INTEGRATION OF SPEECH SEPARATION, DIARIZATION, AND RECOGNITION FOR MULTI-SPEAKER MEETINGS: SYSTEM DESCRIPTION, COMPARISON, AND ANALYSIS

Desh Raj[1], Pavel Denisov[2], Zhuo Chen[3], Hakan Erdogan[4], Zili Huang[1], Maokui He[5], Shinji Watanabe[1], Jun Du[5], Takuya Yoshioka[3], Yi Luo[6], Naoyuki Kanda[3], Jinyu Li[3], Scott Wisdom[4], John R. Hershey[4]

[1]Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD
[2]Institute for Natural Language Processing, University of Stuttgart, Germany
[3]Microsoft Corp, Redmond, WA, [4]Google Research, Cambridge, MA
[5]University of Science and Technology of China, HeFei, China
Department of Electrical Engineering, Columbia University, NY

`draj@cs.jhu.edu`

ABSTRACT

Multi-speaker speech recognition of unsegmented recordings has diverse applications such as meeting transcription and automatic subtitle generation. With technical advances in systems dealing with speech separation, speaker diarization, and automatic speech recognition (ASR) in the last decade, it has become possible to build pipelines that achieve reasonable error rates on this task. In this paper, we propose an end-to-end modular system for the LibriCSS meeting data, which combines independently trained separation, diarization, and recognition components, in that order. We study the effect of different state-of-the-art methods at each stage of the pipeline, and report results using task-specific metrics like SDR and DER, as well as downstream WER. Experiments indicate that the problem of overlapping speech for diarization and ASR can be effectively mitigated with the presence of a well-trained separation module. Our best system achieves a speaker-attributed WER of 12.7%, which is close to that of a non-overlapping ASR.

Index Terms— Speech separation, diarization, speech recognition, multi-speaker

1. INTRODUCTION

Multi-speaker speech recognition is defined as a task wherein, given a long unsegmented recording consisting of a conversation between an unknown number of participants, the expected output contains "who said what and when." Although the last two decades have seen incredible leaps in speech technology — deep learning systems have matched human parity in the single-speaker conversational speech settings [3] — multi-speaker speech processing remains a challenge as such recordings may contain almost 20% overlapped speech [4, 5]. This scenario has implications for both diarization and automatic speech recognition (ASR) — single-speaker diarization systems miss the interfering speaker completely, and ASR systems trained on clean utterances are more error-prone on overlapped regions. These factors, along with the effect of far-field acoustic conditions, may increase the speaker-attributed word error rates by up to 86% [6] in meetings. To tackle such problems, recent editions of DIHARD [7] and CHiME [1] have focused on these tasks in very challenging settings.

A straightforward approach for solving the overlap problem in multi-speaker conversations is through speech separation. Deep learning based separation methods have been progressing consistently over the last few years — a signal-to-distortion ratio of 19.0 dB has been obtained on the popular WSJ0-2mix dataset [8], making the enhanced signals almost indistinguishable from clean utterances [9]. However, these techniques are often evaluated on short, fully overlapping (and often simulated) mixtures, and do not measure distortions introduced in overlap-free regions. They may also make unrealistic assumptions, such as prior knowledge of the number of speakers in the mixture. Recently, there have been efforts towards "continuous speech separation," which situates separation techniques in more realistic settings of long-form conversations containing partially overlapped speech [10, 2], such as those found in multi-speaker conversations.

There is increasing interest towards combining the advances made in speech separation, diarization, and ASR for tackling multi-speaker speech recognition. The CHiME-6 challenge [1] included a track aimed at recognizing unsegmented dinner-party conversations recorded on multiple microphone arrays, and baseline evaluations indicated that replacing oracle segments with diarization outputs could degrade downstream WERs by almost 52%. Chen et al. [2] proposed the new LibriCSS meeting dataset, and combined speaker-independent continuous speech separation [10] with a strong hybrid ASR [11] to recognize long unsegmented
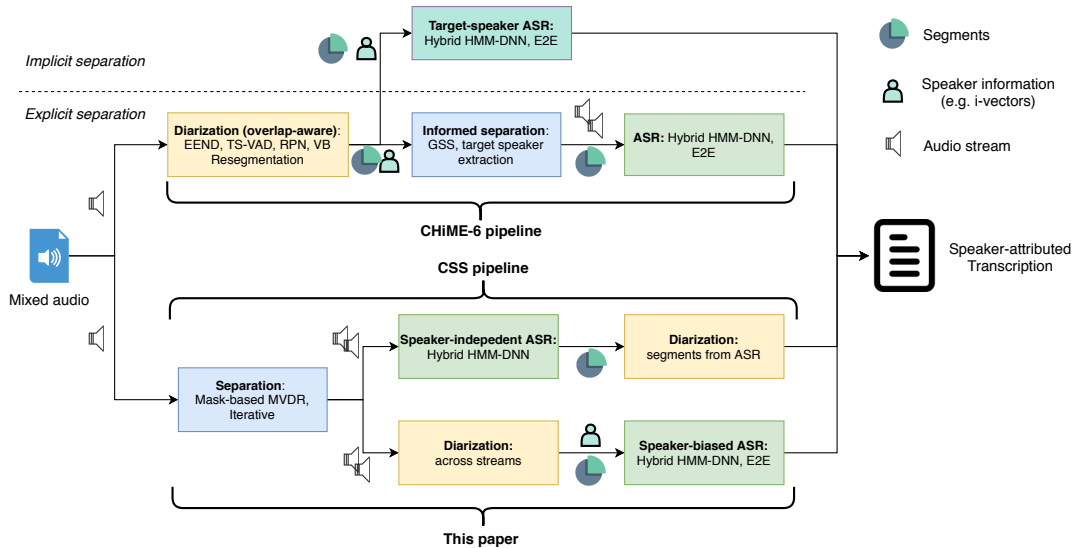
SLT 2021

Fig. 1: An overview of our proposed approach, compared with the CHiME-6 pipeline [1] and the CSS pipeline [2].

meeting recordings. Similar datasets such as LibriMix [12] have been created to satisfy the need for multi-speaker conversational data containing partial overlaps. Joint modeling of speaker counting, speaker identification, and ASR using end-to-end models has also been investigated [13, 14].

Nevertheless, the number of available methods to perform separation, diarization, or recognition are plenty — each with their own sets of requirements and assumptions. Together with the several possibilities of combining them, this creates a daunting task to analyze systems that recognize unsegmented multi-speaker recordings. In this paper, we make a first attempt at tackling this problem, using the LibriCSS dataset. We propose a novel end-to-end modular pipeline that performs separation, diarization, and recognition, in that order. For each of these modules, we compare several existing methods, describing the advantages and disadvantages of each alternative. We also demonstrate the impact of separation on downstream tasks by presenting corresponding results on the original mixed recording. Since the modules only interact at the I/O level, they can be implemented using different tools and optimized independently. Our best system with and without separation achieves a concatenated minimum-permutation WER (cpWER) [1] of 12.7% and 23.9%, respectively, on the LibriCSS evaluation set. We will release code to reproduce (and extend) our pipeline here: https://desh2608.github.io/pages/jsalt/.

2. SYSTEM OVERVIEW

Our pipeline consists of three components: Separation → Diarization → ASR. An unsegmented multichannel audio recording is provided as input to a continuous speech separation system, such as [10]. The separation module works on small windowed segments containing at most 2 or 3 speakers, and produces the corresponding number of separated audio

streams, which are passed to the diarization module. Since a speaker may have been split into different streams across different windows, the diarization is performed by considering all the audio streams simultaneously. We will discuss the implications of this requirement on different diarization methods in Section 4. After diarization, the single-speaker homogenenous segments are fed into an ASR decoder.

Fig. 1 shows our proposed approach, and situates it in comparison with alternative approaches proposed in [1] and [2], which we refer to as the *CHiME-6 pipeline* and the *CSS pipeline*, respectively. All 3 methods perform explicit separation, as opposed to an implicit "separation+recognition" usually performed by target-speaker ASR methods [15]. Stacking the components in different orders elicits unique advantages and limitations for each approach. Since diarization is performed first in the CHiME-6 pipeline, it makes it possible to use the source activity pattern derived from the diarization output to guide the estimation of the mixture model parameters in guided source separation (GSS) [16], or use the speaker information to perform target speaker extraction, as in SpeakerBeam [17]. However, poor diarization can have significant impacts on separation results in this pipeline. The CSS pipeline performs separation in the first stage, followed by ASR and diarization, allowing the system to be deployed in a streaming setting. The diarization performance may itself also benefit from reduced false alarms, since segments are obtained from the ASR output. However, it makes speaker-biasing infeasible at the ASR stage. Furthermore, the ASR output may contain spurious insertions resulting from cross-talk in non-overlap regions of the separated audio streams.

In contrast, our pipeline makes it feasible to employ a speaker-biased ASR, and also to remove extra insertions through simple post-processing of the diarization output (cf. Section 3.3). Unlike the CHiME-6 pipeline, our separation

898

is not dependent on the performance of the diarizer, and speaker-independent separation techniques can be effectively used in the pipeline. Here we demonstrate two of these — mask-based minimum variance distortionless response (MVDR), and sequential multi-frame separation. Furthermore, since the diarization stage sees only single-speaker recordings, it circumvents the need for an overlap-aware diarization component.

We will now describe the performance of different methods in each module – namely separation, diarization, and ASR — in Sections 3, 4, and 5, respectively. All our experiments are conducted on LibriCSS [2], which consists of multi-channel audio recordings of "simulated conversations," generated by mixing Librispeech utterances [18]. It comprises 10, approximately one-hour long, sessions. Each session is made up of six 10-minute-long "mini sessions" that have different overlap ratios, ranging from 0% to 40%. The recordings were made in a regular meeting room by using a seven-channel circular microphone array. We used *session 0* as the development set, and the remaining 9 sessions for evaluation (which we use to report results). Diarization results are reported using diarization error rate (DER), and ASR performance is evaluated in terms of cpWER. It is computed by concatenating all the utterances of a speaker in the reference and hypothesis, scoring all speaker pairs, and then finding the speaker permutation that minimizes the total WER.

3. SPEECH SEPARATION

3.1. Mask-based MVDR

The speech separation module follows the continuous speech separation scheme [19], consisting of three steps. First, the recording is uniformly segmented into smaller chunks with overlap. For each chunk, a multi-channel separation network [2], trained with permutation invariant training (PIT) criterion, is applied to estimate three time-frequency (TF) masks: two for speech sources and one for noise. In this scheme, we used a chunk size of 2.4s with 0.8s hop.

Thereafter, a stitching algorithm [20] is used to track local permutations and glue the masks from each chunk into the meeting-wise mask. This stitching is performed using mask similarity between adjacent chunks on the overlapped region (which is 1.6s in our case), by finding the permutation that has minimum distance between chunks. Based on the resulting permutation, the chunk-wise masks are connected to obtain a mask stream for the long recording. We averaged the overlap region between chunks for this process.

Finally, given the stitched masks for the entire meeting, a mask-based adaptive MVDR beamforming [21] is performed to get the final separation result. Since the noise is mostly stationary, we used the noise mask for the entire meeting to estimate the noise spatial covariance.

Note that the number of output channels in each local separation is highly correlated with the chunk size. As [21] suggests, most 2.4s chunks contains at most 2 speakers, therefore two output channels suffice for local processing. However, when the chunk size is larger (as in the next separation model), it may contain more than 2 speakers, even though LibriCSS contains at most 2-speaker overlaps.

3.2. Sequential multi-frame separation

We also used a novel sequential separation system which has multiple mask-based beamforming steps [22]. A mask-based multi-frame multi-channel Wiener filter (MCWF) beamformer is used. We used three untied sequential steps inside the model by feeding previously beamformed signals into the next mask-prediction step. The output of this model can be the last mask-network output or the one previous beamformed output. This model was trained using 10 second long mixtures of three speaker utterances from Libri-Light database [23]. During training, the utterances were mixed on-the-fly with room impulse responses obtained using an image-method based room simulator that simulates data acquisition in shoebox shaped rooms with an 8-microphone array [22]. This pre-trained model was used to separate sources in the LibriCSS dev and eval datasets using 8 second long blocks with 4 seconds overlap. Since LibriCSS has 7 microphones, we added another microphone signal by shifting the first microphone signal with one sample and adding white Gaussian noise with variance 1e-6. The source estimate outputs were stitched using a stitching loss of magnitude STFT domain mean-squared error between time-domain signals to resolve the permutation across blocks. This is similar to the stitching method described in Section 3.1.

3.3. Experimental results

Table 1 shows the performance of the separation methods on the LibriCSS eval set. Separation performance is calculated on a simulated eval set different from LibriCSS since reference signals are required. Since the models work with varying window sizes, we first mapped the output tracks to the N source tracks for every 0.8s block[1], where N is the number of participating speakers, using magnitude-domain distance and assuming at most two active speakers per 0.8s block, and then generated N speaker tracks from the separated outputs. We call this process "oracle track mapping". We report separation performance in terms of average meeting-level SDR [24]. Sequential multi-frame model achieved a higher SDR of 14.1 dB as compared to the mask-based MVDR's 5.8 dB, since implemented MVDR beamformer does not attempt to reconstruct target signals exactly.

We also evaluated the separated audio with a spectral clustering based diarizer (Section 4.1) and a hybrid HMM-DNN ASR (Section 5.1), and the results are shown in the table in terms of DER and cpWER, respectively. From the table, we note that the diarization performance of the two methods were

---

[1]0.8s is the highest common factor between the chunk size of the models.

Table 1: Performance of separation methods on LibriCSS eval set in terms of resulting downstream diarization (using spectral clustering) and ASR (using TDNN-F model) results. For comparison, we also show results obtained on a "no separation" baseline. Separation performance is reported on a simulated eval set. †MVDR beamformer does not attempt to reconstruct the target signal at the reference microphone directly, so the separation metric reflects this mismatch.

Method	SDR (dB)	DER (%)	cpWER (%)
No separation	-	18.3	31.0
Mask-based MVDR	5.8†	**13.9**	22.8
Sequential multi-frame	**14.1**	14.1	**19.3**

comparable, but the 3-stream sequential multi-frame separation outperformed the 2-stream mask-based MVDR in terms of cpWER results. Although the sequential model was trained with an 8-microphone cubic geometry, it generalized well to LibriCSS, which has a completely different geometry, since the model is inherently geometry-independent.

Furthermore, a simple post-processing trick applied on the diarization output was found to be useful for the sequential model – using this trick improved the cpWER by 15.4% relative. For this post-processing, we removed a segment from a stream if it was completely enclosed within a same-speaker segment in a different stream. This is akin to filtering out cross-talk prior to ASR decoding. This trick was not useful for the mask-based MVDR, providing only a marginal improvement of 0.2% relative in terms of cpWER. Note that although sequential model achieved significantly better SDR, their cpWER are similar. This indicates the potential objective mismatch between speech separation and recognition.

4. SPEAKER DIARIZATION

We can categorize diarization methods based on whether or not they can assign overlapping speaker segments. Since our pipeline separates the recording prior to diarization, overlap-awareness is not a strict requirement. At the same time, since the separation is done on windowed segments, the diarization needs to be performed across audio streams (since a speaker can be present in different streams at different times). Based on these conditions, we selected the following diarization methods for our study.

4.1. X-vector + clustering

This method consists of a speech activity detection (SAD) component followed by clustering of small subsegment embeddings. We used a similar SAD as that described in [1], consisting of a TDNN-Stats based classifier with Viterbi decoding for inference. The speech segments were divided into subsegments with a window size of 1.5s and a stride of 0.75s, and 128-dimensional embeddings were extracted using an x-vector extractor [25] trained on VoxCeleb data [26] with simulated room impulse response [27]. We conducted experi-

ments with 3 variants of clustering: (i) agglomerative hierarchical clustering (AHC) on PLDA scores [28], (ii) spectral clustering (SC) on cosine similarity [29], and (iii) VBx clustering initialized from the AHC system [30]. For these methods, we did not make any assumptions about the number of speakers in the recording. We used the same PLDA trained on Librispeech for both AHC and VBx, and did not use the PLDA interpolation technique from [30].

4.2. Region proposal networks (RPN)

This is a supervised method, which combines the segmentation and embedding extraction steps into a single neural network and jointly optimizes them [31]. The region embeddings are then clustered (using K-means clustering) and a non-maximal suppression is applied. We trained the RPN on simulated meeting-style recordings with partial overlaps generated using utterances from the Librispeech [18] training set. Since we used K-means clustering, we assumed that the oracle number of speakers for each recording is known.

4.3. Target-speaker voice activity detection (TS-VAD)

The TS-VAD model takes conventional speech features (e.g., MFCC) along with i-vectors for each speaker as inputs and produces frame-level activities for each speaker using a neural network with a set of binary classification output layers [32]. Since the number of these binary output nodes are fixed for training, we assumed that the maximum possible number of speakers in any session is at most 8. The initial estimates for the speaker i-vectors were obtained using the SC system. For training, we created simulated meeting-style data similar to that used for training the RPN model.

4.4. Experimental results

We conducted experiments with "mixed" as well as "separated" audio to analyze the impact of separation on diarization performance. For the mixed recording, we selected the first channel as our input. For evaluation on separated streams, we fixed the separation component as mask-based MVDR, and the ASR was chosen to be a speaker-biased hybrid HMM-DNN (described in Section 5.1). Table 2 shows the diarization performance on mixed LibriCSS, with a breakdown by overlap condition. The SAD error for the clustering-based systems was 4.8%. It is immediately evident that assigning overlapping speech is important to perform well on this task – both RPN and TS-VAD outperformed clustering-based methods. Even on low overlap regions, there is a significant difference, and we conjecture that this arises from a mismatch between the training data for the PLDA used for scoring the AHC and VBx models, and the evaluation set. Since SC uses cosine scoring, it performed better than the other clustering methods. For the RPN and TS-VAD systems, creating a simulated mixture which closely resembled the test set was found to be important, and using cepstral mean normalization (CMN) was critical for this performance.

Table 2: Diarization performance on mixed LibriCSS evaluation set, in terms of % DER. 0S and 0L refer to 0% overlap with short and long inter-utterance silences, respectively. RPN and TS-VAD methods assume that oracle number of speakers is known. We also report the corresponding cpWER using a TDNN-F based ASR model. Using an oracle diarization output results in a cpWER of 23.1%.

Method	Overlap ratio in %						DER	cpWER
	0L	0S	10	20	30	40		
AHC	16.1	12.0	16.9	23.6	28.3	33.2	22.6	36.7
VBx	14.6	11.1	14.3	21.5	25.4	31.2	20.5	33.4
SC	10.9	9.5	13.9	18.9	23.7	27.4	18.3	31.0
RPN	**4.5**	9.1	8.3	**6.7**	11.6	14.2	9.5	27.2
TS-VAD	6.0	**4.6**	**6.6**	7.3	**10.3**	9.5	**7.6**	**24.4**

Table 3: Comparison of diarization and downstream ASR performance of different methods on separated audio streams for LibriCSS evaluation set. We used mask-based MVDR for separation and TDNN-F based ASR for these experiments.

Separation	Metric	AHC	VBx	SC	RPN	TS-VAD
✗	**DER**	22.6	20.5	18.3	9.5	**7.6**
✗	**cpWER**	36.7	33.4	31.0	27.2	**24.4**
✓	**DER**	37.5	-	**13.9**	22.4	31.7
✓	**cpWER**	86.7	-	**22.8**	34.9	47.7

Next, we evaluated the systems on separated audio streams obtained using the mask-based MVDR method. Table 3 shows these results, along with the downstream cpWER obtained using a hybrid HMM-DNN ASR model. Since AHC and SC perform subsegment-level clustering, they can be naturally extended to diarization across streams. VBx, on the other hand, estimates speaker changes through HMM state transitions, so it is not directly applicable to this scenario. RPN and TS-VAD can also be extended to this new setting, since RPN uses clustering of region embeddings, and TS-VAD predicts frame-level speaker activity based on the corresponding i-vectors. Our first observation is that among clustering-based methods, SC performed significantly better than AHC, likely because the PLDA used for AHC was trained on clean Librispeech utterances, which are acoustically very different from the separated audio. To verify this, we performed the AHC also on a cosine similarity matrix, and it resulted in an absolute DER improvement of 13.8%. For RPN and TS-VAD, performance on mixed recording was not indicative of results on separated audio. Without any post-processing, we obtained a DER of 26.9% using RPN. This improved to 22.4% on filtering out non-speech segments using our SAD from Section 4.1. Similarly, TS-VAD performance degraded severely on going from mixed to separated audio. We found this degradation to be consistent for missed speech, false alarms, and speaker confusions, indicating a likely mismatch in train vs. test conditions. Consequently, the cpWER for both RPN and TS-VAD was found to be significantly higher than that for SC.

Table 4: ASR performance on mixed LibriCSS evaluation set, in terms of % WER. These results correspond to the "utterance-wise evaluation" of [2], since we used oracle segments. The TDNN-F and Transformer E2E models obtain WERs of 8.8% and 5.5% (beam size 60), respectively, on the Librispeech test-other evaluation set.

Model	Overlap ratio in %						Average
	0L	0S	10	20	30	40	
TDNN-F (*base*)	16.1	16.0	27.7	39.1	49.4	58.3	36.8
+ *fine-tuned*	11.4	11.4	18.2	26.3	33.7	40.7	25.2
Transformer E2E (beam = 5)	5.5	5.6	13.1	21.6	31.1	41.6	21.6
Transformer E2E (beam = 30)	**5.1**	**5.3**	**11.4**	**19.5**	**28.5**	**38.3**	**19.7**

5. SPEECH RECOGNITION

We conducted our ASR experiments on a hybrid TDNN-F based, and a Transformer-based end-to-end ASR model. Pre-trained models (trained on Librispeech) for both of these are publicly available, enabling reproducibility of our results.

5.1. TDNN-F based hybrid HMM-DNN

Following the Kaldi [33] Librispeech recipe, we trained a 17-layer deep neural network consisting of factored TDNN layers [34] using the lattice-free MMI objective [35]. We used 40-dim MFCC features, and additionally appended 100-dim i-vectors estimated online. The model was trained on the 960h Librispeech data with 3x speed perturbation. We call this our *base* model. On the Librispeech test-clean and test-other evaluation sets, this model obtains WERs of 3.8% and 8.8%, respectively. We additionally fine-tuned it for 1 epoch on Librispeech train set augmented with simulated room impulse responses [27] to match the acoustic conditions of mixed LibriCSS recordings; this is referred to as the *fine-tuned* model. For decoding on LibriCSS, we used the 3-gram language model provided with the Librispeech data, and the lattices were rescored using a pruned TDNN-LSTM based RNNLM trained on the training transcripts [36]. We used a 2-pass decoding strategy, where the i-vectors were re-estimated from the non-silence regions for the second pass [37]. The fine-tuned model was used to evaluate the downstream ASR performance of the separation and diarization methods.

5.2. Transformer-based end-to-end ASR

Our end-to-end (E2E) ASR is a state-of-the-art ESPNet-based [38] Transformer encoder-decoder model [39]. It was trained on the 960h Librispeech corpus with SpecAugment [40], using 83-dim log-mel filterbank features with pitch. The encoder consists of 2 convolutional layers and 12 self-attention blocks, and the decoder contains 6 self-attention blocks. The training loss jointly minimizes sequence-to-sequence (S2S) and connectionist temporal classification (CTC) objectives [41]. The decoder predicts subword units generated using SentencePiece [42]. For decoding, we used beam search which combines scores from the S2S, CTC, and an external Transformer-based language model. This

901

Table 5: ASR performance on separated LibriCSS evaluation set, in terms of % cpWER. These results correspond to the "continuous-input evaluation" of [2]. Mask-based MVDR and spectral clustering were used for separation and diarization, respectively.

Model	Without separation (DER = 18.3%)							With separation (DER = 13.9%)						
	0L	0S	OV10	OV20	OV30	OV40	Average	0L	0S	OV10	OV20	OV30	OV40	Average
TDNN-F	17.6	19.1	24.4	33.0	38.4	45.2	31.0	16.9	15.5	18.8	22.7	26.6	29.4	22.3
Transformer	12.4	14.1	20.2	29.5	35.3	41.9	27.1	11.2	8.5	10.6	14.8	15.5	17.5	13.4

ASR model obtains WERs of 2.2% and 5.5% on the Librispeech `test-clean` and `test-other` evaluation sets, respectively, using a beam size of 60.

5.3. Experimental results

Similar to our diarization experiments, we evaluated the ASR models on mixed and separated audio. Table 4 shows the results obtained on the mixed LibriCSS data. For these experiments, we used oracle segments and speaker information. From the table, we see that a strong Librispeech model also performed well on mixed LibriCSS utterances. Among the TDNNF-based hybrid models, fine tuning on reverberated data provided a 31.5% relative WER improvement. For transformer-based E2E models, decoding with larger beams (of size 30) improved WER by almost 10% relative, compared to decoding with smaller beam sizes. We did not get any significant gains by increasing beam sizes further, so we used this setting for further experiments.

Next, we present ASR results in the context of our pipeline, i.e., using separated audio streams from the mask-based MVDR model and segments from spectral clustering based diarization. We used the *fine-tuned* variant of the hybrid HMM-DNN ASR model. Additionally, to emphasize the importance of the separation module, we also show cpWERs obtained on mixed recordings. The results are shown in Table 5. On replacing oracle segments with those obtained from diarization, the downstream cpWERs for hybrid and E2E ASR systems degraded by 23.4% and 37.6%, respectively. Interestingly, this increase was more prominent in the low and medium overlap conditions, which suggests that errors in these cases occured primarily from incorrect speaker assignment. On applying separation, the cpWERs improved significantly (although this was partly due to better diarization). In particular, we found that the E2E model benefited more from separated audio streams, with its cpWER improving from 27.1% to 13.4%. For both the models, the cpWER with separation was found to outperform the corresponding results on mixed recording even using oracle segments.

Finally, with all our experimental evaluations in place, we combined the best performing models at each stage of the pipeline. We prepared two variants: (a) without separation, and (b) with separation. For (a), we selected TS-VAD based diarization and the Transformer-based ASR model. Pipeline (b) consists of sequential multi-frame separation, followed by SC-based diarization and the same ASR. We present the final

Table 6: Pipeline variants: (a) without separation, and (b) with separation, combining best models at each stage.

	Separation	Diarization	ASR	DER(%)	cpWER(%)
(a)	-	TS-VAD	Transformer	**7.6**	23.9
(b)	Sequential	SC	Transformer	14.1	**12.7**

results for both variants in Table 6. It is evident that in the absence of an explicit separation module, even a well performing diarization and ASR combo is hamstrung. We discuss some more implications of these results in the next section.

6. DISCUSSION

Our extensive experiments using a variety of separation, diarization, and ASR methods elicit important lessons for solving the multi-speaker speech recognition problem. First, the importance of explicit separation cannot be overstated — the best cpWER for a pipeline with separation is 46.9% relative better than one without it (Table 6). For separation, we found training and inference with larger chunks to perform better; there is a caveat, however — the improved performance is obtained through diarization tricks that can filter out the increased cross-talk with these models. Second, we observed that although new (supervised) diarization methods like RPN and TS-VAD provide substantial gains on mixed recordings, the performance does not carry over well to separated audio streams, and traditional clustering approaches can still outperform them in these settings (Table 3). In particular, the best diarization without separation was found to be 45.8% relative better than one with separation, which is highly counterintuitive. This is particularly relevant for pipelines such as ours and the CSS pipeline, where separation is performed before diarization, and we regard this as an important direction for further investigation. There have also been concurrent efforts to apply diarization on separated audio streams [43]. Finally, unlike diarization, ASR performance on mixed recordings was strongly indicative of the performance on separated audio streams. We found that a state-of-the-art ASR model trained on clean, single speaker utterances integrates well in the pipeline and results in the best available cpWER on this dataset. This is encouraging, especially in view of recent advances in end-to-end speech recognition.

7. REFERENCES

[1] Shinji Watanabe, Michael Mandel, Jon Barker, and Emmanuel Vincent, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *ArXiv*, vol. abs/2004.09249, 2020.

[2] Zhuo Chen, Takuya Yoshioka, Liang Lu, Tianyan Zhou, Zhong Meng, Yi Luo, J. Wu, and Jinyu Li, "Continuous speech separation: Dataset and analysis," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7284–7288, 2020.

[3] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig, "Achieving human parity in conversational speech recognition," *ArXiv*, vol. abs/1610.05256, 2016.

[4] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner, "The AMI meeting corpus: A pre-announcement," in *MLMI*, 2005.

[5] Elizabeth Shriberg, Andreas Stolcke, and Don Baron, "Observations on overlap: findings and implications for automatic processing of multi-party conversation," in *INTERSPEECH*, 2001.

[6] Takuya Yoshioka, Dimitrios Dimitriadis, Andreas Stolcke, William Hinthorn, Zhuo Chen, Michael Zeng, and Xuedong Huang, "Meeting transcription using asynchronous distant microphones," in *INTERSPEECH*, 2019.

[7] Neville Ryant, Kenneth Ward Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman, "The second DIHARD diarization challenge: Dataset, task, and baselines," *ArXiv*, vol. abs/1906.07839, 2019.

[8] Yi Luo, Zhuo Chen, and Takuya Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 46–50, 2020.

[9] Yi Luo and Nima Mesgarani, "Conv-TasNet: Surpassing ideal time?frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 1256–1266, 2019.

[10] Takuya Yoshioka, Zhuo Chen, Changliang Liu, Xiong Xiao, Hakan Erdogan, and Dimitrios Dimitriadis, "Low-latency speaker-independent continuous speech separation," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6980–6984, 2019.

[11] Liang Lu, Xiong Xiao, Zhuo Chen, and Yifan Gong, "Pykaldi2: Yet another speech toolkit based on Kaldi and Pytorch," *ArXiv*, vol. abs/1907.05955, 2019.

[12] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent, "Librimix: An open-source dataset for generalizable speech separation.," *arXiv: Audio and Speech Processing*, 2020.

[13] Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, Tianyan Zhou, and Takuya Yoshioka, "Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers," *ArXiv*, vol. abs/2006.10930, 2020.

[14] Naoyuki Kanda, Xuankai Chang, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka, "Investigation of end-to-end speaker-attributed ASR for continuous multi-talker recordings," *ArXiv*, vol. abs/2008.04546, 2020.

[15] Marc Delcroix, Kateřina Žmolíková, Keisuke Kinoshita, Atsunori Ogawa, and Tomohiro Nakatani, "Single channel target speaker extraction and recognition with speaker beam," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5554–5558, 2018.

[16] Christoph Böddeker, Jens Heitkaemper, Joerg Schmalenstroeer, Lukas Drude, Jahn Heymann, and Reinhold Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *INTERSPEECH 2018*, 2018.

[17] Kateřina Žmolíková, Marc Delcroix, Keisuke Kinoshita, Tsubasa Ochiai, Tomohiro Nakatani, Lukás Burget, and Jan Cernocký, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, pp. 800–814, 2019.

[18] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[19] Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, and Fil Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5739–5743, 2018.

[20] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[21] Takuya Yoshioka, Igor Abramovski, Cem Aksoylar, Zhuo Chen, Moshe Ben David, Dimitrios Dimitriadis, Yifan Gong, Ilya Gurvich, Xuedong Huang, Yanping Huang, Aviv Hurvitz, Li Jiang, Sharon Koubi, Eyal Krupka, Ido Leichter, Changliang Liu, Partha Parthasarathy, Alon Vinnikov, Lingfeng Wu, Xiong Xiao, Wayne Xiong, Huaming Wang, Zhenghao Wang, Jingjing Zhang, Yong Zhao, and Tianyan Zhou, "Advances in online audio-visual meeting transcription," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 276–283, 2019.

[22] Zhong-Qiu Wang, Hakan Erdogan, Scott Wisdom, Kevin Wilson, Desh Raj, Shinji Watanabe, Zhuo Chen, and John R. Hershey, "Sequential multi-frame neural beamforming for speech separation and enhancement," in *IEEE Spoken Language Technology (SLT) Workshop*, 2021.

[23] Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien

Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, Tatiana Likhomanenko, Gabriel Synnaeve, Armand Joulin, Abdel rahman Mohamed, and Emmanuel Dupoux, "Libri-light: A benchmark for asr with limited or no supervision," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7669–7673, 2020.

[24] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[25] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018.

[26] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: A large-scale speaker identification dataset," *ArXiv*, vol. abs/1706.08612, 2017.

[27] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220–5224, 2017.

[28] Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree, "Speaker diarization using deep neural network embeddings," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4930–4934, 2017.

[29] T. Park, Kyu J. Han, Manoj Kumar, and Shrikanth S. Narayanan, "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap," *IEEE Signal Processing Letters*, vol. 27, pp. 381–385, 2020.

[30] Mireia Diez, Lukás Burget, Shuai Wang, Johan Rohdin, and Jan Cernocký, "Bayesian HMM based x-vector clustering for speaker diarization," in *INTERSPEECH*, 2019.

[31] Zili Huang, Shinji Watanabe, Yusuke Fujita, Paola García, Yiwen Shao, Daniel Povey, and Sanjeev Khudanpur, "Speaker diarization with region proposal network," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6514–6518, 2020.

[32] Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri Y. Khokhlov, Mariya Korenevskaya, Ivan Sorokin, Tatiana V. Timofeeva, Anton Mitrofanov, Andrei Andrusenko, Ivan Podluzhny, Aleksandr Laptev, and Aleksei Romanenko, "Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario," *ArXiv*, vol. abs/2005.07272, 2020.

[33] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukás Burget, Ondrej Glembek, Nagendra Kumar Goel, Mirko Hannemann, Petr Motlícek, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý, "The Kaldi speech recognition toolkit," 2011.

[34] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *INTERSPEECH*, 2018.

[35] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *INTERSPEECH*, 2016.

[36] Hainan Xu, Tongfei Chen, Dongji Gao, Yiming Wang, Ke Li, Nagendra Goel, Yishay Carmiel, Daniel Povey, and Sanjeev Khudanpur, "A pruned RNNLM lattice-rescoring algorithm for automatic speech recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5929–5933, 2018.

[37] Vimal Manohar, Szu-Jui Chen, Zhiqi Wang, Yusuke Fujita, Shinji Watanabe, and Sanjeev Khudanpur, "Acoustic modeling for overlapping speech recognition: JHU CHiME-5 challenge system," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6665–6669, 2019.

[38] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "Espnet: End-to-end speech processing toolkit," *ArXiv*, vol. abs/1804.00015, 2018.

[39] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al., "A comparative study on transformer vs RNN in speech applications," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 449–456.

[40] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *INTERSPEECH*, 2019.

[41] Suyoun Kim, Takaaki Hori, and Shinji Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4835–4839, 2017.

[42] Taku Kudo and John Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *EMNLP*, 2018.

[43] Xiong Xiao, Naoyuki Kanda, Z. Chen, Tianyan Zhou, T. Yoshioka, Sanyuan Chen, Yong Zhao, Gang Liu, Y. Wu, J. Wu, Shujie Liu, Jinyu Li, and Yifan Gong, "Microsoft speaker diarization system for the voxceleb speaker recognition challenge 2020," *ArXiv*, vol. abs/2010.11458, 2020.