OVERLAPPED SPEECH DETECTION FOR IMPROVED SPEAKER DIARIZATION IN MULTIPARTY MEETINGS

Kofi Boakye[1], *Beatriz Trueba-Hornero*[1,2], *Oriol Vinyals*[1], *Gerald Friedland*[1]

[1]International Computer Science Institute, Berkeley, CA, U.S.A.
[2]Polytechnic University of Catalonia, Barcelona, Spain

ABSTRACT

State-of-the-art speaker diarization systems for meetings are now at a point where overlapped speech contributes significantly to the errors made by the system. However, little if no work has yet been done on detecting overlapped speech. We present our initial work toward developing an overlap detection system for improved meeting diarization. We investigate various features, with a focus on high-precision performance for use in the detector, and examine performance results on a subset of the AMI Meeting Corpus. For the high-quality signal case of a single mixed-headset channel signal, we demonstrate a relative improvement of about 7.4% DER over the baseline diarization system, while for the more challenging case of the single far-field channel signal relative improvement is 3.6%. We also outline steps towards improvement and moving beyond this initial phase.

Index Terms— speaker diarization, overlap detection

1. INTRODUCTION

The presence of overlapped, or co-channel, speech in meetings is a common occurrence and a natural consequence of the spontaneous multiparty conversations which arise within these meetings. This speech, in addition, presents a significant challenge to automatic systems that process audio data from meetings, such as speech recognition and speaker diarization systems. In the case of speaker diarization, current state-of-the-art systems assign speech segments to only one speaker, thus incurring missed speech errors in regions where more than one speaker is active. For these systems, such as our own ICSI Diarization System [1], this error may represent a significant portion of the diarization error. For example, in previous RT diarization evaluations, up to 43% relative of the ICSI system diarization error consisted of missed speech errors due to overlap.

To be certain, it is only recently that diarization error rates of systems have been reduced to the point that a large portion of the remaining error is due to overlap. As a result, little work has been done on addressing the issues posed by the phenomenon. Some studies have been reported about the effects of overlap in meetings (e.g.,[2],[3],and [4]), but work on systems for identifying overlapped speech and mitigating its effects in speaker diarization appear to be absent from the literature. As overlapped speech is now a major obstacle in improving the performance of speaker diarization systems, efforts in overlap detection will be of increasing interest and importance.

With this view, we present in this paper our initial efforts toward addressing overlapped speech in automatic speaker diarization. This consists of an overlap detection system along with a segment post-processing procedure for the segmentation generated by the speaker diarization system. The overlap detector is an HMM-based segmenter that operates using features tailored for the task while the post-processing procedure is a speaker assignment method for the identified overlap segments based on speaker posterior probabilities produced by the diarization system.

As with any detection scheme, the overlap system is susceptible to errors of two types: false alarms and misses. These errors impact the diarization system quite differently, with false alarms carrying through to increase the diarization false alarm error and misses having no effect on the baseline diarization error. Because of this difference, the overlap detector is optimized for low false alarms, which corresponds to a high precision (and possibly low recall) operating point.

The remainder of this paper is organized as follows. The diarization system is briefly described in Section 2 and the HMM-based segmenter along with the segmenter features are described in Section 3. The diarization segment post-processing procedure is detailed in Section 4 and we present results on AMI development data in section 5. Finally, conclusions and future work are given in Section 6.

2. THE ICSI DIARIZATION SYSTEM

The goal of speaker diarization is to segment audio into speaker-homogeneous regions, ultimately to answer the question, “Who spoke when?”. In the ICSI diarization system, as with most state-of-the-art systems, this is accomplished through agglomerative clustering of segments with merging

This work was partly supported by the Swiss National Science Foundation through the research network IM2 and the European Union 6th FWP IST Integrated Project AMIDA.

based on Bayesian Information Criterion (BIC) scores. These scores are computed using GMMs of frame-based cepstral features (MFCCs). The clustering approach starts with a large number of initial clusters and proceeds by an iterative procedure of merging, model re-training and re-alignment. In the merging step, a BIC-based merge score is calculated between each two candidates. This measurement is then used to determine which two clusters should be merged or whether the merge should terminate. One major innovation of the system is the elimination of the tunable parameter in this merging procedure by ensuring that, for any given BIC comparison, the difference between the number of free parameters in the two hypotheses is zero. The system is described fully in [1].

System performance is measured using the diarization error rate (DER). This is defined as the sum of the false alarm (falsely identifying speech), missed speech (failing to identify speech), and speaker error (incorrectly identifying the speaker) times, divided by the total amount of speech time in a test audio file. Because the system presently can assign only a single speaker label to a segment, missed speech errors from speaker overlap persist and cannot be reduced. And since these errors presently constitute a substantial portion of the diarization error, overlap detection is an important next step in improving system performance.

3. HMM-BASED OVERLAP SEGMENTER

3.1. HMM architecture

To detect overlapped speech, we use an HMM-based overlap segmenter. The segmenter consists of three classes—nonspeech, speech, and overlapped speech—each being represented with a three-state model. State emission probabilities are modeled using a multivariate Gaussian Mixture Model (GMM) with 32 components and diagonal covariance matrices. For each class HMM, mixtures are shared between the three states, with separate mixture weights for each state.

3.2. Training

The class GMMs are trained using an iterative Gaussian splitting technique with successive re-estimation. The training starts with a single Gaussian and doubles the number of Gaussians at each iteration until the final mixture of 32 is obtained. Model re-estimation occurs at the end of each iteration. Speech, nonspeech, and overlap regions are identified in the training data using ASR forced-alignment times generated from ground-truth transcriptions of the audio.

3.3. Testing

Test audio signals are segmented into regions labeled as one of the three classes using a single Viterbi decoding pass of the full channel waveform. The speech and nonspeech classes are then considered a single "non-overlap" class and the overlap regions obtained are scored against reference overlap regions (again identified using forced-alignment). To measure segmentation performance in isolation (i.e., independent of improvements to the diarization system) we use precision, recall, and F-score values computed based on false alarm, missed detection, and total overlapped speech times.

As previously stated, we desire a low false alarm rate, and thus high precision, overlap detection system. This is achieved by adjusting the transition penalty from the speech to the overlap class in the Viterbi decoding. The penalty is determined by a parameter which is tuned using held-out data.

3.4. Overlap Detection Features

A key consideration in the overlap detection system is the selection of features used in the HMM-based segmenter. We have explored about 40 features (prosodic, short-term, long-term, etc.) and list below those that yielded the best performance in our experiments. The features were computed over sliding windows (with window sizes stated below) advanced by 20 ms.

Baseline MFCCs
The baseline features used consist of 12th-order Mel-frequency cepstral coefficients (MFCCs) along with first differences. Cepstral mean subtraction (CMS) is performed as a waveform-level normalization. MFCCs are common to various speech-related tasks (speech recognition, speaker recognition, speaker diarization, etc.) and as such served as a natural baseline feature for the system. The MFCCs were computed over a window of 60 ms.

RMS energy (Eg)
The energy content of a speech segment will likely be affected by the presence of additional speakers; specifically, we anticipate that overlapped speech will have a higher energy content than single-speaker speech in general. The short-time root-mean-squared (RMS) energy was computed over a window of 20 ms. To compensate for potential channel gain differences, signal waveforms were normalized based on overall RMS channel energy estimates.

LPC residual energy (LPC)
The linear predictive coding (LPC) coefficients of a speech signal encode the formants of a speaker while the residual signal represents the portion of the speech signal that cannot be attributed to this formant model—typically the excitation source. In the case of more than one speaker, a fixed-order LPC representation will not be able to model the spectrum (shaped by formants of multiple speakers) well. This potentially leads to more energy content in the residual signal. 12th-order LPC residual energy values were computed over a window of 25 ms.

Diarization posterior entropy (DPE)
Using frame-level speaker likelihoods from the diarization system, we compute the posterior probability for each speaker on every frame and subsequently a frame-level entropy from these posteriors. To reduce the effects of noise in the posterior values, we filter these probabilities with a 500 ms Hamming window. In single-speaker regions we expect one model to have the highest probability and the remainder to have significantly lower values. In overlap segments, however, there should be lower, more evenly distributed probabilities among the overlapping speakers and, as a result, the entropy should be higher.

4. DIARIZATION SEGMENT POST-PROCESSING

Having identified regions of overlapped speech, this information can then be used to modify segment and label information output by the diarization system. The procedure is as follows. In an overlapped segment, the frame-level speaker posteriors mentioned in Section 3.4 are summed over the frames of the segment to obtain a single "score" for each speaker. Typically the diarization system will have assigned the segment to the speaker with the highest score, in which case the speaker with the second highest score is chosen as the other speaker. In the event that the system has chosen another speaker, then this highest scoring speaker is selected as the additional speaker. Note that this procedure limits the number of possible overlapping speakers to two, but that two-speaker overlap typically comprises 80% or more of the instances of overlapped speech. A diagram of the final system is shown in Figure 1.

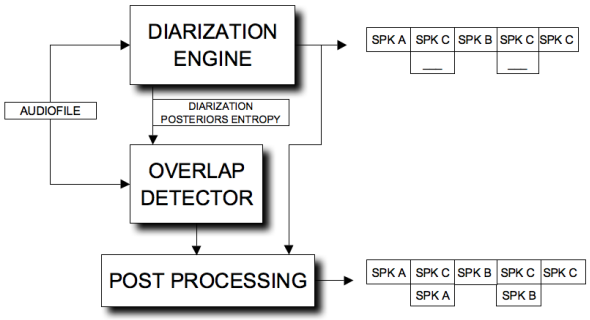

Fig. 1. *Diagram of integrated overlap detector and diarization system.*

5. RESULTS ON AMI DEVELOPMENT SET

Experiments evaluating the overlap detection system were performed using audio data from the AMI Meeting Corpus. The data consisted of 16 kHz-sampled single-channel signals, one per meeting, from the IDIAP subset ('IS' meetings) of the corpus. This subset comprises 38 meetings, each involving four participants engaged in a scenario-based meeting ranging in duration from 13 to 40 minutes. The meetings contain approximately 18% overlapped speech. Of the 38 meetings, 12 were used for test as in [5], 22 used for training and 4 were used as a development set for tuning parameters.

Near-field mixed-headset results
As a first step, experiments were performed using high-quality single-channel signals obtained by mixing the audio signals from the four individual headset channels in each meeting. This was done to facilitate the development of the system and the feature selection process, as well as to determine an upper limit to the performance of any far-field system. This is analogous to the meeting speech recognition task, where the near-field recognition condition provides a performance bound for the far-field one.

Table 1. *Performance comparisons for systems using AMI development data and near-field audio data.*

System	Prec.	Rec.	F-score	DER
Baseline Diarization	-	-	-	32.28
MFCC+Δ	0.7	0.27	0.39	30.84
MFCC+Eg+LPC+DPEΔ	0.72	0.25	0.37	30.46
MFCC+DPE+Δ	0.73	0.32	0.45	30.13
MFCC+Eg+DPE+Δ	0.76	0.34	0.47	29.90

The results for various systems are given in Table 1, with the best overall system appearing in the last row. The precision, recall, and F-score values are for the overlap detector in isolation while the DER refers to the diarization system performance after post-processing using the identified overlap segments. The first row, "Baseline Diarization", gives the baseline performance of the diarization system without the use of overlap information. It should be noted that reference speech/nonspeech information was used in the diarization system so as not to confound the false alarm error contributions of the speech/nonspeech detector, which is presently an independent system, with those of the overlap detector (see below).

From the results we see that, for this simplified case, the overlap detector indeed improves the diarization system. In addition, a strong correlation between improved overlap detection precision and reduced DER for the diarization system exists, as we anticipated. Lastly, the best feature combination—MFCCs, RMS energy, diarization posterior entropy, and first differences—yields a DER reduction of 2.38%, a relative improvement of about 7.4%.

Far-field results
Having demonstrated system functionality using the mixed-headset signals, we subsequently conducted experiments for the more realistic scenario of single-channel far-field microphone signals. In this case the overlap detector must contend

with a poorer signal-to-noise (SNR) ratio as well as convolutive effects from room responses. The results for various systems are given in Table 2 and are presented in the same fashion as Table 1. Observe that the performance of the system degrades significantly owing to far-field conditions. Nevertheless, reductions in DER are made by the overlap detector in this case as well. The best feature combination—the same one as in the near-field case—yields a DER reduction of 1.36%, a relative improvement of about 3.6% over the baseline. Lastly, here, too, we see the correlation between precision and DER reductions.

Table 2. *Performance comparisons for systems using AMI development data and far-field audio data.*

System	Prec.	Rec.	F-score	DER
Baseline Diarization	-	-	-	38.11
MFCC+Δ	0.54	0.15	0.24	38.09
MFCC+Eg+LPC+DPE+Δ	0.61	0.33	0.42	37.26
MFCC+DPE+Δ	0.64	0.31	0.42	36.83
MFCC+Eg+DPE+Δ	0.66	0.26	0.37	36.75

Error analysis

As mentioned in Section 2, the DER is composed of false alarm (FA), missed speech (MS), and speaker errors (SE). By decomposing the DER into its constituent errors we can better analyze the effect of the overlap system. Table 3 gives a breakdown of the DER for the baseline and best performing systems for both the near-field and far-field conditions. Observe the baseline false alarm rate of zero in both cases; this is due to the use of reference speech activity regions. The overlap detection system introduces false alarm errors in both cases, though the number is small due to the relatively high precision. The effectiveness of the overlap segmenter is shown clearly in the reduction of the missed speech error. In addition, the small increase in speaker error indicates the post-processing speaker assignment algorithm is largely effective as well.

Table 3. *Breakdown of diarization error rate for baseline and best overlap-detecting diarization systems. Error measures consist of false alarm (FA), missed speech (MS), and speaker error (SE).*

System	Near-field			Far-field		
	FA	MS	SE	FA	MS	SE
Diarization	0.0	18.3	14.0	0.0	18.3	19.8
+ Overlap detection	1.4	13.5	14.9	1.8	14.6	20.3

6. CONCLUSIONS AND FUTURE WORK

In this paper we have motivated the need for overlapped speech detection in the speaker diarization task and have described our first efforts toward developing a system to perform this detection. In the case of near-field mixed-headset audio we obtained a 7.4% relative improvement in DER for the diarization system, while for the more challenging far-field case a 3.6% relative improvement was obtained. In both cases we observed the importance of a high-precision detector in achieving improvements.

As this work represents the beginning of the system development—and of the overall effort in automatic speaker overlap detection—the potential amount of future work is considerable. There are, however, a few key directions we intend to pursue. One is to continue to identify features useful to overlap detection, with a particular emphasis on robustness to environmental variations—a major concern in the meetings domain. Another is using speech and nonspeech detection information from the segmenter to more fully integrate the overlap detection and diarization systems; at present, the diarization system uses a separate speech activity detector. Lastly, we intend to investigate the possibility of using the overlap detector in a diarization pre-processing step (added to the post-processing step described above) to exclude overlapped speech from the training data to achieve purer speaker models. The overlap detector may then be able to help reduce not only missed speech errors, but speaker errors as well.

7. REFERENCES

[1] C. Wooters and M. Huijberts, "The ICSI RT07s speaker diarization system," in *Proc. of the Rich Transcription 2007 Meeting Recognition Evaluation Workshop*, 2007, Baltimore, MD.

[2] N. Morgan *et al.*, "Meetings about meetings: Research at ICSI on speech in multiparty conversations," in *Proc. ICASSP 2003*, 2003, pp. 740–743.

[3] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversations," in *Proc. Eurospeech 2001*, 2001, pp. 1359–1362, Aalborg, Denmark.

[4] O. Çetin and E. Shriberg, "Speaker overlaps and ASR errors in meetings: Effects before, during, and after the overlap," in *Proc. ICASSP 2006*, 2006, pp. 357–360, Toulouse, France.

[5] H. Hung *et al.*, "Using audio and video features to classify the most dominant person in meetings," in *Proceedings of the 15th International ACM Conference on Multimedia*, New York, NY, USA, 2007, pp. 835–838, ACM Press.