Listening to Multi-talker Conversations:
Modular and End-to-end Perspectives

Desh Raj
Department of Computer Science
Johns Hopkins University
draj2@jhu.edu

1 Introduction

1.1 Motivation

Since the first large vocabulary continuous speech recognition (LVCSR) systems were built more than 30 years ago, improvement in voice technology has enabled applications such as smart assistants on mobile phones, semi-automated customer support, and embedded systems for appliances (such as smart refrigerators, for instance). Nevertheless, today's systems are passive listeners which transcribe single-speaker utterances to feed downstream language understanding components. **Conversational intelligence** of the future is expected to comprise systems that can actively participate in human conversations, including scenarios such as note-taking or fact-checking in meetings, collaborative learning in education, or simply recommending grocery items in households. While such systems would require intelligence in diverse modalities — dialog systems for context handling, emotion recognition from speech and video, common sense reasoning, to name a few — their ability to recognize **free-flowing multi-party conversations** is a crucial and challenging component that still remains unsolved.

1.2 Goal

Multi-talker speech recognition is a broad task with several applications; in this dissertation I focus specifically on the "meeting" setting with the following salient features: (i) long-form unsegmented recordings with duration up to a few hours, (ii) unknown number of speakers (usually 2 to 8), (iii) moderate overlapped speech conditions (often 20-30% per session), and (iv) either single or multi-channel far-field recording conditions, with varying degrees of ambient noise. The task is formulated as *speaker-attributed speech recognition*, which comprises two underlying tasks: (a) automatic speech recognition (ASR), or the task of "what was said," and (b) speaker diarization (SD), or the task of "who spoke when." Formally, given a recording $\mathbf{X} \in \mathbb{R}^{D \times T}$, where D denotes channels (microphones) and T is the number of samples in the audio, the goal of speaker-attributed ASR is to compute $\hat{\mathbf{Y}} = \{\hat{Y}^1, \ldots, \hat{Y}^K\}$, where K is the number of speakers in the recording, and $\hat{Y}^k$ represents the sequence of words (or other transcription units) from speaker k, i.e., $\hat{Y}^k = (y_1^k, \ldots, y_{n_k}^k)$. Since neither the number K of speakers, nor speaker identities are not known in advance, $\{1, \ldots, K\}$ may be any permutation of the speakers which minimizes the overall error rate between $\hat{\mathbf{Y}}$ and the ground truth $\mathbf{Y}$.

In this work, I will explore two solution paradigms for this problem — modular, and end-to-end. Broadly speaking, modular approaches aim to decouple the tasks of transcription and speaker attribution, and solve each of these using independently optimized components. This may be in the form of a "pipeline" strategy, where we first estimate time-marked speaker segments $\hat{\mathbf{X}} = \{\hat{X}^1, \ldots, \hat{X}^K\}$ from $\mathbf{X}$, followed by transcribing each segment (§ 3.1), or in the form of a "simultaneous" strategy, where the two tasks are processed independently (and in parallel), followed by combination to produce $\hat{\mathbf{Y}}$ (§ 3.3). In the end-to-end framework, the sequence of speaker-annotated tokens is estimated jointly and a single model is trained for this task (§ 4.2). Both these paradigms have their own advantages and limitations, which elicit several questions and outcomes, such as:

1. design and analysis of diarization methods to handle overlapping speech, by using overlap assignment to extend clustering-based diarization algorithms, or by the application of speech

separation;
2. discovery of mechanisms to mitigate errors caused by the mismatch of train-test conditions when using ASR or diarization systems with speech separation in the modular framework;
3. design and development of end-to-end streaming multi-talker systems based on the transducer model, including techniques for segmentation and speaker attribution; and
4. comparison of frameworks along several axes — accuracy, latency, system complexity — on popular benchmarks, namely LibriCSS, AMI, and AliMeeting [8, 30, 50].

2 Background & Related Work

Multi-talker speech recognition of free-flowing conversations is a well-known problem. In the offline setting, a long-form audio recording (ranging between several minutes up to a few hours) is provided, and the expected output is a speaker-attributed transcription with time marks. When deployed online, streaming audio is provided as input with the same transcription requirements. In addition to their use in conversational agents, these systems have several other applications — such as real-time meeting transcription for hearing-impaired participants, and generating automatic subtitles for movies or video streams, to name a few. When used in conjunction with language understanding or dialog systems, they also enable real conversational AI. However, systems for solving this problem are still far from human parity, often achieving between 30% and 50% error rates on the task.

In the 2000s, several advances were made as a result of NIST evaluations [12] and the AMI project [5] that were aimed at tackling the multi-talker recognition problem, primarily in the offline setting. More recently, challenges such as DIHARD [38] and CHiME [43] have focused on diarization and speech recognition tasks in very challenging scenarios. These efforts have highlighted several problems that need to be addressed. Specifically, careful selection of deep learning techniques for acoustic modeling, language modeling, and system combination was shown to reach human transcription accuracy for conversational telephone speech [47], but this evaluation was limited to conversations between two speakers. Although this is an important development, real multi-talker conversations raise many additional issues, most notably the case of overlapping speech. Studies have shown that meetings can contain up to 20% overlapping speech [5], which has implications for both diarization and ASR — diarization systems which make single-speaker assumptions miss at least one of the two speakers completely, and ASR systems trained on clean utterances are more error-prone on overlapped regions. Combined with the effect of non-stationary noise and reverberation in real recordings, the error rates in these settings may be degraded by up to 86% [48] in meetings, and 52% in dinner-party settings [43]. In popular literature, the task is often referred to as the "cocktail party problem", and solutions to addres it comprises several modules, each with a long history of research in speech processing (Figure 1).

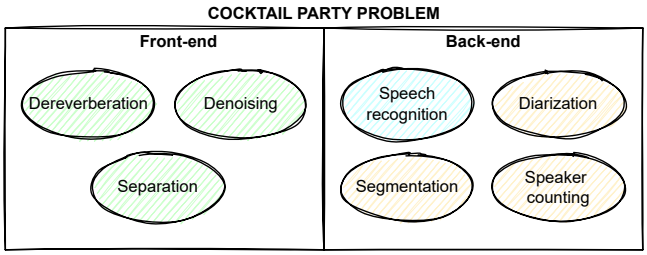


Figure 1: Traditional solutions the cocktail party problem, which attempt to enhance the signal (front-end) to support the eventual task (back-end).

The speaker diarization and ASR communities have independently sought to develop methods that handle multi-talker overlapping speech. For diarization, existing approaches to solve the problem involve using externally trained overlap detectors to identify frames that contain overlapping speech. Once overlaps are detected, an "overlap assignment" stage assigns multiple speaker labels to those frames [4]. A second set of methods train end-to-end neural systems to perform overlapping diarization in a supervised setting [13]. For multi-talker ASR, permutation-invariant training (PIT), which was first proposed for speech separation, has successfully been employed [49]; however, the transcriptions produced by such a method are unordered *across* utterances, which makes them dependent on an external speaker tracking or diarization module. A new framework called serialized output training [23] aims to mitigate some of the issues with PIT, and has been used to jointly perform ASR, speaker

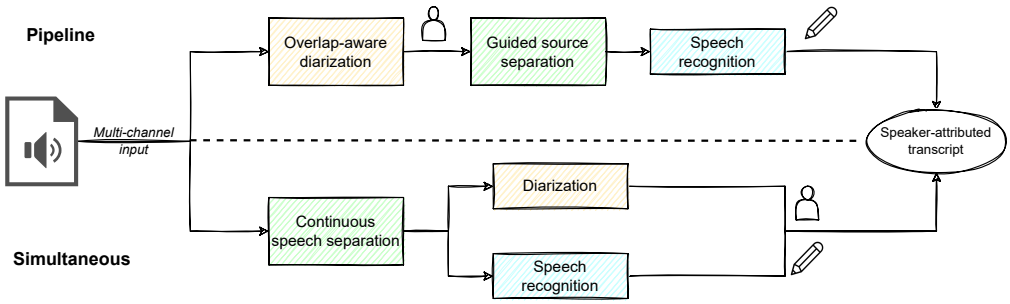


Figure 2: The architecture of modular systems for multi-talker speech recognition. The "pipeline" approach (*top* and § 3.1) relies on accurate overlap-aware diarization to produces segments for speech enhancement (GSS) and ASR. The "simultaneous" approach (*bottom* and § 3.3) decouples transcription and speaker attribution by using a continuous speech separation front-end.

identification, and counting [22, 21].

With these advancements in diarization and ASR systems, there is increasing interest to develop systems that perform multi-talker speech recognition for unsegmented recordings. The latest edition of the CHiME challenge [43] included a track for evaluating systems for dinner-party conversations. An iterative training strategy combining clustering-based diarization with target-speaker ASR has been proposed recently [24]. The need for controlled, but realistic, data to investigate such systems has resulted in several new datasets such as LibriCSS [8] and LibriMix [10]. The research agenda at the JSALT 2020 workshop[1] included a project aimed at "building fully contained multi-talker audio transcription systems based on speech separation and extraction". The M2MeT challenge [50] was organized at ICASSP 2022 for the specific task of multi-channel multi-party meeting transcription.

3 A modular perspective on multi-talker speech recognition

3.1 Pipeline approach: the CHiME-6 challenge

Track 2 of the CHiME-6 challenge[2] was the first community challenge activity which tackled the task of unsegmented multi-talker speech recognition under far-field conditions. The challenge baseline [43] proposed a pipeline system comprising 3 modules. First, a clustering-based diarization system identifies homogeneous speaker segments in the recording. A guided source separation (GSS) system uses the speaker activity to generate masks for a beam-former that enhances the speech [3]. Finally, an ASR system then processes the enhanced speech segments to obtain the speaker-annotated transcripts.

A naive stitching of the modules in sequence introduces several sources of error. The presence of overlapping speech in the CHiME-6 data (approx. 40% of total speech in the evaluation set) degrades clustering-based diarization, which assumes single-speaker recordings. This error propagates to the GSS-based enhancement, and finally to the ASR, which is trained only on single-speaker segments. Furthermore, even with perfect (oracle) diarization, enhancement leads to a train-test mismatch for the ASR system, which negatively impacts word error rates (WER). To alleviate these issues, our prior introduced overlap assignment using the VBx resegmentation method [4] to improve the hierarchical clustering based diarizer, and included GSS-enhanced segments as data augmentation during ASR training. Combined with multi-channel fusion strategies, these resulted in absolute error rate improvements of over 10% in the challenge evaluation.

Status. The pipeline approach for the CHiME-6 corpus was published at the CHiME-6 workshop at IEEE ICASSP 2020 [1], but has not yet been evaluated on meeting tasks (LibriCSS, AMI, and AliMeeting). Implementations for individual components (diarizer[3] and GSS[4]) are complete — I am

[1] https://www.clsp.jhu.edu/workshops/20-workshop/
[2] https://chimechallenge.github.io/chime6/
[3] https://github.com/desh2608/diarizer
[4] https://github.com/desh2608/gss

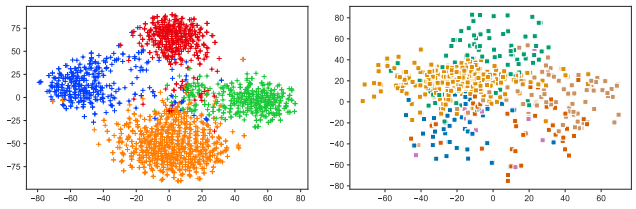
working on a manuscript comparing modular approaches (including the simultaneous approach that we will see next), where detailed results on the meeting tasks will be presented.

Relevance. This work serves as an introduction to the task of multi-talker ASR with a traditional pipeline approach. While the CHIME-6 challenge was limited to the provided dinner-party conversations, it introduced the *concatenated minimum-permutation word error rate* (cpWER) evaluation metric, which has subsequently been adopted for the task of speaker-attributed ASR. I will apply the pipeline approach developed in this project for meeting transcription tasks described earlier, and all proposed systems in the dissertation will be evaluated using cpWER.

3.2 Overlap-aware speaker diarization

Previously, we mentioned that handling overlapping speech in the diarization module can be achieved in one of several ways. The conventional approach to speaker diarization [14] formulates the task as a clustering problem — the speech regions in the recording are split into small segments, an embedding is obtained for each segment using a speaker embedding extractor [40], followed by clustering of the embeddings. Evidently, this process assumes the absence of overlapping speech, and is thus prone to errors on meeting tasks. In § 3.1, we used overlap assignment based on VB-resegmentation, which is based on the assumption that the speaker activity matrix generated during the process encodes speaker likelihoods in the segment. However, this arbitrary heuristic may often lead to an incorrect assignment of the second speaker, resulting in increased speaker assignment errors (i.e. confusions) in the overlapped regions. Furthermore, VBx requires initialization with a different clustering-based diarizer, which may not be desirable in practice.

In Raj et al. [34], we proposed an overlap-aware diarization algorithm based on multi-class spectral clustering (SC). Using an alternative formulation of SC [51], we relax the discrete clustering task into a convex optimization problem solvable by eigen-decomposition. Thereafter, the convex solution set is discretized by alternating between singular value decomposition (SVD) and modified non-maximal suppression which encodes the overlap constraints. This principled scheme of overlap assignment is found to reduce speaker confusion errors on overlapped speech regions, compared to VB-resegmentation (from 18% to 14% on the AMI evaluation set, for instance).



(a) **(b)**

Figure 3: T-SNE plots of x-vector embeddings for (a) non-overlapping, and (b) overlapping segments for the recording `EN2002a` in the AMI eval set (containing 4 speakers). Colors denote the speaker assigned to the segment. For (b), each color represents a distinct pair of speaker labels, resulting in 6 differently colored clusters.

Nevertheless, this method of overlap-aware diarization has several limitations. Training an accurate overlap detector is a challenging task in itself — models are often required to be trained on synthetic force-aligned overlapping speech due to sparsity of real annotated data, causing a mismatch between train and test conditions. False alarms in overlap detection may lead to the diarizer "hallucinating" extra speakers, resulting in degradation of error rates for both diarization and downstream ASR. Even assuming oracle overlap detection, second speaker assignment is often error-prone because speaker embedding extractors are trained on single-speaker utterances. From Fig. 3, we can observe that embeddings for speaker-pairs from overlapping regions are not well-separated, leading to errors in speaker assignment [34]. Most importantly, while these methods are well suited for the pipeline approach described in § 3.1, they add needless complexity for frameworks based on separation, that I will describe next.

Status. The method for overlap-aware spectral clustering was published at IEEE SLT 2021 [34]. Its impact on the pipeline approach is under investigation (cf. § 3.1).

Relevance. This work builds upon the pipeline framework described earlier, and proposes new

designs and methods for overlap-aware diarization. By comparing cpWERs of pipelines built on different diarizers, I will show the impact of diarization on GSS-based enhancement and ASR performance.

Other related work. Recently, several methods for overlap-aware diarization have been proposed that formulate the problem as frame-level multi-label classification, and train neural networks for the task. Among these is target-speaker voice activity detection (TS-VAD), which uses speaker embeddings as auxiliary inputs to the classifier to avoid speaker permutation issues. While the original TS-VAD was proposed for a fixed number of speakers in the CHiME-6 condition [31], we extended it to be applicable to an unknown number of speakers and demonstrated its application for meeting scenarios [19]. I have also developed DOVER-Lap [33], a method for combining the outputs from several such overlap-aware diarization systems, and proposed approximation algorithms (based on greedy and randomized local search) for speaker mapping between the systems [35]. These projects are not intended to be a core part of the dissertation.

3.3 Simultaneous systems based on CSS

Perhaps the most important drawback of the pipeline approach (§ 3.1) is the ASR's dependence on the diarization output. Since the diarizer is inherently an offline system, this dependence implies that transcripts can only be obtained once the entire recording has been processed, which greatly limits applicability of these systems to real-time meeting transcription. To alleviate this problem, we developed modular systems that decouple the diarizer and ASR into simultaneous components built upon continuous speech separation [8] — the framework is designed to produce word tokens in "real-time" as the meeting progresses, and attribute speakers to tokens (diarization) at the end of the session [32]. Such a system has also been referred to as *separate-recognize-diarize* in literature [48].

Continuous speech separation (CSS) transforms input speech into a fixed number (usually 2 or 3) of output streams such that each output stream is overlap-free. These streams are then used for performing diarization and ASR, simultaneously. We have shown that this framework performs well for a simple meeting task (LibriCSS). But there remain several questions to be addressed. First, we would like to extend clustering-based diarization to apply to the case of multiple audio streams. While this is straightforward for methods such as agglomerative hierarchical or spectral clustering, it is not immediately obvious how to modify VBx [26] for this task. The VBx model assumes that the input sequence of speaker embeddings is generated by a Bayesian HMM with speaker-specific state distributions, producing one embedding at every time step. In ongoing work, we are reformulating VBx such that the HMM states encode speaker tuples (instead of single speakers) that produce multiple speaker embeddings (one per audio stream). Alternatively, a simpler modification may be considered where each stream is processed independently but speaker states are shared across all streams.

Similar modifications are also required for ASR decoding strategies to ensure optimal coalition with CSS. Since CSS is designed to work with short windows (approx. 2-3 seconds), it may often break up long utterances into the component streams. This poses a challenge for ASR systems which rely on utterance context for best results. As such, I propose to improve the CSS-ASR synergy by designing "cross-decoders" which process all the separated streams jointly such that state (i.e. speaker) transitions are allowed to occur across the streams.

Status. A system describing the simultaneous approach for LibriCSS was published at IEEE SLT 2021 [32], but it has not yet been evaluated for more realistic meeting tasks such as AMI and AliMeeting. A manuscript for CSS-based diarization is in preparation for IEEE SLT 2022. Detailed investigations of the full system will be a part of the journal submission mentioned earlier in Section 3.1.

Relevance. This work explores an alternate framework for the design of modular multi-talker ASR systems. Compared with the pipeline approach, the simultaneous framework allows real-time transcription by decoupling the ASR from the diarization output. My objective in this work is to propose

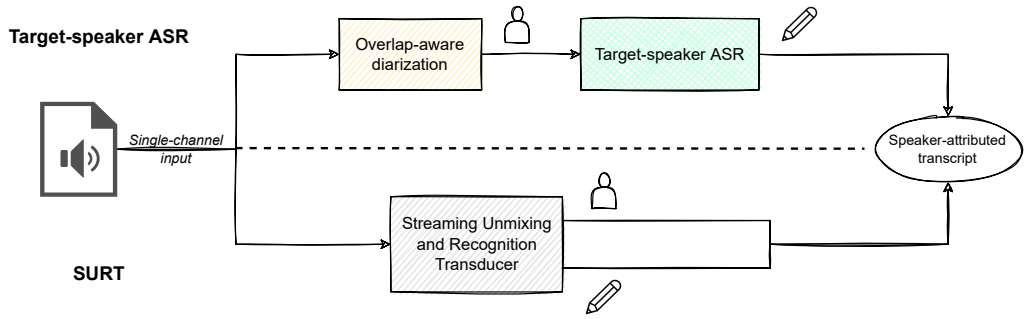


Figure 4: An overview of *separation-free* systems for multi-talker speech recognition. The top approach (§ 4.1) combines overlap-aware diarization with target-speaker ASR. SURT (§ 3.3; *bottom*) provides end-to-end speaker-annotated transcripts, and may be thought of as a separation-free analogue for the simultaneous modular framework of Figure 2.

extensions to diarization and ASR models that allow them to work within the CSS context.

4 Towards end-to-end multi-talker speech recognition

The modular systems described in Section 3 deal with overlapping speech present in multi-talker meetings by explicitly separating the recordings into single-speaker streams. This approach has several disadvantages. First, speech separation models based on neural networks often produce distortions in the target speech—even in non-overlapping speech regions—resulting in degradation of diarization and ASR performance [45]. Second, explicit separation adds unnecessary model complexity, not the least arising from the conversion of frequency-domain signals to time-domain and back. Furthermore, most single-channel speech separation methods work reasonably on clean speech mixtures, but are severely impacted by spectral smearing caused by noise and reverberations in realistic mixtures [29]. We will therefore explore "separation-free" frameworks that may be considered analogous to the pipeline and simultaneous approaches described earlier, and which can be used with single-channel recordings.

4.1 Combining overlap-aware diarization with target-speaker ASR

Biased (or informed) speech separation methods (such as SpeakerBeam [58] and VoiceFilter [42]) extract a *target* speaker's utterance from an overlapped mixture using auxiliary information about the target (usually in the form of speaker embeddings or enrolment audio). Similarly, biased ASR models, also known as target-speaker ASR (TS-ASR), transcribe the utterance of a target speaker in a mixture [11, 37]. As mentioned, these models have the advantage of being conceptually simpler and directly optimized for the downstream ASR task, as opposed to models based on explicit separation. We propose to pair TS-ASR with overlap-aware diarization as a separation-free alternative to the pipeline framework described in Section 3.1.

However, such a pairing comes with its own set of challenges. TS-ASR models are often trained with synthetically overlapped mixtures, using oracle enrolment audios or speaker embeddings. When used in conjunction with imperfect diarizers, incorrect speaker labels and real overlaps (such as backchannels) may be detrimental to ASR performance. We conjecture that some of these problems may be mitigated using auxiliary losses and training strategies to regularize the TS-ASR. For instance, a variation of scheduled sampling may be employed during training where the TS-ASR system is fed speaker segments obtained from a diarizer in lieu of enrolment embeddings/audio.

From a modeling perspective, we are interested in building TS-ASR systems based on the kind of encoder-decoder models popularly known as neural transducers (RNN-Transducer [15] or transformer-transducer [52]). These models combine the streaming capabilities of conventional CTC based models [16] with trainable neural decoders found in attention-based encoder-decoders (AEDs) [6], and as such, have received significant interest from both the academic and industrial research communities [7]. Their combination with time-domain target speech extraction has also been investigated [39].

For our separation-free setting, we will investigate methods to supply target speaker information to these models, such as auxiliary networks trained jointly with the transducer, and parameter-efficient fine-tuning strategies, such as adapters [19, 25].

Status. Implementation of overlap-aware diarization is complete (as described in earlier sections). Target-speaker ASR will be implemented in the k2 framework[5] which provides building blocks for transducer models. Results from the combined system will be submitted to IEEE ICASSP 2023 (submission deadline is October 2022).

Relevance. This work lays the foundations for going from modular towards end-to-end-trainable multi-talker systems. The framework demonstrates how the separation module can be removed by using target-speaker ASR in conjunction with overlap-aware diarization. By investigating best practices for informing the TS-ASR using the diarizer output, this work also builds on the idea of optimal synergies between speech processing components that has been a recurring theme in this dissertation.

Other related work. Some of the ideas proposed in this section are similar to our work on extending SpeakerBeam to be applied to diarization outputs [59] (contribution as a co-author). However, while SpeakerBeam generates target speech which is then decoded with a hybrid HMM-DNN ASR to produce speaker transcripts, my goal in this work is to obtain these transcripts "end-to-end," i.e., without going through intermediate speaker-specific audio signals.

4.2 Continuous streaming recognition based on neural transducers

Although target-speaker ASR mitigates the requirement for explicit separation, the overall framework in Section 4.1 is still dependent on the diarizer output, and as such, suffers from the same limitations as the pipelined modular systems described earlier. In the modular framework, this problem was solved by applying continuous speech separation to remove overlapping speech from the recording.

We adopt a similar strategy in the separation-free setting by applying the Streaming Unmixing and Recognition Transducer (SURT) [27], originally proposed for 2-speaker overlapping ASR. The model consists of an *unmixing* and a *recognition* component, where the former extracts speaker-specific features from overlapped speech using convolutional layers and the latter transcribes it using a neural transducer (Figure 5). This setup is conceptually similar to the CSS-ASR model (§ 3.3), with the major difference being that it is trained end-to-end with an ASR objective, such that no explicit separation is performed. While the original SURT was shown to obtain competitive WERs for 2-speaker single-turn sessions, we extended it to Multi-SURT for multi-talker meeting transcription by proposing several modifications [36]. Since meeting sessions may be arbitrarily long and ASR encoders are known to be unreliable on long-form recordings [9], we replace regular LSTMs with dual-path LSTM and transformer encoders in the transducer [28] to facilitate modeling of these long sequences. Training tricks such as chunk-width randomization and curriculum learning were found to improve the WERs on long sessions while retaining single-turn performance, and the resulting model demonstrated competitive results on the LibriCSS task.

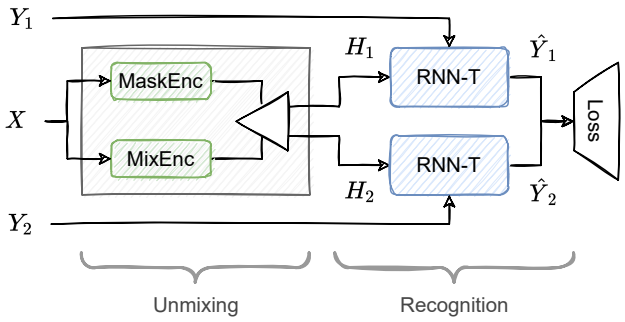


Figure 5: Streaming Unmixing and Recognition Transducer (SURT).

However, several important questions remain to be addressed. Since our objective in this work is *speaker-attributed* ASR, we are interested in using the SURT model to also perform speaker diarization along with transcription. This may be accomplished in one of several ways. If we have access to

[5] https://github.com/k2-fsa/k2

speaker enrolment audio, their corresponding embeddings may be provided as auxiliary information to the network, such that the transducer is now tasked to predict a `(token, speaker)` pair at every time step. Such a system has been referred to as a *speaker inventory* in literature [18]. Alternatively, the transducer can be tasked to predict speaker embeddings that can then be clustered to obtain speaker identities, similar to the *turn-to-diarize* strategy [46]. Even when design issues are solved, the transducer-based multi-talker ASR may require substantive engineering to make training feasible under resource constraints. For instance, the common industrial practice for training RNN-Ts suggests using subwords (such as BPEs) as modeling units, and allowing arbitrarily many labels per time-step, both of which contributes to a large memory requirement for storing the output logits. Following recent work that decouples modeling units from label topologies for CTC/MMI systems [54], one may ask whether the memory blow-up may be prevented by using smaller modeling units (such as graphemes/chenones or monophones/senones) with a monotonically constrained topology (i.e., one output label per time step).

Status. The multi-turn SURT model (evaluated on LibriCSS) is accepted for publication at IEEE ICASSP 2022 [36], but has not yet been evaluated on AMI or AliMeeting. Since this work was done during an internship at Microsoft (Summer 2021), the code for SURT is not publicly available, and as such, needs to be reimplemented with open-source toolkits, such as Icefall/k2/Lhotse [57]. The extension of SURT for speaker diarization is planned for Fall/Winter 2022.

Relevance. This work will complete the framework shift that started with pipeline-based modular systems, by designing and analyzing a fully end-to-end trainable system based on neural transducers. It will also detail the practical issues involved in implementing such systems, which make modular frameworks still a popular choice for multi-talker ASR.

4.3 Advances in SURT: graph-PIT, multi-channel models, and self-supervised learning

A common problem that occurs in any system that processes multi-talker input is that of label permutation ambiguity. For example, for a speech separation system that separates two mixed utterances, the correspondence between the output speaker assignments and the ground truths is indeterminate. The conventional approach for solving this issue is through utterance-level permutation invariant training (uPIT), which minimizes the total loss between the pairs of references and hypotheses by computing the optimal permutation between them. Although this strategy has been widely adopted for single-turn (2 utterance) mixtures, it quickly becomes intractable for longer meeting-like sessions containing arbitrarily many utterances. While extending SURT to multi-turn sessions [36], we proposed the use of heuristic error assignment training (HEAT), which mitigates this problem by fixing the assignment of references to output channels using utterance delays as a heuristic. However, this constraint unfairly penalizes models in the case when long pauses occur during/between utterances. Recently, the graph-PIT objective has been proposed for continuous speech separation, which solves this problem by generalizing PIT using graph coloring [41]. The SURT model may benefit from these advances by adapting the graph-PIT objective to be used with the underlying transducer loss.

The SURT models we have discussed so far are designed to process single-channel recordings. Since meetings are often recorded using array microphones — all 3 datasets under consideration (LibriCSS, AMI, AliMeeting) provide array microphone recordings — it is natural to ask how performance may be improved by incorporating ideas from the substantial research on multi-channel signal processing. In particular, the minimum variance distortionless response (MVDR) filter has been used to reduce distortions produced by neural network based spectral masking methods, which otherwise negatively impact ASR performance [53]. Recently, an all-neural MVDR method was proposed for target speech extraction [55], and has since been adapted for continuous speech separation [56]. This component may be plugged into SURT to improve the unmixed streams that are fed to the transducer.

The end-to-end models in the SURT framework are trained from scratch using simulated/real meeting-like sessions, and often require large amounts of data and compute to converge. This is

because the model has to simultaneously learn the basics of speech recognition, while also learning to separate and transcribe multi-talker mixtures. Recently, speech processing (like NLP) has greatly benefited from large, pre-trained, self-supervised models (SSL) such as Wav2Vec 2.0 [2] and HuBERT [20], with successful applications in tasks such as ASR, speech translation, and speaker recognition [44]. We propose to make SURT training easier by integrating these pre-trained models into the transducer component of the framework. Such an integration will involve careful interfacing between the unmixing and transducer components, since the former produces time-frequency (TF) masked signals, whereas SSL models are usually trained on raw waveforms. One approach to resolve this mismatch is by constructing the waveform from the STFT signals using differentiable Griffin-Lim estimates [17]. However, such a conversion may be computationally expensive (since it is required for each training sample) and contrary to the goal of our separation-free framework; therefore, we are also interested in tighter integration of SSL models into SURT in the frequency domain.

Status. The explorations described in this chapter have not begun yet, and are planned for Spring 2023, with possible submission to IEEE ICASSP 2024 and/or IEEE/ACM TASLP.

Relevance. This work will build upon the multi-talker SURT framework developed recently, and proposes several advances to bridge the gap between end-to-end and modular systems — including novel training objectives, extension to multi-channel processing, and integrating large self-supervised models, which have revolutionized single-speaker ASR. More broadly, it will help to provide a conclusion to the goal of this dissertation, while also laying the foundation for advancing the framework by incorporating ideas from adjacent speech procession research.

5 Timeline

The following is a tentative timeline of the completed and proposed work that will constitute this dissertation.

- **Modular frameworks** (§ 3)
  - *Simple pipeline approach* (§ 3.1): Published at CHiME-6 workshop at IEEE ICASSP 2020; investigation for meeting data ongoing (for TASLP submission).
  - *Overlap-aware diarization* (§ 3.2): Published at IEEE SLT 2021.
  - *Simultaneous approach using CSS* (§ 3.3): Published at IEEE SLT 2021; CSS-based diarization in preparation for IEEE SLT 2022; investigation of full system in progress (for TASLP submission).
- **End-to-end frameworks** (§ 4)
  - *Systems based on TS-ASR* (§ 4.1): Planned for IEEE ICASSP 2023.
  - *Systems based on SURT* (§ 4.2): Published at IEEE ICASSP 2022; further work in Fall/Winter 2022 (for InterSpeech 2023).
  - *Advances in SURT* (§ 4.3): Planned for Spring 2023.
- **Dissertation writing**: Summer/Fall 2023

References

[1] A. Arora, D. Raj, A. S. Subramanian, K. Li, B. Ben-Yair, M. Maciejewski, P. Zelasko, P. Garcia, S. Watanabe, and S. Khudanpur. The JHU multi-microphone multi-speaker ASR system for the CHiME-6 challenge. *ArXiv*, 2020.

[2] A. Baevski, H. Zhou, A. rahman Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*, 2020.

[3] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach. Front-end processing for the chime-5 dinner party scenario. *5th International Workshop on Speech Processing in Everyday Environments (CHiME 2018)*, 2018.

[4] L. Bullock, H. Bredin, and L. P. García-Perera. Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection. *IEEE ICASSP*, 2020.

[5] 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. L. Masson, I. McCowan, W. Post, D. Reidsma, and P. Wellner. The AMI meeting corpus: A pre-announcement. In *MLMI*, 2005.

[6] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *IEEE ICASSP*, 2016.

[7] X. Chen, Y. Wu, Z. Wang, S. Liu, and J. Li. Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. *IEEE ICASSP*, 2021.

[8] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, and J. Li. Continuous speech separation: Dataset and analysis. *IEEE ICASSP*, 2020.

[9] C.-C. Chiu, W. Han, Y. Zhang, R. Pang, S. Kishchenko, P. Nguyen, A. Narayanan, H. Liao, S. Zhang, A. Kannan, R. Prabhavalkar, Z. Chen, T. N. Sainath, and Y. Wu. A comparison of end-to-end models for long-form speech recognition. *IEEE ASRU*, 2019.

[10] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent. Librimix: An open-source dataset for generalizable speech separation. *ArXiv*, 2020.

[11] M. Delcroix, K. Žmolíková, K. Kinoshita, A. Ogawa, and T. Nakatani. Single channel target speaker extraction and recognition with speaker beam. *IEEE ICASSP*, 2018.

[12] J. G. Fiscus, J. Ajot, and J. S. Garofolo. The rich transcription 2007 meeting recognition evaluation. In *CLEAR*, 2007.

[13] Y. Fujita, S. Watanabe, S. Horiguchi, Y. Xue, and K. Nagamatsu. End-to-end neural diarization: Reformulating speaker diarization as simple multi-label classification. *ArXiv*, 2020.

[14] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree. Speaker diarization using deep neural network embeddings. *IEEE ICASSP*, 2017.

[15] A. Graves. Sequence transduction with recurrent neural networks. *ArXiv*, abs/1211.3711, 2012.

[16] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *ICML*, 2006.

[17] D. W. Griffin and J. S. Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32, 1984.

[18] C. Han, Y. Luo, C. Li, T. Zhou, K. Kinoshita, S. Watanabe, M. Delcroix, H. Erdogan, J. R. Hershey, N. Mesgarani, and Z. Chen. Continuous speech separation using speaker inventory for long recording. *INTERSPEECH*, 2021.

[19] M.-K. He, D. Raj, Z. Huang, J. Du, Z. Chen, and S. Watanabe. Target-speaker voice activity detection with improved i-vector estimation for unknown number of speaker. *INTERSPEECH*, 2021.

[20] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 2021.

[21] N. Kanda, X. Chang, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka. Investigation of end-to-end speaker-attributed ASR for continuous multi-talker recordings. *IEEE ICASSP*, 2021.

[22] N. Kanda, Y. Gaur, X. Wang, Z. Meng, Z. Chen, T. Zhou, and T. Yoshioka. Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers. *ArXiv*, 2020.

[23] N. Kanda, Y. Gaur, X. Wang, Z. Meng, and T. Yoshioka. Serialized output training for end-to-end overlapped speech recognition. *ArXiv*, 2020.

[24] N. Kanda, S. Horiguchi, Y. Fujita, Y. Xue, K. Nagamatsu, and S. Watanabe. Simultaneous speech recognition and speaker diarization for monaural dialogue recordings with target-speaker acoustic models. *IEEE ASRU*, 2019.

[25] A. Kannan, A. Datta, T. N. Sainath, E. Weinstein, B. Ramabhadran, Y. Wu, A. Bapna, Z. Chen, and S. Lee. Large-scale multilingual speech recognition with a streaming end-to-end model. *ArXiv*, 2019.

[26] F. Landini, J. Profant, M. Díez, and L. Burget. Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks. *Computer, Speech, and Language*, 2021.

[27] L. Lu, N. Kanda, J. Li, and Y. Gong. Streaming end-to-end multi-talker speech recognition. *IEEE Signal Processing Letters*, 28, 2021.

[28] Y. Luo, Z. Chen, and T. Yoshioka. Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation. *IEEE ICASSP*, 2020.

[29] M. Maciejewski, G. Sell, L. P. García-Perera, S. Watanabe, and S. Khudanpur. Building corpora for single-channel speech separation across multiple domains. *ArXiv*, 2018.

[30] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. D. Wellner. The AMI meeting corpus. 2005.

[31] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko. Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario. *INTERSPEECH*, 2020.

[32] D. Raj, P. Denisov, Z. Chen, H. Erdogan, Z. Huang, M.-K. He, S. Watanabe, J. Du, T. Yoshioka, Y. Luo, N. Kanda, J. Li, S. Wisdom, and J. Hershey. Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis. *IEEE SLT*, 2021.

[33] D. Raj, L. P. García-Perera, Z. Huang, S. Watanabe, D. Povey, A. Stolcke, and S. Khudanpur. DOVER-Lap: A method for combining overlap-aware diarization outputs. *IEEE SLT*, 2021.

[34] D. Raj, Z. Huang, and S. Khudanpur. Multi-class spectral clustering with overlaps for speaker diarization. *IEEE SLT*, 2021.

[35] D. Raj and S. Khudanpur. Reformulating DOVER-Lap label mapping as a graph partitioning problem. *INTERSPEECH*, 2021.

[36] D. Raj, L. Lu, Z. Chen, Y. Gaur, and J. Li. Continuous streaming multi-talker ASR with dual-path transducers. *IEEE ICASSP*, 2022.

[37] R. V. Rikhye, Q. Wang, Q. Liang, Y. He, and I. McGraw. Multi-user voicefilter-lite via attentive speaker embedding. *IEEE ASRU*, 2021.

[38] N. Ryant, K. W. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman. The second DIHARD diarization challenge: Dataset, task, and baselines. *ArXiv*, 2019.

[39] J. Shi, C. Zhang, C. Weng, S. Watanabe, M. Yu, and D. Yu. An investigation of neural uncertainty estimation for target speaker extraction equipped rnn transducer. *Computer, Speech, and Language*, 73, 2022.

[40] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. X-vectors: Robust DNN embeddings for speaker recognition. *IEEE ICASSP*, 2018.

[41] T. von Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach. Graph-pit: Generalized permutation invariant training for continuous separation of arbitrary numbers of speakers. *ArXiv*, abs/2107.14446, 2021.

[42] Q. Wang, H. Muckenhirn, K. W. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. Lopez-Moreno. Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking. *INTERSPEECH*, 2019.

[43] S. Watanabe, M. Mandel, J. Barker, and E. Vincent. CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. *ArXiv*, 2020.

[44] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T. hsien Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. rahman Mohamed, and H. yi Lee. Superb: Speech processing universal performance benchmark. *ArXiv*, 2021.

[45] J. Wu, Z. Chen, S. Chen, Y. Wu, T. Yoshioka, N. Kanda, S. Liu, and J. Li. Investigation of practical aspects of single channel speech separation for ASR. *INTERSPEECH*, 2021.

[46] W. Xia, H. Lu, Q. Wang, A. Tripathi, I. Lopez-Moreno, and H. Sak. Turn-to-diarize: Online speaker diarization constrained by transformer transducer speaker turn detection. *ArXiv*, 2021.

[47] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. Achieving human parity in conversational speech recognition. *ArXiv*, 2016.

[48] T. Yoshioka, D. Dimitriadis, A. Stolcke, W. Hinthorn, Z. Chen, M. Zeng, and X. Huang. Meeting transcription using asynchronous distant microphones. In *INTERSPEECH*, 2019.

[49] D. Yu, X. Chang, and Y. Qian. Recognizing multi-talker speech with permutation invariant training. *ArXiv*, 2017.

[50] 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. *ArXiv*, 2021.

[51] S. X. Yu and J. Shi. Multiclass spectral clustering. *IEEE ICCV*, 2003.

[52] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar. Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7829–7833, 2020.

[53] W. Zhang, C. Boeddeker, S. Watanabe, T. Nakatani, M. Delcroix, K. Kinoshita, T. Ochiai, N. Kamo, R. Haeb-Umbach, and Y. Qian. End-to-end dereverberation, beamforming, and speech recognition with improved numerical stability and advanced frontend. *IEEE ICASSP*, 2021.

[54] X. Zhang, V. Manohar, D. Zhang, F. Zhang, Y. Shi, N. Singhal, J. Chan, F. Peng, Y. Saraf, and M. L. Seltzer. On lattice-free boosted MMI training of HMM and CTC-based full-context ASR models. *IEEE ASRU*, 2021.

[55] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu. Adl-mvdr: All deep learning mvdr beamformer for target speech separation. *IEEE ICASSP*, 2021.

[56] Z. Zhang, T. Yoshioka, N. Kanda, Z. Chen, X. Wang, D. Wang, and S. E. Eskimez. All-neural beamformer for continuous speech separation. *ArXiv*, abs/2110.06428, 2021.

[57] P. Żelasko, D. Povey, J. Trmal, and S. Khudanpur. Lhotse: a speech data representation library for the modern deep learning ecosystem. *ArXiv*, abs/2110.12561, 2021.

[58] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. H. Cernocký. Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures. *IEEE Journal of Selected Topics in Signal Processing*, 13, 2019.

[59] K. Žmolíková, M. Delcroix, D. Raj, S. Watanabe, and J. H. Cernocký. Auxiliary loss function for target speech extraction and recognition with weak supervision based on speaker characteristics. *INTERSPEECH*, 2021.