Adapting Multi-Lingual ASR Models for Handling Multiple Talkers

Chenda Li[1,2,†], Yao Qian[2], Zhuo Chen[2], Naoyuki Kanda[2],
Dongmei Wang[2], Takuya Yoshioka[2], Yanmin Qian[1], Michael Zeng[2]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]Microsoft, Redmond, WA, USA

[1]{lichenda1996, yanminqian}@sjtu.edu.cn,
[2]{yaoqian, zhuc, nakanda, dongmei.wang, tayoshio, nzeng}@microsoft.com

Abstract

State-of-the-art large-scale universal speech models (USMs) show a decent automatic speech recognition (ASR) performance across multiple domains and languages. However, it remains a challenge for these models to recognize overlapped speech, which is often seen in meeting conversations. We propose an approach to adapt USMs for multi-talker ASR. We first develop an enhanced version of serialized output training to jointly perform multi-talker ASR and utterance timestamp prediction. That is, we predict the ASR hypotheses for all speakers, count the speakers, and estimate the utterance timestamps at the same time. We further introduce a lightweight adapter module to maintain the multilingual property of the USMs even when we perform the adaptation with only a single language. Experimental results obtained using the AMI and AliMeeting corpora show that our proposed approach effectively transfers the USMs to a strong multilingual multi-talker ASR model with timestamp prediction capability.

Index Terms: multi-talker speech recognition, transfer learning, speaker counting

1. Introduction

Overlapping speech is omnipresent in natural speech communications [1,2]. Previous studies showed that the duration of overlapped speech during meetings could range from 6% to 15% [3]. Automatic conversation transcription systems must be able to recognize overlapping speech and distinguish individual speakers. Additionally, it is desirable to estimate the timestamps of each utterance to enhance usability.

Single-talker automatic speech recognition (ASR) systems often struggle when they are presented with multi-talker overlapped speech [4, 5]. Permutation invariant training (PIT) [6, 7] enables building ASR systems for multi-talker data [8–12]. Most of the PIT-based models have a fixed number of output branches, making them less useful for processing a long audio signal containing an unknown number of speakers. Target speech recognition [13–15] is another approach, which only transcribes the speech of a specified talker. Serialized output training (SOT) [16] was recently proposed to map the multi-talker audio including speech overlaps into a single output sequence representing the transcriptions of multiple speakers. A special token indicating a speaker change is inserted in the output sequence to distinguish the transcriptions of individual speakers. While the SOT-based ASR model can transcribe the overlapping speech of a variable number of speakers, the originally proposed scheme does not offer the ability to predict the

utterance timestamps, limiting the usefulness of the generated transcriptions. It is also worth noting that SOT has only been tested in a monolingual setting.

The timestamp prediction of ASR hypothesis was conventionally achieved by using the forced alignment algorithm based on the phoneme posterior probability [17], which was common for hybrid ASR systems. However, it cannot be directly applicable for end-to-end (E2E) ASR models [18] because the ASR model is usually not designed to produce phoneme posterior probability. It becomes further challenging when the audio contains speech overlaps, for which even a conventional forced alignment tool is no longer applicable. Only limited studies were conducted to estimate timestamps of the multi-talker ASR hypothesis, where a small neural network was introduced to estimate word timestamps [19] or special symbols indicating the start and end of each utterance were introduced [20].

In this work, we aim to solve the multi-talker ASR, speaker counting, and timestamp prediction problem with a single multilingual E2E model. Thanks to the advanced deep learning architecture (e.g. [21]) and the rapid growth of data and computing resources in recent years, large-scale pre-trained foundation models [22–27] have shown their strengths in many speech processing tasks [28], including speech separation [29, 30], speaker identification/diarization [25,28,31], and speech recognition [22–27]. This inspires us that, with proper transfer learning, a single large-scale multi-lingual pre-trained model could be extended to handle the multi-talker ASR, speaker counting, and timestamp prediction problems all at once. With this perspective, we choose the recently proposed Whisper [27] as our foundation model, and fine-tune it by using an enhanced SOT framework that includes an utterance timestamp prediction task. We also introduce a lightweight adapter [32] to alleviate catastrophic forgetting of languages that are not included in the fine-tuning data. Based on the experiments using the AMI [33] and AliMeeting [34] corpora, we demonstrate that the proposed framework provides a single multilingual model that can recognize overlapping speech with timestamp prediction, while accurately counting the number of speakers.

2. Methods

We leverage Whisper, one of the popular USMs, as a foundation model and adapt it to multi-talker speech recognition. The proposed enhanced SOT and adapter-based transfer learning methods are used to perform the adaptation.

2.1. Whisper as a foundation model

Whisper is based on an attention-based encoder-decoder (AED) [35, 36] architecture. Both the audio encoder and text decoder comprise Transformer [21] blocks. The Whisper audio encoder
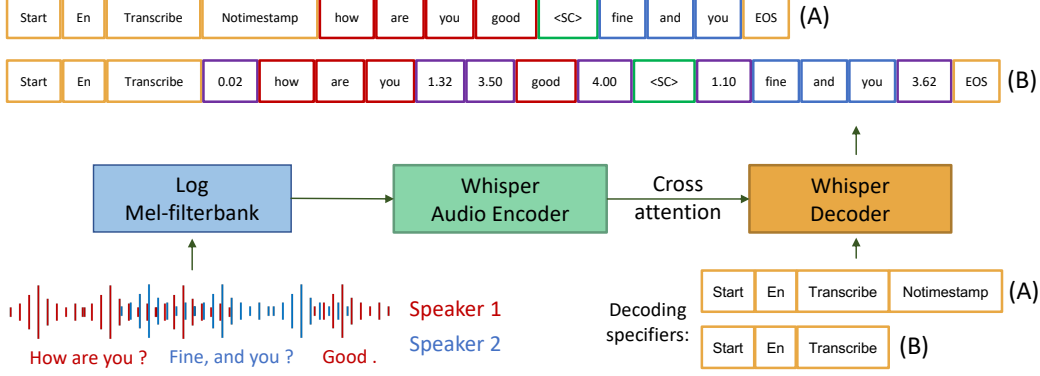

Figure 1: *An illustration of Whisper multi-talker transfer learning with SOT. (A): Training labels in SOT without timestamp; (B) Training labels in time-stamped SOT.*

encodes log mel-filterbank feature $\mathbf{X} \in \mathbb{R}^{f_a \times l_a}$ into deep hidden emebddings $\mathbf{H} \in \mathbb{R}^{f_h \times l_h}$:

$$\mathbf{H} = \text{Encoder}(\mathbf{X}), \quad (1)$$

where f_a, f_h and l_a, l_h are the dimension and sequence length of input feature X and hidden emebddings H, respectively. The text decoder can be regarded as an audio-conditional autoregressive language model:

$$\mathbf{o}_n = \text{Decoder}(\mathbf{t}, Y_{1:n-1}, \mathbf{H}), \quad (2)$$

where $\mathbf{o}_n \in \mathbb{R}^{|\mathcal{V}|}$ is an output distribution for estimating the n-th token y_n from token vocabulary $\mathcal{V}$, $Y_{1:N} = [y_1, \cdots, y_N | y_n \in \mathcal{V}]$ represents the estimated token sequence with length of N, and $\mathbf{t} = \{t_1, \cdots, t_I | t_i \in \mathcal{V}\}$ is a sequence of special tokens[1] that specifies a decoding task.

Whisper was trained with multi-task learning on multilingual speech data. As a result, it showed remarkable multilingual ASR capabilities and thus can be seen as a USM. However, its ASR accuracy severely deteriorates when we apply it to a multi-talker speech recognition task. This could be because the multi-talker task was not explicitly included in the multi-task training.

2.2. SOT with time-stamp prediction

To address the multi-talker ASR challenge, we consider adapting a USM to build a single AED model that jointly performs multi-talker ASR, speaker counting, and timestamp prediction. To this end, we propose an enhanced version of SOT to incorporate the timestamp prediction task during the adaptation.

SOT was proposed to train an AED model to perform multi-talker ASR [16]. Although other multi-talker ASR models usually have multiple output branches [6, 8, 12], SOT assumes an AED model with one output branch. Therefore, it can be used to convert a pre-trained USM into a multi-talker model without changing the model structure. With SOT, the utterances of all speakers are concatenated to form a single token sequence by inserting a special token $\langle sc \rangle$ denoting a speaker change. Fig.1.A illustrates how SOT organizes its training labels. The reference token sequence can be given as $R = [r_1^1, \cdots, r_{N1}^1, \langle sc \rangle, r_1^2, \cdots, r_{N2}^2, \langle eos \rangle]$, where r_i^j is the i-th token of the j-th speaker and $\langle eos \rangle$ denotes the end of the

output sequence that is only used when the utterances of all speakers are transcribed. We use the speaker-wise first-in-first-out (FIFO) style [16] to form the training labels. That is, the reference transcriptions of individual speakers are sorted by the start times of their first utterances. As a consequence, the model is expected to transcribe all the utterances of the first speaker, then output a $\langle sc \rangle$ token, and then continue transcribing the remaining speakers one by one.

We propose to incorporate the timestamp prediction task into SOT by leveraging the timestamp tokens of Whisper [27]. Whisper includes a set of additional tokens in the token vocabulary $\mathcal{V}$ to represent discretized timestamps at a 20-ms resolution. These timestamp tokens are predicted by the decoder at the same time as the tokens representing the transcriptions. We incorporate the timestamp tokens into SOT as illustrated in Fig. 1.B. During training, the timestamp tokens are inserted in the training labels at the beginning and the ending of each speaker's homogeneous segment. The speaker homogeneous segment consists of consecutive tokens spoken by the same speaker where the silence periods between the neighboring tokens do not exceed 2 seconds. Note that, unlike the case with Whisper, the timestamps do not increase monotonically within the token sequence and have overlapped regions among different speakers.

2.3. Adapter for cross-lingual transfer learning

Whisper was pre-trained on 680K hours of multilingual audio, of which a 117K hours subset covered 96 non-English languages [27]. It is desirable if a model adapted with multi-talker English data can also recognize the multi-talker speech of other languages that are not seen during the adaptation. In our preliminary experiments, when we fine-tuned the entire Whisper model with the enhanced SOT method using English data, the obtained model performed poorly for other languages.

To improve multilingual generalization, we employ a parameter-efficient transfer learning method proposed in [32]. Specifically, we insert bottleneck adapter modules with a limited number of parameters into the Transformer basic blocks of both the encoder and decoder. We optimize only the newly introduced parameters and freeze most of the original parameters with the aim that the newly added parameters could learn to handle multiple talkers in a less language-dependent fashion. The detailed structure of the adapter module will be described in Sec.3.3.

---

[1]For example, $\mathbf{t}$ contains $[\langle start \rangle, \langle English \rangle, \langle transcribe \rangle, \langle notimestamp \rangle]$ for an English ASR task. Detailed descriptions can be found in [27, Sec. 2.3].

Table 1: *WER (%) and LDER (%) for AMI-SDM evaluation set.*

Exp. ID	Pre-training model	Pre-training data	Fine-tuning data	Timestamp prediction	WER (w.r.t. # of talkers) (%)					LDER (w.r.t. # of talkers) (%)				
					avg.	1	2	3	4	avg.	1	2	3	4
-	Multi-talker AED [37]	900k hrs multi-talker simulation	AMI-SDM	-	21.2	14.7	19.6	25.7	35.5				-	
1	Whisper small	680k hrs Internet	-	-	47.3	25.4	46.3	68.0	82.5				-	
2			AMI-SDM	w/o	32.5	14.9	27.0	45.7	68.9				-	
3			AMI-SDM	w	28.5	14.7	25.9	41.2	54.7	10.4	0.9	6.8	21.7	35.3
4	Whisper medium	680k hrs Internet	-	-	44.1	21.1	43.4	65.4	78.6				-	
5			AMI-SDM	w/o	26.8	12.3	21.6	36.8	62.2				-	
6			AMI-SDM	w	23.6	12.8	21.8	32.5	45.9	8.2	0.9	5.9	16.5	27.6
7	Whisper large	680k hrs Internet	-	-	44.2	22.5	44.6	65.0	78.5				-	
8			AMI-SDM	w/o	28.4	13.0	23.0	38.7	66.0				-	
9			AMI-SDM	w	21.4	12.0	20.0	29.3	40.6	6.2	0.8	4.5	11.5	22.3

3. Experiments

3.1. Dataset

3.1.1. English Data

We use the AMI meeting corpus [33], which contains approximately 100 hours of English meeting recordings, for both SOT-based fine-tuning and testing. We utilize the audio signals from the first channel of the microphone array, a.k.a. the single distant microphone (SDM) audio. We segment each recording at silence positions or non-overlapping utterance boundaries into shorter *utterance groups* for both training and evaluation. Further details of *utterance groups* and the dataset specification can be found in [37]. After the segmentation, there is a total of 66 hours of SDM training data.

We also utilize 120-hour or 360-hour meeting-style English data simulated using LibriSpeech [38] in some experiments. Our simulation setup follows the one used in [39, 40], except that we use *utterance groups* of 30 seconds or shorter and that there are at most 4 speakers in each *utterance group*. To promote the learning of multi-talker recognition, these simulated data have more utterance overlaps than the AMI-SDM segments. The overlap ratio of each *utterance group* is between 60% and 80%.

3.1.2. Mandarin Chinese Data

We use AliMeeting [34] to test our fine-tuned multi-talker models in different languages with a zero-shot approach. The AliMeeting corpus contains approximately 108 hours of real meeting recordings in Mandarin Chinese. We use the audio signals from the third channel of the far-field microphone array. The meeting-long recordings are segmented into *utterance groups* in the same manner as with AMI-SDM. The *utterance groups* that are longer than 30 seconds are not used. As a result, there are 31.3, 1.5, and 3.9 hours of data for training, validation, and testing, respectively. The training set is only used in Exp. 10 of Table 3 to provide a reference point for the cross-lingual task.

3.2. Evaluation

At inference time, the multi-talker fine-tuned models generate utterances for one or multiple speakers. For the English test set, we calculate word error rate (WER) by following the procedure described in [37]. Namely, for each utterance group, the best speaker alignment between the hypotheses and the references is selected among all possible permutations. The WER is calculated by dividing the total error count from all utterance groups by the total number of reference words. For the Mandarin test set, we compute the character error rate (CER) by using the same procedure except that we use the characters as the recognition unit. The text normalization scheme of Kaldi [41] AMI recipe is applied in the AMI-SDM evaluation. For AliMeeting, we use Whisper's default text normalizer and then convert all the traditional Chinese characters into simplified Chinese[2].

In addition, to assess the quality of speaker and timestamp prediction, we compute the diarization error rate (DER) within each *utterance group*, which we call local DER (LDER). We also report the speaker counting accuracy for each *utterance group* by following [37].

Table 2: *AMI-SDM speaker counting accuracy comparison between SOT baseline (Exp. 5) and joint timestamp prediction (Exp. 6).*

Exp. ID	Actual # of talkers	Estimated # of talkers (%)				
		1	2	3	4	≥5
5	1	**98.6**	1.4	0.1	0.0	0.0
	2	16.9	**76.2**	6.8	0.2	0.0
	3	4.0	40.9	**46.6**	7.9	0.6
	4	1.5	17.7	42.9	**28.6**	9.3
6	1	**97.7**	2.1	0.2	0.0	0.0
	2	12.6	**72.3**	14.0	1.0	0.0
	3	1.6	24.1	**56.0**	15.9	2.3
	4	0.0	8.2	39.2	**35.6**	16.9

3.3. Training configuration

We fine-tune Whisper models based on the AMI-SDM training data with the enhanced SOT-style reference transcriptions. Three Whisper models, including small, medium, and large (v1) [27], are examined. We conduct fine-tuning experiments without and with the adapter module.

For the experiments without the adapter module, we first extend the token embedding set of the decoder by adding a randomly-initialized embedding representing ⟨sc⟩ token. All the Whisper model parameters are then updated with the AdamW optimizer. We use a linear decay learning rate (LR) scheduler with the initial LR of $1e-6$ (Exp. {2,3,5,6,8,9,10}). We fine-tune the small and medium models for 2 epochs with a batch size of 2 while a batch size of 1 is used for the large model.

For the experiments using the adapter module, we freeze most of the original Whisper model parameters, and only update the parameters in the layer normalization modules in the

---

[2]A traditional Chinese character and its simplified version differ in writing but have the same pronunciation and meaning.

Table 3: *Comparison of generalization ability for unseen language (Mandarin Chinese) in multi-talker processing.*

Exp. Id	Fine-tuning data	Adapter	AMI-SDM (English)							AliMeeting (Mandarin)					
			WER	LDER	SC Acc in # talkers				CER	LDER	SC Acc in # talkers			
					1	2	3	4			1	2	3	4
4	-	-	44.0	-	100.0	0.0	0.0	0.0	59.1	-	100.0	0.0	0.0	0.0
10	AliMeeting	w/o	39.4	22.7	99.3	12.3	1.5	0.5	28.3	12.8	95.3	80.2	52.6	28.8
6	AMI-SDM	w/o	23.6	8.2	97.7	72.3	56.0	35.6	50.7	35.2	91.8	65.4	26.9	11.1
11	AMI-SDM	w	25.0	9.3	97.6	73.9	46.5	29.3	43.7	26.9	95.7	60.4	21.0	5.2
12	+ LibriSpeech 120h	w	24.7	9.0	97.4	74.4	49.3	28.7	40.1	22.1	89.1	72.3	32.9	20.0
13	+ LibriSpeech 360h	w	24.3	8.7	98.4	73.9	47.3	27.1	36.9	18.0	91.9	73.2	28.7	16.3

Transformer blocks, the token embedding for ⟨sc⟩, and the parameters constituting the adapter module. The initial LR of 1e-5 is used. The adapter module is a two-layer bottleneck feedforward network with a ReLU activation function. The input and output dimensions are the same as the Whisper model's width and the bottleneck dimension is 256. We add two adapter modules to each Transformer block contained in the Whisper encoder and decoder: one after the self-attention layer and one after the feed-forward layer. There is a residual connection [42] within each adapter module, and near-identity initialization [32] is applied to the adapter modules at the start of fine-tuning.

3.4. Results on AMI-SDM

We first evaluate models fine-tuned without adapter modules by using the AMI-SDM evaluation set. Table 1 shows the WER comparison of the original Whisper models (Exp. {1,4,7}), the models fine-tuned with the original SOT without timestamp prediction (Exp. {2,5,8}) and the models fine-tuned with the enhanced SOT (Exp. {3,6,9}). Large WER margins between the original models and the SOT fine-tuned models are consistently observed for all three model sizes. This shows the SOT fine-tuning makes the Whisper models capable of handling the overlapped multi-talker speech. By comparing the experiments with/without timestamp prediction, we find that the models fine-tuned with timestamp prediction greatly reduce the WERs for the utterance groups including 3- and 4-speakers. We further calculate the speaker counting accuracy and report it as a confusion matrix in Table 2. From the result, we can observe that the addition of the timestamp prediction task helps improve the speaker counting accuracy for the utterance groups with 3 and 4 speakers, which aligns with the significant WER improvements. Our hypothesis for these improvements is that the timestamp tokens can serve as an explicit indicator for the SOT decoder to identify the location of the overlapping region. As a result, the learning of the cross-attention in the decoder may have been facilitated. Further investigation is desirable to better understand this phenomenon, and we will leave it to future work.

We also include the state-of-the-art (SOTA) system on the same testing setup (i.e. utterance-group-level evaluation that does not allow the use of oracle utterance boundaries) from previous work [37] for better comparison. Our best system (Exp. 9) achieves very close overall WER to the SOTA results but there is still a gap for the 3- and 4-speakers breakdown WER. The reason might be that large-scale simulated multi-talker speech is used for SOT pre-training in [37], and their pre-trained model already has some good knowledge for multi-talker processing.

The LDERs of the enhanced SOT-based systems are also listed in Table 1. We can see a clear trend that the LDER becomes better as the model size increases, especially for the utterance groups with many speakers. These results show that the

obtained model can perform ASR and timestamp prediction for multiple talkers while counting them at the same time.

3.5. Results on Zero-shot Multilingual Transfer

Table 3 reports the WERs and LDERs of different systems on the AliMeeting testset to see if the models adapted by the English multi-talker data are applicable to the Mandarin testset. We first evaluate the results on the original Whisper medium model (Exp. 4) and the AliMeeting multi-talker fine-tuned model (Exp. 10). These results can be regarded as the lower- and upper-bound for other evaluations in the table.

The model without the adapter module (Exp. 6) achieves the best performance on the AMI-SDM dataset, but it underperforms all the other fine-tuned models on the AliMeeting dataset. It shows that naively fine-tuning the model with the English-only multi-talker data impairs the multi-lingual capability which the original model is equipped with.

On the other hand, we observe significant improvements in both WER and LDER for the AliMeeting testset when we introduce the adapter module, while mostly preserving the same WER and LDER for the AMI-SDM testset (Exp. 6 vs Exp. 11). This result shows that the adapter module can help the model acquire the multi-talker recognition capability in a less language-dependent way so that the obtained model can handle the multi-talker data in multiple languages.

We further fine-tune the model with additional simulated LibriSpeech meeting-style data with high overlap ratios. Comparing Exp. {12,13} with Exp. 11, there are large improvements in speaker counting accuracy for the 3- and 4-speaker utterance groups. It shows that using more overlapped English training data facilitates the learning of multi-talker processing, and the learned knowledge from English can be well transferred into Chinese multi-talker meetings.

4. Conclusion

In this paper, we propose an effective approach to adapt USMs for multi-talker ASR. The enhanced version of serialized output training with timestamp prediction is first introduced for handling multiple-talkers with accurate speaker and time prediction. We also introduce the adapter module to transfer cross-lingual knowledge for multi-talker processing from a single language data. Our experiments show that the enhanced SOT successfully converts Whisper to multi-talker ASR that can recognize overlapping speech while predicting the speaker and timestamps with high accuracy. By introducing the adapter module, our proposed model becomes able to handle multi-talker audio even when the language spoken in the audio is unseen in the adaptation data.

5. References

[1] J. Barker, S. Watanabe *et al.*, "The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines," in *Proc. ISCA Interspeech*. ISCA, Sep. 2018, pp. 1561–1565.

[2] S. Watanabe, M. Mandel *et al.*, "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings," in *6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*. ISCA, May 2020, pp. 1–7.

[3] 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 *Proc. ISCA Interspeech*. ISCA, 2006.

[4] Y. Qian, X. Chang *et al.*, "Single-channel multi-talker speech recognition with permutation invariant training," *Speech Communication*, vol. 104, pp. 1–11, Nov. 2018.

[5] T. Yoshioka, D. Dimitriadis *et al.*, "Meeting Transcription Using Asynchronous Distant Microphones," in *Proc. ISCA Interspeech*. ISCA, Sep. 2019, pp. 2968–2972.

[6] D. Yu, M. Kolbæk *et al.*, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE ICASSP*, Mar. 2017, pp. 241–245.

[7] M. Kolbæk, D. Yu *et al.*, "Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," *IEEE/ACM Trans. ASLP.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[8] D. Yu, X. Chang *et al.*, "Recognizing Multi-Talker Speech with Permutation Invariant Training," in *Proc. ISCA Interspeech*. ISCA, Aug. 2017, pp. 2456–2460.

[9] H. Seki, T. Hori *et al.*, "A Purely End-to-End System for Multi-speaker Speech Recognition," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2620–2630.

[10] X. Chang, Y. Qian *et al.*, "End-to-end Monaural Multi-speaker ASR System without Pretraining," in *Proc. IEEE ICASSP*, May 2019, pp. 6256–6260.

[11] X. Chang, W. Zhang *et al.*, "MIMO-Speech: End-to-End Multi-Channel Multi-Speaker Speech Recognition," in *Proc. IEEE ASRU*, Dec. 2019, pp. 237–244.

[12] W. Zhang, X. Chang *et al.*, "Improving End-to-End Single-Channel Multi-Talker Speech Recognition," *IEEE/ACM Trans. ASLP.*, vol. 28, pp. 1385–1394, 2020.

[13] M. Delcroix, K. Zmolikova *et al.*, "Single Channel Target Speaker Extraction and Recognition with Speaker Beam," in *Proc. IEEE ICASSP*, Apr. 2018, pp. 5554–5558.

[14] N. Kanda, S. Horiguchi *et al.*, "Auxiliary Interference Speaker Loss for Target-Speaker Speech Recognition," in *Proc. ISCA Interspeech*. ISCA, Sep. 2019, pp. 236–240.

[15] Z. Huang, D. Raj *et al.*, "Adapting self-supervised models to multi-talker speech recognition using speaker embeddings," Nov. 2022.

[16] N. Kanda, Y. Gaur *et al.*, "Serialized Output Training for End-to-End Overlapped Speech Recognition," in *Proc. ISCA Interspeech*. ISCA, Oct. 2020, pp. 2797–2801.

[17] M. McAuliffe, M. Socolof *et al.*, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Proc. ISCA Interspeech*, vol. 2017, 2017, pp. 498–502.

[18] J. Li *et al.*, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[19] N. Kanda, X. Xiao *et al.*, "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed asr," in *Proc. IEEE ICASSP*. IEEE, 2022, pp. 8082–8086.

[20] I. Sklyar, A. Piunova *et al.*, "Separator-Transducer-Segmenter: Streaming Recognition and Segmentation of Multi-party Speech," in *Proc. ISCA Interspeech*, 2022, pp. 4451–4455.

[21] A. Vaswani, N. Shazeer *et al.*, "Attention is All you Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[22] S. Schneider, A. Baevski *et al.*, "Wav2vec: Unsupervised Pretraining for Speech Recognition," *arXiv:1904.05862 [cs]*, Sep. 2019.

[23] A. Baevski, H. Zhou *et al.*, "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *arXiv:2006.11477 [cs, eess]*, Oct. 2020.

[24] W.-N. Hsu, B. Bolte *et al.*, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Trans. ASLP.*, vol. 29, pp. 3451–3460, 2021.

[25] S. Chen, C. Wang *et al.*, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *arXiv:2110.13900 [cs, eess]*, Oct. 2021.

[26] C. Wang, Y. Wu *et al.*, "UniSpeech: Unified Speech Representation Learning with Labeled and Unlabeled Data," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 10 937–10 947.

[27] A. Radford, J. W. Kim *et al.*, "Robust Speech Recognition via Large-Scale Weak Supervision," Dec. 2022.

[28] S.-w. Yang, P.-H. Chi *et al.*, "SUPERB: Speech processing Universal PERformance Benchmark," Oct. 2021.

[29] Z. Huang, S. Watanabe *et al.*, "Investigating Self-Supervised Learning for Speech Enhancement and Separation," in *Proc. IEEE ICASSP*, May 2022, pp. 6837–6841.

[30] H. Song, S. Chen *et al.*, "Exploring WavLM on Speech Enhancement," in *Proc. IEEE SLT*, Jan. 2023, pp. 451–457.

[31] Z. Chen, S. Chen *et al.*, "Large-Scale Self-Supervised Speech Representation Learning for Automatic Speaker Verification," in *Proc. IEEE ICASSP*. IEEE, 2022, pp. 6147–6151.

[32] N. Houlsby, A. Giurgiu *et al.*, "Parameter-Efficient Transfer Learning for NLP," Jun. 2019.

[33] J. Carletta, S. Ashby *et al.*, "The AMI Meeting Corpus: A Pre-announcement," in *Machine Learning for Multimodal Interaction*, ser. Lecture Notes in Computer Science, S. Renals and S. Bengio, Eds. Berlin, Heidelberg: Springer, 2006, pp. 28–39.

[34] F. Yu, S. Zhang *et al.*, "M2Met: The Icassp 2022 Multi-Channel Multi-Party Meeting Transcription Challenge," in *Proc. IEEE ICASSP*, May 2022, pp. 6167–6171.

[35] J. K. Chorowski, D. Bahdanau *et al.*, "Attention-Based Models for Speech Recognition," in *Advances in Neural Information Processing Systems*, vol. 28. Curran Associates, Inc., 2015.

[36] W. Chan, N. Jaitly *et al.*, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE ICASSP*, Mar. 2016, pp. 4960–4964.

[37] N. Kanda, G. Ye *et al.*, "Large-Scale Pre-Training of End-to-End Multi-Talker ASR for Meeting Transcription with Single Distant Microphone," in *Proc. ISCA Interspeech*. ISCA, Aug. 2021, pp. 3430–3434.

[38] V. Panayotov, G. Chen *et al.*, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE ICASSP*, Apr. 2015, pp. 5206–5210.

[39] C. Li, Z. Chen *et al.*, "Dual-Path Modeling for Long Recording Speech Separation in Meetings," in *Proc. IEEE ICASSP*, Jun. 2021, pp. 5739–5743.

[40] ——, "Dual-Path Modeling With Memory Embedding Model for Continuous Speech Separation," *IEEE/ACM Trans. ASLP.*, vol. 30, pp. 1508–1520, 2022.

[41] "The Kaldi Speech Recognition Toolkit," in *Proc. IEEE ASRU*, D. Povey, A. Ghoshal *et al.*, Eds. IEEE Signal Processing Society, 2011.

[42] K. He, X. Zhang *et al.*, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.