the longer context could even be useful.

Hypothesis		Coincidence rate [%]					
		Words and sil.			Words only		
		#act. speakers			#act. speakers		
		1	2	Avg.	1	2	Avg.
Spoken	Spoken	0.0	0.2	0.0	0.0	0.2	0.0
1-best	Spoken	3.1	0.6	2.7	0.0	0.6	0.1
Lattice		4.8	0.9	4.1	0.1	0.9	0.2
Spoken	1-best	8.5	0.7	7.1	8.2	0.7	6.9
	Lattice	15.2	1.3	12.8	14.7	1.3	12.5

Table 3: Analysis of leakage between separated channels. Coincidences are counted once for both words and silence, once for matching words only. The hypotheses (lattices and 1-best) are obtained with the 4gram LM. "Spoken" refers to the forced alignment of the ground truth transcription. Results on LibriCSS test.

Transcript for diarization	Error types eliminated	ORC WER [%]	
		4gram LM	Trafo LM
Whisper	-	6.7	5.4
	Leaks	6.5	5.2
	Missing	6.1	4.8
	Merges	6.7	5.3
	Boundaries	6.7	5.3
	All	5.9	4.4
Ours	-	6.5	5.2
	Leaks	6.4	5.2
	Missing	6.0	4.7
	Merges	6.5	5.2
	Boundaries	6.5	5.1
	All	6.0	4.5

Table 4: Effect of different segmentation error types on ORC WER evaluated based on their elimination using oracle-informed heuristics. Results on LibriCSS test.

Adapting the boundaries should mainly only affect silence at the beginning and end of segments which is not expected to have a strong impact on the recognized hypothesis.

However, adding missing segments significantly improves the ORC WER. This error type accounts for more than half of the performance gap to the oracle segmentation. In combination, fixing the observed error types allows closing around 90% of the gap for the Whisper-informed diarization and around 75% for our diarization. An absolute difference of only 0.1% or 0.2% respectively remains. Future work might therefore concern improved VAD to address these errors.

6 Conclusion

This work studies remaining errors in a strong modular meeting transcription system. For this, we extend an existing leakage analysis framework with proper sensitivity to temporal locality and show that significant leakage to the cross channel can be measured in regions where only the primary speaker is active. After identifying typical segmentation error types, we evaluate their effect on the performance gap to the oracle segmentation. The results show that missing speech segments are the main contributor and that the identified leakage is not a major problem. By adding advanced diarization systems, we close around a third of the gap to the oracle segmentation and achieve state-of-the-art single-microphone results on the LibriCSS task among systems that only use LibriSpeech data for ASR training.

7 Acknowledgement

This research was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project No. 448568305. Computational Resources were provided by BMBF/ NHR/ PC2.

References

- [1] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, "Continuous speech separation with conformer," in *Proc. ICASSP*, 2021, pp. 5749–5753.
- [2] S. Berger, P. Vieting, C. Boeddeker, R. Schlüter, and R. Haeb-Umbach, "Mixture encoder for joint speech separation and recognition," in *Proc. Interspeech*, Dublin, Ireland, Aug. 2023, pp. 3527–3531.
- [3] I. Sklyar, A. Piunova, X. Zheng, and Y. Liu, "Multi-turn rnn-t for streaming recognition of multi-party speech," in *Proc. ICASSP*, Singapore, May 2022, pp. 8402–8406.
- [4] N. Kanda, J. Wu, Y. Wu, X. Xiao, Z. Meng, X. Wang, Y. Gaur, Z. Chen, J. Li, and T. Yoshioka, "Streaming multi-talker ASR with token-level serialized output training," Preprint arXiv:2202.00842, 2022.
- [5] 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 *Proc. ICASSP*, Barcelona, Spain, May 2020, pp. 7284–7288.
- [6] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.
- [7] C. Boeddeker, A. S. Subramanian, G. Wichern, R. Haeb-Umbach, and J. L. Roux, "TS-SEP: Joint diarization and separation conditioned on estimated speaker embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1185–1197, 2024.
- [8] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng, Z. Du, W. Huang, P. Guo, Z. Yan, B. Ma, X. Xu, and H. Bu, "M2Met: The ICASSP 2022 multi-channel multi-party meeting transcription challenge," in *Proc. ICASSP*, Singapore, May 2022, pp. 6167–6171.
- [9] K. Iwamoto, T. Ochiai, M. Delcroix, R. Ikeshita, H. Sato, S. Araki, and S. Katagiri, "How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR," in *Proc. Interspeech*, Incheon, Korea, Sep. 2022, pp. 5418–5422.
- [10] S. Araki, A. Yamamoto, T. Ochiai, K. Arai, A. Ogawa, T. Nakatani, and T. Irino, "Impact of residual noise and artifacts in speech enhancement errors on intelligibility of human and machine," in *Proc. Interspeech*, Dublin, Ireland, Aug. 2023, pp. 2503–2507.
- [11] T. Cord-Landwehr, C. Boeddeker, T. Von Neumann, C. Zorilă, R. Doddipatla, and R. Haeb-Umbach, "Monaural source separation: From anechoic to reverberant environments," in *Proc. IWAENC*, Bamberg, Germany, 2022.
- [12] F. Wessel, R. Schlüter, and H. Ney, "Explicit word error minimization using word hypothesis posterior probabilities," in *Proc. ICASSP*, Salt Lake City, UT, USA, May 2001, pp. 33–36.
- [13] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 288–298, Mar. 2001.
- [14] P. Vieting, S. Berger, T. v. Neumann, C. Boeddeker, R. Schlüter, and R. Haeb-Umbach, "Combining TF-GridNet and mixture encoder for continuous speech separation for meeting transcription," in *Proc. SLT*, Macao, China, Dec. 2024, pp. 160–167.
- [15] T. Yoshioka, H. Erdogan, Z. Chen, X. Xiao, and F. Alleva, "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 3038–3042.
- [16] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Making time-frequency domain models great again for monaural speaker separation," in *Proc. ICASSP*, Rhodes, Greece, Jun. 2023.
- [17] T. von Neumann, C. Boeddeker, T. Cord-Landwehr, M. Delcroix, and R. Haeb-Umbach, "Meeting recognition with continuous speech separation and transcription-supported diarization," in *Proc. ICASSP Workshop on Hands-free Speech Communications and Microphone Arrays*, Seoul, Korea, 2024, pp. 775–779.
- [18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*. Waikoloa, HI, USA: PMLR, Jul. 2023, pp. 28 492–28 518.
- [19] D. Raj, P. Denisov, Z. Chen, H. Erdogan, Z. Huang, M. He, S. Watanabe, J. Du, T. Yoshioka, Y. Luo, N. Kanda, J. Li, S. Wisdom, and J. R. Hershey, "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *Proc. SLT*, Shenzhen, China, Jan. 2021, pp. 897–904.
- [20] C. Boeddeker, T. Cord-Landwehr, and R. Haeb-Umbach, "Once more diarization: Improving meeting transcription systems through segment-level speaker reassignment," in *Proc. Interspeech*, 2024, pp. 1615–1619.
- [21] S.-T. Niu, J. Du, R.-Y. Wang, G.-B. Yang, T. Gao, J. Pan, and Y. Hu, "DCF-DS: Deep cascade fusion of diarization and separation for speech recognition under realistic single-channel conditions," Preprint arXiv:2411.06667, 2024.
- [22] T. von Neumann, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach, "MeetEval: A toolkit for computation of word error rates for meeting transcription systems," in *Proc. CHiME Workshop*, Dublin, Ireland, 2023, pp. 27–32.
- [23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 5206–5210.
- [24] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," Preprint arXiv:1910.13934, 2019.
- [25] K. Irie, A. Zeyer, R. Schlüter, and H. Ney, "Language modeling with deep transformers," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 3905–3909.
- [26] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 5036–5040.
- [27] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [28] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.