ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446589

SYNTHETIC CONVERSATIONS IMPROVE MULTI-TALKER ASR

Thai-Binh Nguyen[1], *Alexander Waibel*[1,2]

[1]Karlsruhe Institute of Technology
[2]Carnegie Mellon University
thai-binh.nguyen@kit.edu

ABSTRACT

In recent times, automatic speech recognition (ASR) has seen remarkable progress, particularly in recognizing dominant speakers. Nevertheless, the challenge of multi-talker scenarios involving distinguishing between speakers and transcribing their speech accurately remains unsolved due to limited data constraining model effectiveness. In this study, We propose a novel methodology called Systematic Synthetic Conversations (SSC), which leverages conventional ASR datasets to help an end-to-end (E2E) multi-talker ASR model establish new state-of-the-art results across synthetic and authentic multi-talker datasets. Notably, we achieved a 3.47% word error rate (WER) for the Libri2Mix [1] set, and WERs of 13.96% and 19.51% for the AMI-IHM and AMI-SDM [2] sets, respectively. These outcomes underscore the hidden potential of existing resources in tackling the complicated multi-talker problems within the domain of ASR.

Index Terms— multi-talker, asr, synthetic conversation

1. INTRODUCTION

Multi-talker speech recognition is an emerging research in the speech community due to its great potential in applications such as conversation and meeting transcriptions. The issue here is that the sounds we're dealing with are complicated. There are multiple people talking, sometimes all at once, and this can be mixed with background noise and reverberation. In this complex acoustic condition, a recognition system must differentiate persons and transcribe their utterances. Many studies proposed to handle this task. Figure 1 reveals a few main approaches.

The initial approach, as depicted in Figure 1a, involves a straightforward separation of mixture signals into multiple channels, with each channel corresponding to a distinct speaker (referred to as the speech separation model). Following this, a conventional ASR model is applied to transcribe the signals [3, 4]. This method yields discernible speech and text outputs, each attributed to a specific speaker. Nevertheless, the training data for the separation model is usually synthesized data. Furthermore, the training of the speech separation system often relies on a signal-level criterion, which might not inherently align with the optimal criteria for ASR performance.

Instead of explicitly isolating the speaker at the signal level, certain studies achieve this within hidden layers by employing a speaker modeling model (refer to figure 1b). In [5], authors use additional information (speaker embedding) of the person involved in the input signal. However, relying on a fixed speaker embedding model can potentially bottleneck the system when encountering unknown acoustic conditions. Alternatively, in [6], speakers are implicitly separated within the mixed speech embedding, avoiding the need for speaker information that might not be available in practical scenarios. With PIT [7], a joint multi-talker inference using permutation invariant training was also considered for this strategy. In essence, this method resembles E2E multi-talker ASR, which mitigates artifacts in the output signal that may be introduced like in the first approach(figure 1a). Nonetheless, gathering data for this method remains challenging, leading most studies to conduct experiments solely on limited or synthetic datasets.

We can consider using a speaker diarization model to detect "who speaks when" and then do ASR after that (figure 1c). This method [8, 9], however, could suffer from accuracy degradation in overlapped regions because the ASR system is usually designed to recognize single-speaker speech. In addition, data used to train the diarization model is also expensive to annotate. [10] show that taking roughly 2 hours for a single annotator to annotate 10 minutes of audio for one pass.

The approaches in figure 1(a, b, c) share a common trait: they yield explicit outputs that identify speakers and their dialogues. However, their training relies on limited real or simulated data. Suppose we ignore the need to identify the speaker, an alternate approach shown in figure 1d, an E2E solution using a single ASR model is feasible. This strategy employs a sequence-to-sequence model to produce token sequences, with special tokens marking speaker changes in transcripts. In previous work [11], specialized tokens (<spk:dr> and <spk:pt>) were inserted into ASR transcripts to distinguish doctor and patient speech. However, such a system cannot be used for generic purposes. In [12], authors employed 75K hours of single-talker data to simulate 900K hours of multi-talker audio for pretraining. Both [11, 12] use private large-scale data, which has been considered a limitation.

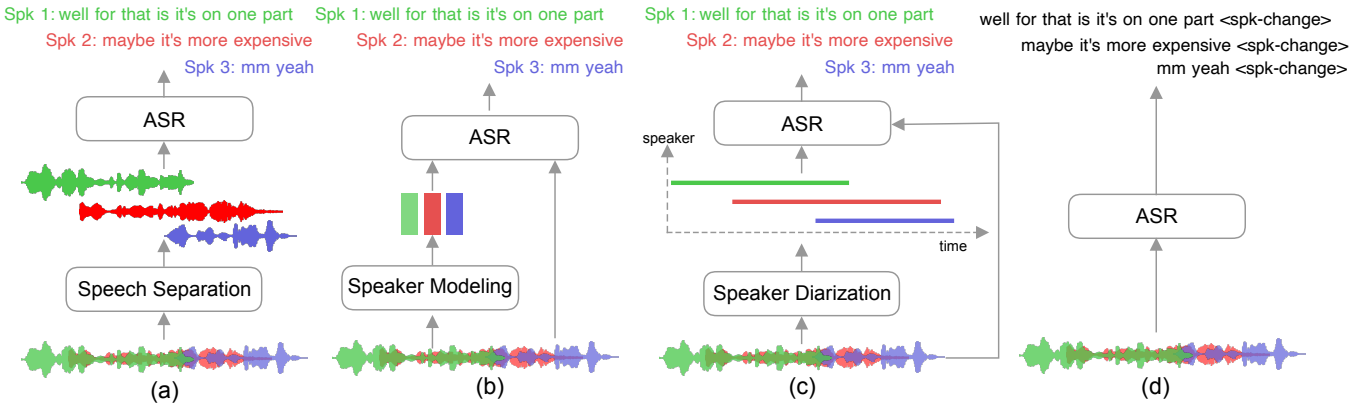

Fig. 1. ASR multi-talker approaches

Our study introduces an innovative method for leveraging existing public ASR data. We establish a comprehensive dialog dataset by employing label force alignment and conversation audio reconstruction. We named this process Systematic Synthetic Conversations (SSC). Our focus lies in the E2E approach for multi-talker transcription generation. Surpassing prior state-of-the-art work [12], our results excel in AMI-SDM and AMI-IHM benchmarks, achieving this with a dataset smaller than five times their size. Moreover, we establish a new state-of-the-art benchmark in the Libri2Mix dataset. We openly share our label alignment to foster further research, facilitating engagement with this challenge. `hf.co/datasets/nguyenvulebinh/asr-alignment`.

2. SYSTEMATIC SYNTHETIC CONVERSATIONS

Conversation is a concept when multiple people express their thoughts, ideas, and opinions back and forth. Intuitively, we can combine utterances from multiple people with a relative meaning to construct a conversation. In the AMI meeting corpus, if we segment the recording at silence positions or non-overlapping utterance boundaries, around 60% of the time, a segment contains a single speaker, 25% and 10% involve two and three speakers, respectively. Inside a segment that has multi-speakers, around 20.5% of the time is overlapping speech.

Based on that intuition, here we want to reconstruct a conversation to make data for training the ASR multi-talker model. To do that, we need a set of utterances that contains only a single speaker. Many datasets fit that requirement, like Librispeech [13], MuST-C [14], TED-LIUM [15], VoxPopuli [16], and Common Voice [17]. However, most are the corpus of read speech, so the diversity is not strong enough. A dataset from YouTube like GigaSpeech [18] is more diverse, but it can contain multiple speakers. We take a random 50.000 utterances from GigaSpeech and do the speaker diarization [19], which shows that 29.8% of it contains more than one speaker. Although containing multi-speaker, the overlapping speech portion is small, just around 0.5%.

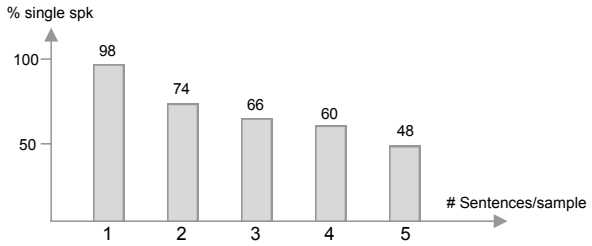

Fig. 2. Correlation between the number of sentences in an utterance in GigaSpeech and the proportion of single speaker.

We explored various approaches for extracting segments with just one speaker from the GigaSpeech dataset. Initially, we considered employing speaker diarization to isolate individual speakers' text. However, due to a reported Diarization Error Rate of 24.3% on average [19], we found this method to be less dependable. An alternative strategy we pondered involved segmenting based on the semantic structure of the transcriptions, achieved through sentence segmentation. In practice, breaking a sequence into sentences can often be guided by punctuation. Fortunately, GigaSpeech's transcriptions include punctuation marks. Even datasets lacking punctuation can still benefit, as research [20] has made strides in information recovery techniques. For datasets without punctuation, we turned to a pre-trained model from Nvidia Nemo framework [21] to restore these cues. The relationship between sentence count and the prevalence of single-speaker instances in the GigaSpeech dataset is depicted in Figure 2. Notably, when an utterance comprises a solitary sentence, it is highly likely to feature only one speaker.

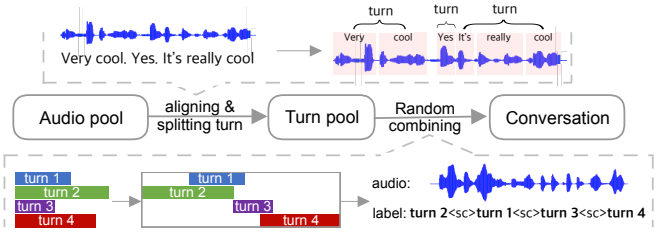

Fig. 3. Overview of Systematic Synthetic Conversations

Dataset	Size (samples/hours)	# Turns
GigaSpeech	8283K / 10K	9680K
Common Voice	945K / 1.7K	955K
MuST-C	248K / 0.5K	304k
TED-LIUM	268K / 0.5K	370K
VoxPopuli	182K / 0.5K	218K
LibriSpeech	281K / 0.9K	551K

Table 1. List of pretraining dataset

To align the textual content and audio, we leveraged the open-source MMS developed by Facebook [22]. This has been trained using the wav2vec2 CTC model, encompassing 31K hours covering 1,130 languages. Following the alignment, we re-evaluated the number of speakers in each audio fragment (segmented by sentence). The percentage of single-speaker now is 97%. This confirms the efficacy of text-based alignment and sentence segmentation for isolating individual speaker utterances from diverse audio sources.

Our SSC approach is depicted in Figure 3. The audio pool includes various datasets (see Table 1). GigaSpeech accommodates multiple speakers, while other datasets feature a solitary speaker per sample. In total, this covers 14K hours. Each sample undergoes alignment using the MMS, followed by sentence tokenizer, facilitated by Spacy Sentencizer[23]. These segmented units, referred to as "turns", are anticipated to emanate from individual speakers. This aligning and splitting turn procedure is executed as a preliminary step preceding the actual training process. Post-segmentation, a total of 12 million turns is obtained.

During the random combination phase, we pick a maximum of n turns from the turn pool and arrange them randomly along the time axis. This arrangement maintains an approximate 20% overlap rate, inspired by insights from the AMI dataset. To avoid excessive length, we enforce a strict 20-second cap on the total combined audio duration. A turn from the n turns is included only if its inclusion doesn't result in the total length exceeding this duration. This dynamic combination takes place during model training. Labels are generated by concatenating the transcriptions of the turns, with a unique $\langle sc \rangle$ token (indicating speaker change) inserted between consecutive turns. The order of transcriptions aligns with the appearance sequence of the corresponding signals within the mixture audio.

3. EXPERIMENTS SETUP

3.1. Modeling

The base model we use to benchmark is a typical ASR sequence-to-sequence model. The encoder is WavLM$_{large}$ [16], and the decoder is BART-decoder$_{base}$ [24]. Formally, given an input sequence $X \in \mathbb{R}^l$ (l is the length of the signal), the goal of the ASR model is to estimate transcription $Y = (y_t \in \{1, ..., |\mathcal{V}|\} | t = 1, ..., T)$ where $|\mathcal{V}|$ is the size of the vocabulary $\mathcal{V}$, and T is the number of estimated tokens. When we deal with multi-talker, the output sequence should be $Y = \{y_1^1, .., y_{T^1}^1, \langle sc \rangle, y_1^2, .., y_{T^2}^2, \langle sc \rangle, y_1^3, .., y_{T^3}^3, \langle eos \rangle\}$, where y_i^j represents the i-th token of the j-th speaker. Here, $\langle eos \rangle$, a token for sequence end, is used only at the end of the entire sequence.

The model advances using Speech-Encoder-Decoder from the HuggingFace library [25] through two training phases. Firstly, the pre-training phase employs data from the SSC (refer to figure 2). This phase spans 300K training steps with a batch size of 120. Following this, the second phase entails fine-tuning with a standard multi-talker ASR dataset for roughly ten epochs.

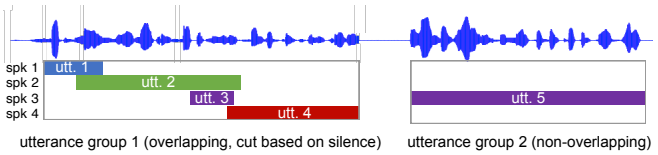

Fig. 4. Utterance group-based

3.2. Dataset

For fine-tuning and evaluation, we employ LibriMix (synthetic) and AMI Meeting Corpus (real multi-talker dataset). LibriMix uses the "2-spk 16kHz max" setup (denoted as Libri2Mix), where two random Lirispeech samples are mixed, testing the model in fully overlapping scenarios. With the AMI dataset, we follow [12], adopting an utterance group-based approach. These groups can contain up to 4 speakers (see example in figure 4), segmented from original audio based on silence or non-overlapping utterance boundaries.

The Libri2Mix training dataset encompasses 58 hours of speech mixtures from LibriSpeech train-clean-100, with 11 hours each for the dev and test sets. The AMI benchmark exists in two versions: AMI-IHM records meetings via closed microphones, while AMI-SDM employs distant microphones. Using the Kaldi toolkit's scripts [26], we allocate 76.9 hours for training and around 8.9 hours for both the dev and test sets (applicable to both AMI-IHM and AMI-SDM).

3.3. Evaluation metric

During training, the sequence of each speaker's transcription aligns with their appearance time. However, during inference, speaker order may be randomized, lacking specific constraints. We focus solely on completing content for each speaker, not the order. Similar to other studies [12, 29, 7], we employ the concatenated minimum permutation word error rate metric. This involves concatenating all speakers' reference transcriptions, generating permutation lists for hypotheses, and calculating WER for each permutation against

Table 2. Comparison WER of different systems on Libri2Mix

Model	dev	test
JSM + WavLM-CTC [5]	11.04	10.68
Conv-TasNet + Transformer [27]	21.00	21.9
PIT + Transformer [27]	26.58	26.55
Wav2vec2-Sidecar [6]	7.68	8.12
Whisper (large)	49.43	49.99
WavLM-BART (ours)	**3.36**	**3.47**

Table 3. Comparison WER of different systems on AMI (utterance group)

Model	IHM	SDM
JSM + WavLM-CTC [5]	25.88	-
PIT + WavLM [5]	27.02	-
Conformer [12]	14.9	21.2
Multi-talker Whisper (large) [28]	-	21.4
Whisper (large)	34.21	50.71
WavLM-BART (ours)	**13.96**	**19.51**

Table 4. Impact of Data Processing Efficiency on WER in AMI-IHM

WavLM-BART	WER	Differences
(0) finetune	42.64	-28,68
(1) utterance data	33.69	-19.73
(1) + finetune	25.72	-11,76
(2) conversation data	26.31	-12.35
(2) + finetune	13.96	0

the reference, selecting the lowest WER. For simplicity, we'll refer to this as normal WER going forward.

4. EXPERIMENTS RESULT

Table 2 presents benchmark results obtained from various models evaluated on the Libri2Mix dataset. All models (except Whisper) were pre-trained on a bigger dataset (including full LibriSpeech) but finetuned only on Libri2Mix with the train-clean-100 set. In this context, JSM refers to the Joint Speaker Modeling, which combines speaker embedding with the WavLM-CTC ASR model. The Conv-TasNet model is a dedicated speech separation architecture, while PIT involves multi-talker inference through permutation invariant training. The Wav2vec2-Sidecar method involves the separation of mixed speech embeddings and subsequent transcription for distinct speakers. Whisper [30] here to demonstrate the limitations of a robust ASR model (trained extensively with 680k hours of audio) which can only handle the dominant speaker. During our study period, the most recent state-of-the-art (SOTA) achievement within the Libri2Mix dataset was attributed to the Wav2vec2-Sidecar model. Our model WavLM-BART shows outstanding result and beat other models by a large margin.

AMI meeting benchmark is shown in table 3. Along with JSM and PIT models, we add a few more architectures to compare. Firstly, a Confromer model [12] was pre-trained with 75k hours in the multi-talker style. From that, they finetuned with the AMI utterance group data and got a new SOTA (WER 14.9 and 21.2 in IHM and SDM). In [28], authors utilize the Whisper model and finetune them with the AMI utterance group (result in WER 21.4 in SDM). It helps to improve the default version of Whisper a lot (WER 50.71 in SDM). Our WavLM-BART model, which pre-trained with only 14k hours of public audio (using our proposed SSC) and fine-tuned with the AMI, outperformed both Whisper and Conformer (which have much more private data).

We do an ablation experiment (table 4) on the AMI-IHM dataset to show the efficiency of the SSC. The bottom line showcases the outcomes of our prime system, which underwent pre-training with conversation data and fine-tuning using in-domain AMI data, resulting in a WER of 13.96%. However, when pre-trained with original utterance data and subsequently fine-tuned with AMI, the WER exhibits a marked deterioration (33.69%), increasing by 11.76% in absolute terms. In cases where no pre-training data is employed, the model's performance reaches its poorest performance, with training solely on AMI data yielding a WER of 42.64% (increasing by 28.68% absolute).

Model	# of talkers				Total
	1	2	3	4	
Conformer [12]	14.7	19.6	25.7	35.5	21.2
Multi-talker Whisper [28]	12.0	20.2	29.3	40.6	21.4
WavLM-BART (ours)	12.9	20.3	29.1	30.7	19.5

Table 5. WER for AMI-SDM evaluation set w.r.t the number of speakers. All systems fine-tuned on utterance group data.

Table 5 details the pros and cons of systems based on how models were trained. In this table, we can see how good each system is depending on how many speakers exist in the sample. Whisper in [28] works best in the single-speaker setting since it is trained with much more data in that type, but the performance is down when the number of speaker increase. Our model and the Conformer model in [12] work better when a sample has multiple talkers because of those trained with dialog-orient data. Although [12] has bigger data (more than five times compared with us), we gain better performance in 1, and 4-speaker settings and competitive results in 2-speaker setting. It proves the help of our proposed SSC.

5. CONCLUSION

In summary, our novel approach effectively harnesses common public ASR datasets to enhance multi-talker end-to-end ASR systems. Our method demonstrates superior performance on the AMI and LibriMix datasets, even with considerably less data than prior studies, and we have openly shared our label alignment to foster continued research. While our current random conversation constructor is promising, future advancements could involve integrating retrieval conversation knowledge-based mechanisms for potentially superior results.

6. ACKNOWLEDGMENT

The authors gratefully acknowledge the support provided by Carl Zeiss Stiftung under the project Jung bleiben mit Robotern (P2019-01-002).

7. REFERENCES

[1] Joris Cosentino and et al., "Librimix: An open-source dataset for generalizable speech separation," 2020.

[2] Jean Carletta, Simone Ashby, and et al., "The ami meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.

[3] Zhuo Chen and et al., "Continuous speech separation: Dataset and analysis," in *ICASSP*, 2020, pp. 7284–7288.

[4] Desh Raj and et al., "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *2021 IEEE SLT Workshop*, 2021, pp. 897–904.

[5] Zili Huang, Desh Raj, and et al., "Adapting self-supervised models to multi-talker speech recognition using speaker embeddings," in *ICASSP*, 2023, pp. 1–5.

[6] Lingwei Meng, Jiawen Kang, and et al., "A sidecar separator can convert a single-talker speech recognition system to a multi-talker one," in *ICASSP*, 2023, pp. 1–5.

[7] Dong Yu and et al., "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP*, 2017, pp. 241–245.

[8] S.E. Tranter and D.A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[9] Tae Jin Park and et al., "A review of speaker diarization: Recent advances with deep learning," 2021.

[10] Wei Xia and et al., "Turn-to-diarize: Online speaker diarization constrained by transformer transducer speaker turn detection," in *ICASSP*, 2022, pp. 8077–8081.

[11] Laurent Shafey and et al., "Joint speech recognition and speaker diarization via sequence transduction," 2019.

[12] Naoyuki Kanda, Guoli Ye, and et al., "Large-scale pre-training of end-to-end multi-talker asr for meeting transcription with single distant microphone," 2021.

[13] Vassil Panayotov, Guoguo Chen, and et al., "Librispeech: An asr corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[14] Mattia A. Di Gangi and et al., "MuST-C: a Multilingual Speech Translation Corpus," in *NAACL*. 2019, ACL.

[15] Franç ois Hernandez and et al., "TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *Speech and Computer*, pp. 198–208. Springer International Publishing, 2018.

[16] Changhan Wang and et al., "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL-IJCNLP*, Online, Aug. 2021, pp. 993–1003, ACL.

[17] R. Ardila and et al., "Common voice: A massively-multilingual speech corpus," in *LREC*, 2020.

[18] Guoguo Chen, Shuzhou Chai, and et al., "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," 2021.

[19] Hervé Bredin and Antoine Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. Interspeech 2021*, 2021.

[20] Thai Binh Nguyen and et al., "Improving vietnamese named entity recognition from speech using word capitalization and punctuation recovery models," *Proc. Interspeech 2020*, pp. 4263–4267, 2020.

[21] Monica Sunkara, Srikanth Ronanki, and et al., "Multimodal Semi-Supervised Learning Framework for Punctuation Prediction in Conversational Speech," in *Proc. Interspeech 2020*, 2020, pp. 4911–4915.

[22] Vineel Pratap, Andros Tjandra, and et al., "Scaling speech technology to 1,000+ languages," 2023.

[23] Matthew Honnibal and Mark Johnson, "An improved non-monotonic transition system for dependency parsing," in *EMNLP*. Sept. 2015, pp. 1373–1378, ACL.

[24] Mike Lewis, Yinhan Liu, and et al., "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019.

[25] Thomas Wolf nd et al., "Transformers: State-of-the-art natural language processing," in *EMNLP: System Demonstrations*, Online, Oct. 2020, pp. 38–45, ACL.

[26] Daniel Povey, Arnab Ghoshal, and et al., "The kaldi speech recognition toolkit," *ASRU*, 01 2011.

[27] Song Li, Beibei Ouyang, Fuchuan Tong, Dexin Liao, Lin Li, and Qingyang Hong, "Real-Time End-to-End Monaural Multi-Speaker Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 3750–3754.

[28] Chenda Li and et al., "Adapting multi-lingual asr models for handling multiple talkers," 2023.

[29] Naoyuki Kanda and et al., "Investigation of end-to-end speaker-attributed asr for continuous multi-talker recordings," in *SLT*, 2021, pp. 809–816.

[30] Alec Radford and et al., "Robust speech recognition via large-scale weak supervision," 2022.