interspeech 2008 incorporating SST 08

Two's a Crowd: Improving Speaker Diarization by Automatically Identifying and Excluding Overlapped Speech

Kofi Boakye[1,2], *Oriol Vinyals*[1], *Gerald Friedland*[1]

[1] International Computer Science Institute, Berkeley, CA, U.S.A.
[2] University of California, Berkeley, U.S.A.
{kaboakye,vinyals,fractor}@icsi.berkeley.edu

Abstract

We present an update to our initial work [1] on overlapped speech detection for improving speaker diarization. Specifically, we describe the addition of new features and feature warping techniques that improve segmenter and, consequently, diarization performance. We also demonstrate improved diarization performance by additionally using overlap segment information in a new diarization pre-processing step which excludes overlap segments from speaker clustering. On a subset of the AMI Meeting Corpus we show that this overlap exclusion step nearly triples the relative improvement of diarization error rate as compared to overlap segment post-processing alone.
Index Terms: speaker diarization, overlap detection

10.21437/Interspeech.2008-6

1. Introduction

Speaker diarization—the task of determining "Who spoke when?"—has in recent years received considerable attention from the speech community, particularly as a result of the NIST Rich Transcription (RT) meeting recognition evaluation. A fundamental limitation of current systems which try to perform this task, however, is that, for any given time, the "who" in question can be only a single speaker. The presence of overlapped speech, though, is common in multiparty meetings and, consequently, presents a significant challenge to these automatic systems. Specifically, in regions where more than one speaker is active, missed speech errors will be incurred and, given the high performance of some state-of-the-art systems, this can be a substantial portion of the overall diarization error.

In previous work [1], we described our initial efforts towards addressing this issue by using an HMM-based overlap detector along with a segment-based post-processing scheme for the segmentation output of the speaker diarization system. The detector utilized features found to work well for the task—namely, MFCCs, RMS energy, and diarization posterior entropies—to identify regions of overlapped speech, while the post-processing procedure performed speaker assignment based on speaker posterior probabilities output by the diarization system. In addition, we motivated the focus on a high precision operating point for the overlap detector, since false alarms from the detector increase the baseline diarization error rate, whereas misses have zero effect on this error rate.

A less direct, but also significant, effect of overlapped speech in diarization pertains to speaker clustering and modeling. Because overlap segments contain speech from multiple speakers, they should not be assigned to any individual speaker cluster nor included in any individual speaker model. Doing so adversely affects the quality of the speaker models, which potentially reduces diarization performance. In [2], for example, the authors, using an oracle system, demonstrated an improvement in diarization performance by excluding overlap regions from the input to the diarization system.

In this paper, we both add upon our previous work in developing an overlapped speech detector for diarization and address this issue of overlapped speech in clustering and modeling. With the addition of new features—namely, spectral flatness, harmonic energy ratio, and modulation spectrogram features—as well as better modeling and feature warping techniques, we are able to make significant gains in performance over the previous segmenter. By using overlap segment information to exclude overlapped speech segments in the speaker diarization step, too, significant additional gains can be made, even in the automatic case.

The paper is organized as follows. We briefly describe the baseline diarization system in Section 2 and detail updates to the HMM-based overlap segmenter in Section 3. The diarization segment post-processing procedure is described in Section 4 and we present results on a number of experiments using data from the AMI Meeting Corpus in Section 5. Conclusions are given in Section 6 and we identify future work in Section 7.

2. Baseline diarization system

As in our previous work, we evaluated speaker diarization improvement using the system fielded by ICSI in the NIST 2007 Rich Transcription meeting recognition evaluation (RT07s). This state-of-the-art system performs diarization by agglomerative clustering of segments with merging 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 process of merging, model re-training and re-alignment. In the merge step, BIC-based merge scores are used to determine which two clusters should be merged or whether merging 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 [3].

The standard metric for system performance is the diarization error rate (DER), defined as the sum of the false alarm (falsely identifying speech), missed speech (failing to identify speech), and speaker error (incorrectly identifying the speaker)

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 along with DAAD.

times, divided by the total amount of speech time in a test audio file. As with every other state-of-the-art diarization system, the baseline system assigns only a single speaker label to a segment. Consequently, missed speech errors from speaker overlap persist and cannot be reduced. Presently, these errors make up a large portion of the diarization error. For example, in previous RT diarization evaluations, up to 22% of the ICSI system diarization error consisted of missed speech errors due to overlap [3]. Overlap detection, therefore, is clearly crucial to improving the performance of the system.

3. Overlap detector

3.1. System overview

The overlap detector consists of a three-class (nonspeech, speech, and overlapped speech) HMM-based segmenter in which each class is represented with a three-state model and emission probabilities are modeled using a multivariate Gaussian Mixture Model (GMM) with diagonal covariances. In the original system, 32 mixture components were utilized, but we have found that 256 components gives improved segmentation performance, most likely due to better modeling. For each class HMM, mixtures are shared between the three states, with separate mixture weights for each state. The models are trained using an iterative Gaussian splitting technique with successive re-estimation. Training data is obtained by using ASR forced-alignment times generated from ground-truth transcriptions of the audio to identify speech, nonspeech, and overlap regions. For testing, a single Viterbi decoding pass of the full channel waveform is performed. The overlap regions obtained are scored against reference overlap regions (again identified using forced-alignment) and segmentation performance is measured using precision, recall, and F-score values computed based on false alarm, missed detection, and total overlapped speech times. To tune for low false alarms, we adjust the transition penalty from the speech to the overlap class in the Viterbi decoding and set a value based on held-out data.

3.2. Features

Our previous best-performing system utilized a feature vector representation consisting of 12-th order MFCCs, RMS energy, and diarization posterior entropy (DPE) along with their first differences. In our continued search for features useful for overlap detection, we have identified the following and incorporated them into the system. The feature values are computed every 10 ms with window sizes given below.

Spectral flatness
The spectral flatness measure SFM_{dB} in dB is given as:

$$SFM_{dB} = 10 \log_{10} \frac{Gm}{Am} \tag{1}$$

$$Gm = \sqrt[\frac{1}{N}]{\prod_{i=0}^{N-1} \text{mag}(i)} \text{ and } Am = \sum_{i=0}^{N-1} \text{mag}(i) \tag{2}$$

Where Gm is the geometric mean, Am is the arithmetic mean, $\text{mag}(i)$ is the magnitude of the spectral line i and N is the number of FFT points or spectral lines.

Spectral flatness is often used as a voicing measure (e.g., [4]), but since it is related to the shape of the spectrum (which differs for single-speaker and overlapped speech), it may also be of use in detecting overlap. The feature is computed over a window of 60 ms.

Harmonic energy ratio (HER)
In the case of voiced overlapped speech, the energy distribution in the spectrum will not be as concentrated in the bands associated with the detected pitch—typically the pitch of the dominant (i.e., more audible) speaker—as in the single-speaker scenario. We have, therefore, included the harmonic energy ratio as a feature for detection. The harmonic energy ratio represents the ratio of harmonic to non-harmonic energy for a frame as determined by a pitch detector. The energy is computed by selecting each harmonic FFT band plus two adjacent bands, computing the energy and dividing by the energy of the remaining bands. In the event no pitch is detected, the average pitch value is used and the ratio is computed accordingly. HER is computed over a window of 50 ms.

Modulation Spectrogram (MSG) features
The modulation spectrogram provides an alternative and complementary representation of the speech signal with a focus on temporal structure. Developed by Kingsbury et al. and detailed in [5], it represents a filtered version of the spectrogram of a speech signal. The spectrogram of the signal is computed using an FFT with step size of 10 ms and an analysis window of 25 ms. In contrast to MFCC features, where for each frame the DCT coefficients of the Mel log-FFT amplitudes are computed, the MSG analyzes the spectrogram using 18 bands from 0 to 8 kHz, filtering the resulting 18 temporal signals with two different filters: a 0-8 Hz filter and an 8-16 Hz filter. For each frame, the MSG features capture the low-pass and band-pass behavior of the spectrogram of the signal within each of the 18 sub-bands, resulting in a total of 36 features per frame.

As we stated, the modulation spectrogram provides information about longer temporal phenomena in contrast with MFCCs, as it uses 0.21 seconds of analysis to extract the features. Thus, we expect that, jointly with MFCCs, this representation of the spectrum of the signal will be richer and perform better in overlap detection.

3.3. Feature warping

A common issue for classification systems is robustness to data mismatches. For speech and audio data, the mismatch typically relates to the channel (e.g., headset vs. distant microphone) or the recording environment (e.g., different rooms or microphone locations). The most severe form of this mismatch occurs between training and test sets, but another form can arise within either (or both) of these sets and also results in reduced performance. In our case, for example, having training or test data from a variety of sites—and, hence, different recording environments—creates an internal mismatch that can affect the performance of the trained models.

To address this, we employ a component-level Gaussianization procedure to the feature vectors on a per-meeting basis. Each feature component is normalized to have a zero-mean, unity-variance Gaussian distribution. This is achieved using a non-linear warping constructed from histograms and an inverse Gaussian function. Saon et al. in [6] demonstrate improved performance in speech recognition tasks with significant channel and environment variability using Gaussianization. In addition, they motivate the choice of normal target distribution by suggesting it facilitates the use of diagonal covariance Gaussians in the acoustic models. To further facilitate this, we also apply a decorrelation procedure via the KLT.

4. Diarization segment post-processing

To apply the segment information obtained from the overlap detector to the diarization system, we use a procedure as follows. For each frame, the diarization system produces speaker likelihoods based on each speaker model. Using these likelihoods, frame-level speaker posteriors are calculated and summed over the frames of the identified overlap segment and a single "score" for each speaker is obtained. 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, for the corpora of interest, two-speaker overlap typically comprises 80% or more of the instances of overlapped speech [7]. A diagram of the overall system is shown in Figure 1.

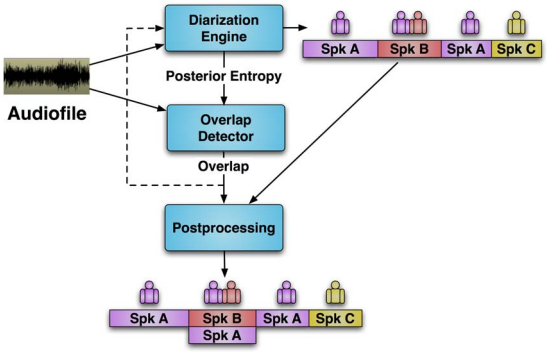

Figure 1: *Diagram of integrated overlap detector and diarization system.*

5. Experimental results

A number of experiments were carried out to evaluate the performance of the overlap detection system. The audio data used consisted of meetings from the AMI Meeting Corpus, a collection of 100 hours of scenario-based meetings each involving four participants. The data is sampled at 16 kHz and single-channel far-field microphone signals are used, one per meeting. We chose this single-channel scenario rather than the multi-channel one because our system operates on a single feature stream and the features, when obtained from delay-and-sum beamformed audio data, performed much more poorly. This is a consequence of the delay-and-sum procedure, which can potentially suppress overlapped speakers.

5.1. AMI data: single-site

For this first set-up, we used data obtained at a single site, specifically, the IDIAP subset of the corpus. This subset consists of 38 meetings containing approximately 18% overlapped speech. Just as in [1], the 38 meetings were divided into 12 for testing, 22 for training, and 4 for parameter tuning.

Overlap detection system comparison

In this experiment, we compare the performance of the previous and new overlap detection systems on this acoustically uniform data set. The results are given in Table 1, with precision, recall, and F-score values referring to the detector in isolation and the DER to the diarization system performance with segment post-processing. The first row, "Baseline Diarization" gives the baseline performance of the diarization system with no use of overlap information. For this and subsequent experiments, unless otherwise stated, we used reference speech/nonspeech information to separate the overlap detection performance from that of the speech/nonspeech detector.

From the results we see that the new system significantly outperforms the previous one in terms of precision, recall and F-score. Consequently, the relative DER improvement increases as well, from 3% to 5%.

Table 1: *Performance comparisons for overlap segmenter systems using single-site AMI data.*

System	Prec.	Rec.	F-score	DER
Baseline Diarization	-	-	-	30.74
Previous	0.68	0.19	0.30	29.82
New	0.76	0.25	0.38	29.17

5.2. AMI data: multi-site

For this set-up, we used data obtained from multiple sites to provide greater acoustic variability within the training, testing, and tuning sets and observe system performance under these conditions. This multi-site data set consists of 60 meetings, divided as follows: 10 for testing, 40 for training and 10 for tuning. The composition of the sets were obtained by randomly selecting (without replacement) meetings from the corpus, which contains 170 meetings in total. This subset of 60 meetings contains approximately 15% overlapped speech.

Overlap detection system comparison

We again compare the performance of the previous and new overlap detection systems, but now for this more acoustically variable—and, hence, more challenging—data set. The results are given in Table 2 and are presented in the same fashion as Table 1. In terms of precision, the two systems produce similar performance, but the new system significantly outperforms the previous in recall and, consequently, F-score. The relative DER improvement for the previous system is under 1% while for the new is 2.3%.

Table 2: *Performance comparisons for overlap segmenter systems using multi-site AMI data.*

System	Prec.	Rec.	F-score	DER
Baseline Diarization	-	-	-	32.77
Previous	0.68	0.11	0.18	32.48
Current	0.67	0.26	0.38	32.02

Overlap exclusion

In addition to using overlap segment information in a post-processing procedure as described in Section 4, we may also use this information in a pre-processing step in which these segments are excluded from the speaker clustering process of the diarization system. The expectation is that this will result in purer speaker clusters which may improve the diarization system performance by reducing speaker error. Such an improvement, for example, was observed in [2] when using the perfect overlap detection of an oracle system. Because the speaker label assignment in the post-processing step utilizes speaker posteriors—which may improve as a result of the purer clusters—this may benefit from the pre-processing as well.

In this experiment, we use the output of the new overlap detector on the multi-site data to identify segments for exclusion from speaker clustering (this corresponds to the dashed line of Figure 1). We then apply the post-processing to this new speaker diarization output. The results are presented in Table 3.

Using the overlap exclusion pre-processing alone, we obtain a 4.8% relative improvement in diarization error rate—more than twice the improvement of the post-processing alone. Including the post-processing yields even more gains, giving the final system a relative improvement of 6.8%—nearly three times the improvement of the post-processing alone.

Table 3: *Performance improvement for new system with overlap exclusion.*

System	DER	Rel. Imp.(%)
Baseline Diarization	32.77	-
Overlap exclusion	31.20	4.8
+Overlap detection	30.54	6.8

Automatic speech/nonspeech segmenter

In all previous experiments, reference speech activity regions were used in the diarization system so as not to confound the false alarm error contributions of the speech/nonspeech detector with those of the overlap detector. In an operational system, however, speech/nonspeech detection must be performed automatically, and here we analyze the performance of the overlap detector in this context. Speech activity regions were determined using the standard speech/nonspeech detector for the ICSI diarization system. The detector performs iterative training and re-segmentation of the audio into three classes: speech, silence, and audible nonspeech. To bootstrap the process, an initial segmentation is created with an HMM trained on broadcast news data. A detailed description can be found in [3].

Results for the new overlap detector are presented in Table 4. Missed speech (MS), false alarm (FA), and speaker error (SE) rates are presented for reference and automatic speech/nonspeech detection. From the baseline results, we see that the automatic detection introduces both misses and false alarms, though very small (and similar) amounts. Curiously, the automatic system has a lower speaker error, suggesting the reference segmentation is not optimal for speaker modeling. Excluding overlaps yields a slightly greater improvement for the automatic case, with a reduction of 2% absolute, as compared to 1.5% absolute for the reference speech regions. The speaker error increases less, too, for the automatic system when post-processing is applied, possibly due to the improved modeling. Because of the increased false alarms and misses, however, the automatic system benefits slightly less from post-processing, with a larger increase in false alarms and smaller reduction in misses as compared to using reference segmentation.

Table 4: *Breakdown of diarization error rate using reference and automatic speech/nonspeech regions. Errors consist of missed speech (MS), false alarm (FA), and speaker error (SE).*

System	Reference			Automatic		
	MS	FA	SE	MS	FA	SE
Diarization	9.9	0.0	22.8	12.1	1.9	21.0
Overlap exclusion	9.9	0.0	21.3	12.1	1.9	19.1
+ Overlap detection	7.0	1.7	21.8	10.6	4.0	19.3

6. Conclusions

In this paper we have described our continued progress towards developing an overlapped speech detection system for speaker diarization. By including additional features, performing feature warping, and enhancing modeling with more Gaussians, we have produced a system which significantly outperforms the previous version under single-site and the more challenging multi-site evaluation conditions. In addition, we have demonstrated that additional gains in performance can be made using overlap information by excluding overlap segments from the diarization clustering process. Lastly, we have observed that automatic speech/nonspeech detection has only a minor negative effect on overall performance.

7. Future Work

As we continue to make improvements to the overlap detector, it is important to have an idea of the extent to which such a system can help improve diarization. This assists in providing motivation for such efforts as well as a yardstick of progress. To demonstrate, we incorporate the overlap exclusion and post-processing procedures into the overall diarization process as in Section 5, but using the reference overlap segments. This is similar to [2], but with the use of our diarization posterior-based speaker assignment procedure rather than heuristic or perfect assignment. The results are given in Table 5. From these results we see that the use of overlap segment information can yield very large gains in speaker diarization performance. Excluding overlap segments alone can give up to a 12.5% relative improvement in DER, and this increases to 27.6% when segment post-processing is performed as well. Clearly, there is much motivation to continue, as the automatic segmenter lags greatly behind the oracle system in terms of DER improvement.

Table 5: *Performance improvement for oracle system with overlap exclusion.*

System	DER	Rel. Imp.(%)
Baseline Diarization	32.77	-
Overlap exclusion	28.68	12.5
+Overlap detection	23.71	27.6

With regard to future work, there are a few primary directions we intend to pursue. Firstly, identifying additional features for overlap detection should continue to improve the system. Because the detector segments speech and nonspeech as well as overlap, we are also interested in using it as a replacement for the current speech/nonspeech segmenter, and so better integrating the overlap detection and speaker diarization systems.

8. References

[1] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved diarization in multiparty meetings," in *Proc. ICASSP 2008*, 2008, pp. 4353–4356, Las Vegas, Nevada.

[2] S.Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *Proc. ASRU 2007*, 2007, pp. 683–686, Kyoto, Japan.

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

[4] R. Yantorno, K. Krishnamachari, and J. Lovekin, "The spectral autocorrelation peak valley ratio (SAPVR) - a usable speech measure employed as a co-channel detection system," in *Proc. of IEEE Workshop on Intelligent Signal Processing*, 2001.

[5] B.E.D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, pp. 117–132, August 1998.

[6] G. Saon, S. Dharanipragada, and D. Povey, "Feature space gaussianization," in *Proc. ICASSP 2004*, 2004, vol. 1, pp. 329–332, Montreal, Canada.

[7] B. Trueba-Hornero, "Handling overlapped speech in speaker diarization," M.S. thesis, Universitat Politecnica de Catalunya, May 2008.