Improved Overlapped Speech Handling for Speaker Diarization

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

[1]Lawrence Livermore National Laboratory, Livermore, CA, U.S.A.
[2]International Computer Science Institute, Berkeley, CA, U.S.A.
kofi@llnl.gov, vinyals@icsi.berkeley.edu, fractor@icsi.berkeley.edu

Abstract

We present our ongoing work in addressing the issue of overlapped speech in speaker diarization through the use of overlap segmentation, overlapped speech exclusion, and overlap segment labeling. Using feature analysis, we identify the most salient features from a candidate list including those from our previous system and a set of newly proposed features. In addition, through independent optimization of overlap exclusion and labeling, we obtain a relative diarization error rate improvement of 15.1% on a sampled subset of the AMI Meeting Corpus, more than double our previous result. When analyzed independently, we show that the performance improvement due to overlapped speech exclusion now rivals that of an oracle system using reference overlap segments.

Index Terms: speaker diarization, overlap detection

10.21437/Interspeech.2011-382

1. Introduction

The presence of overlapped speech i.e., speech containing acoustic content from more than one speaker, poses a challenge to a variety of automatic human language technologies, particularly in the context of multiparty meeting recordings. Speaker diarization—which aims to determine "Who spoke when?"—is one such technology in which, increasingly, overlapped speech represents a significant portion of system errors. While performance of state-of-the-art diarization systems has improved greatly within recent years, these systems are largely incapable of handling overlapped speech. Indeed, the standard approach for diarization has been developed with the underlying assumption that segments are speaker-homogeneous.

In previous work [1, 2], we have demonstrated that handling these non-homogeneous, overlapped speech regions results in lower diarization error rates: By applying a system to identify speaker overlap and a segment post-processing procedure to determine the speaker labels in these regions, we obtained a reduction in missed speech error compared to the single speaker assignment. Furthermore, excluding these overlapped speech segments from the speaker clustering process in diarization yielded a reduction of the speaker error rate.

Central to our approach is the identification of features suited to the task of overlap segmentation. In this work we, add to the list of candidate features and analyze them according to their discriminant capability in order to select those most salient for incorporation into the segmentation system. In addition, as overlapped segment exclusion and labeling have been observed by us and others [3] to differ in sensitivity to overlap precision and recall, we optimize for the two procedures independently. The results are significant gains in performance—in particular, via overlap exclusion—with a reduction in number of features utilized.

2. Baseline ICSI speaker diarization engine

To facilitate comparison with previous work, speaker diarization improvement was evaluated using the 2007 ICSI speaker diarization engine as fielded in the NIST 2007 Rich Transcription evaluation (RT07s). It performs diarization by agglomerative hierarchical clustering of segments with merging based on Bayesian Information Criterion (BIC) scores. These BIC scores are computed using GMMs of 19th-order MFCC features. The system is described fully in [4].

The standard metric for system performance is the diarization error rate (DER), defined as the sum of the false alarm, missed speech, and speaker error times, divided by the total amount of speech time in a test audio file. As previously noted, a limitation of this and other state-of-the-art systems is that they assign only a single speaker label to a segment. Consequently, missed speech errors from speaker overlap cannot be reduced. In addition, including overlapped speech in the clustering process can adversely affect the quality of the speaker models. Dealing with overlapped speech, then, is a major step in improving the performance of the system.

3. Overlapped speech handling

Our approach to handling overlapped speech in diarization consists of three main parts: overlap detection, overlapped speech exclusion, and overlap segment labeling, all described below.

3.1. Overlap detection

The overlap detection system 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 256-component multivariate Gaussian Mixture Model (GMM) with diagonal covariances. For each class HMM, mixtures are shared between the three states, while mixture weights remain separate. 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. A precision parameter—the transition penalty from the speech to the overlap class in the Viterbi decoding—can be adjusted for precision and recall, and was determined using held-out tuning data.

3.2. Overlapped speech exclusion

As a pre-processing step, the identified overlap segments are excluded from the speaker clustering process of the diarization

Copyright © 2011 ISCA

system. This approach has previously been shown to yield improved performance by reducing speaker error, likely due to the increased purity of the speaker clusters. Because the speaker label assignment in the post-processing step utilizes speaker posteriors—which may also improve as a result of the purer clusters—it was hypothesized that this procedure would benefit from the exclusion pre-processing as well.

3.3. Overlap segment labeling

To apply the segment information obtained from the overlap detector to the diarization system, the following procedure is employed. Frame-level speaker posteriors derived from the diarization system are summed over the frames of the identified overlap segment to produce a single "score" for each speaker. The two highest-scoring speakers are then selected for labeling. One exception is if the original segment label was associated with a third speaker, in which case this speaker and the highest-scoring speaker are chosen. Though this procedure limits the number of possible overlapping speakers to two, we have observed that two-speaker overlap typically comprises some 80% of the instances of overlapped speech for the corpora of interest [2]. A diagram of the overall system is shown in Figure 1.

4. Features

The choice of features is of critical importance to the performance of a segmentation system such as that described in Section 3.1. While traditional acoustic features such as MFCCs may be sufficient for speech/nonspeech segmentation, the addition of overlapped speech—which in many ways acoustically resembles single-speaker speech—requires the use and exploration of additional features.

4.1. Candidate features

The set of features considered for the overlapped speech segmenter was largely based on our previous work in [1, 2]. This includes 12th-order MFCCs, RMS energy, LPC residual energy, diarization posterior entropy, spectral flatness, harmonic energy ratio, and modulation spectrogram features. In addition to these features, we explored three new features: kurtosis, zero-crossing rate, and harmonicity, described below. For these features, a window of 50 ms was applied prior to computation. For all features, values are computed every 10 ms and first differences are incorporated into the final feature vector as well.

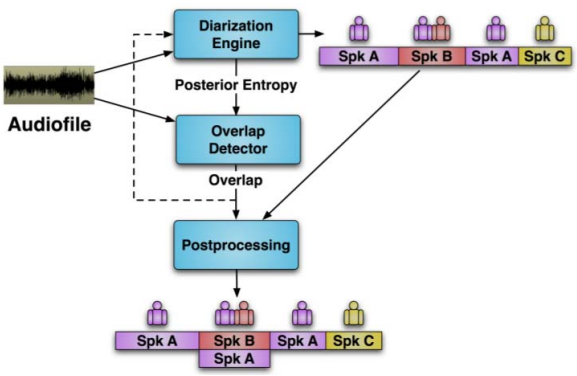


Figure 1: *Diagram of integrated overlapped speech handling system for diarization.*

Kurtosis
The kurtosis of a zero-mean random variable x is defined as

$$\kappa_x = \frac{E[x^4]}{[E[x^2]]^2} - 3 \tag{1}$$

where $E[\cdot]$ is the expectation operator. This serves as a measure of the "Gaussianity" of a random variable. Speech signals, which are typically modeled as having a Laplacian or Gamma distribution, tend to be super-Gaussian—i.e., have a kurtosis above zero. Furthermore, the sum of such distributions—in line with the central limit theorem—has lower kurtosis (i.e., is more Gaussian) than individual distributions. This was observed, for example, by Krishnamachari et al. in [5]. As such, signal kurtosis could serve as an effective feature for detecting overlapped speech.

Zero-crossing rate
The zero-crossing rate (ZCR)—i.e., the rate of sign changes along a signal—is commonly used in speech processing for audio classification, including tasks such as voiced/unvoiced classification [6]. The periodicity of voiced speech tends to produce a lower zero-crossing rate than unvoiced speech (e.g., fricatives) and background noise, making the value useful for such a task. In the case of simultaneously voiced speech from overlapping speakers, it is plausible that the superposition of the speech signals, in addition to increasing energy, will increase the zero-crossing rate. This feature was computed according to

$$\text{ZCR} = \frac{1}{2N} \sum_{i=1}^{N-1} |\operatorname{sgn}(x[i]) - \operatorname{sgn}(x[i-1])| \tag{2}$$

where N is the length of the applied Hamming window.

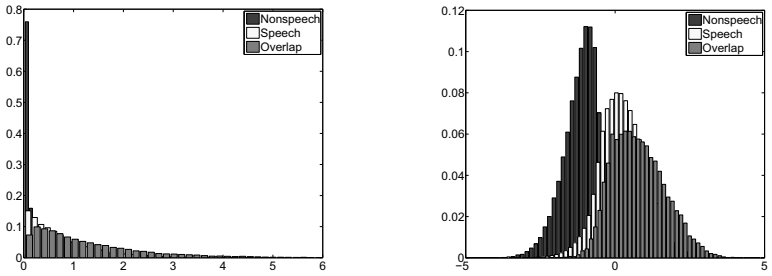
Harmonicity
The harmonicity feature measures the degree of acoustic periodicity within a window of the speech signal. Specifically, we compute the standard deviation of the pitch peaks obtained via the maximum likelihood pitch estimation procedure proposed by Wise et al. in [7]. In the case of voiced overlapped speech, multiple sets of peaks should be present, with one set associated with each overlapping speaker. These peak values will also differ based on the harmonic energy of each speaker, and, consequently, the pitch peak standard deviation will be higher than in the case of single-speaker voiced speech.

4.2. Feature warping

To address the potential variability induced by having training and test audio data from different sites, a non-linear feature normalization was performed. Specifically, we employed a component-level Gaussianization procedure to the acoustic feature vectors on a per-meeting basis. Gaussianization has yielded improved performance in cases of significant channel and environment variability in speech recognition tasks, as shown, for example, in [8]. In addition to Gaussianization, we apply a decorrelation procedure via the KLT to facilitate the use of diagonal covariance Gaussians in the acoustic models. The effect of the Gaussianization procedure is illustrated in Figure 2 for the RMS energy feature on a sample meeting.

4.3. Feature analysis and selection

Given the diversity of features we have proposed and utilized for segmenting overlapped speech, it is likely that some aid more



(a) Raw (b) Gaussianized

Figure 2: *Normalized histograms of the (a) raw and (b) Gaussianized RMS energy for a meeting in the AMI corpus.*

than others in segmentation performance. Poorly discriminating features, then, could be eliminated from the feature set, with little or no negative impact—and possibly positive impact—on the overall result. Though several forms of such feature analysis exist, we employed a KL divergence based feature discriminant capability analysis as proposed by Zhou et al. in [9]. The discriminant capability of each feature component for each acoustic event is given as

$$d_{ij} = D(p_{ij}||q_i) \tag{3}$$

where p_{ij} is the distribution of the ith feature component given the jth acoustic event and q_i is the distribution of the ith feature component given all audio parts. In our case, the sole acoustic event is overlapped speech. In addition, as the overlap segmenter features have been warped via Gaussianization, we compute the KL divergence using the Gaussian parametric model.

Figure 3 presents results of the analysis using the labeled training data from the experiments in Section 5 below. The results suggest that the most salient features are LPC residual energy, spectral flatness, RMS energy, and harmonicity. The 36 modulation spectrogram feature components have a discriminative capability slightly higher than MFCCs, but markedly lower than the aforementioned most salient features. Harmonic energy ratio and diarization posterior entropy are comparable to MFCCs, while kurtosis and zero-crossing rate, being the least salient, have divergences lower than even the cepstral features. Based on these results, the features used in the overlap segmenter were limited to the 12 MFCCs and the four high-saliency overlap features.

5. Experimental results

To evaluate the performance of our overlapped speech handling approach, experiments were conducted using data from the AMI Meeting Corpus, which consists of 100 hours of meeting recordings. As our overlap detection system operates on a single feature stream, single-channel far-field microphone signals are used, one per meeting. In addition, while previous efforts examined both single-site and multi-site data for evaluation, here we focus only on the latter, more challenging, scenario. The data set consists of 40 training, 10 tuning, and 10 test meetings, obtained by randomly selecting (without replacement) meetings from the corpus, which contains 170 meetings in total. This subset of 60 meetings contains roughly 15% overlapped speech. Table 1 lists the meetings used for tuning and testing.

	Meetings				
Tune	EN2004a	ES2013c	IS1001c	IS1001d	IS1005a
	IS1007b	IS1007c	TS3006a	TS3007c	TS3012b
Test	EN2003a	EN2009b	ES2008a	ES2015d	IN1008
	IN1012	IS1002c	IS1003b	IS1008b	TS3009c

Table 1: *Tuning and test meetings for the AMI multi-site data.*

5.1. System performance comparison

Reference segmentation

In this first experiment, we compare the performance of our previous and new overlapped speech handling systems using reference speech activity regions for the diarization system. This is done to maintain separate the false alarm error contributions of the speech activity and overlap detectors. In addition, we present two versions of the new system: one in which a single tuning operating point—optimized for overlap segment labeling—is used for both overlap detection and overlapped speech exclusion (i.e., "old tuning") and a second in which the two are tuned independently (i.e., "new tuning"). This latter approach was motivated by observations of the differing sensitivities of the two procedures to precision and recall and by the work in [3]. Finally, oracle results using reference overlap segments are presented as an additional point of comparison. These results are given in Table 2 and are separated so as to show the impact of overlap exclusion, segment labeling, and the two techniques in combination.

Baseline	32.77		
Overlap det.:	+Exclusion	+Labeling	+Both
Previous	31.20/+4.8	32.02/+2.3	**30.54/+6.8**
New (old tuning)	30.72/+6.3	32.49/+0.9	**30.31/+7.5**
New (new tuning)	28.15/+14.1	32.49/+0.9	**27.82/+15.1**
Oracle	28.68/+12.5	27.61/15.8	**23.71/+27.6**

Table 2: *Diarization error rate and relative improvement over baseline (in %) using reference speech/nonspeech regions*

The first thing to note is that, for the new system with the new tuning strategy, the diarization performance has improved substantially. Overlapped speech exclusion improvement has increased nearly three-fold to 14.1% and overall improvement has more than doubled to 15.1%. This difference highlights another notable trend: the contribution of exclusion to diarization improvement far exceeds that of segment labeling. This contrasts the oracle scenario, in which labeling slightly exceeds exclusion. For the new system with the old tuning strategy, the difference in overall performance is negligible. This, too, is notable, though, given that only five of the ten candidate features (and 32 of the 110 candidate feature components) were

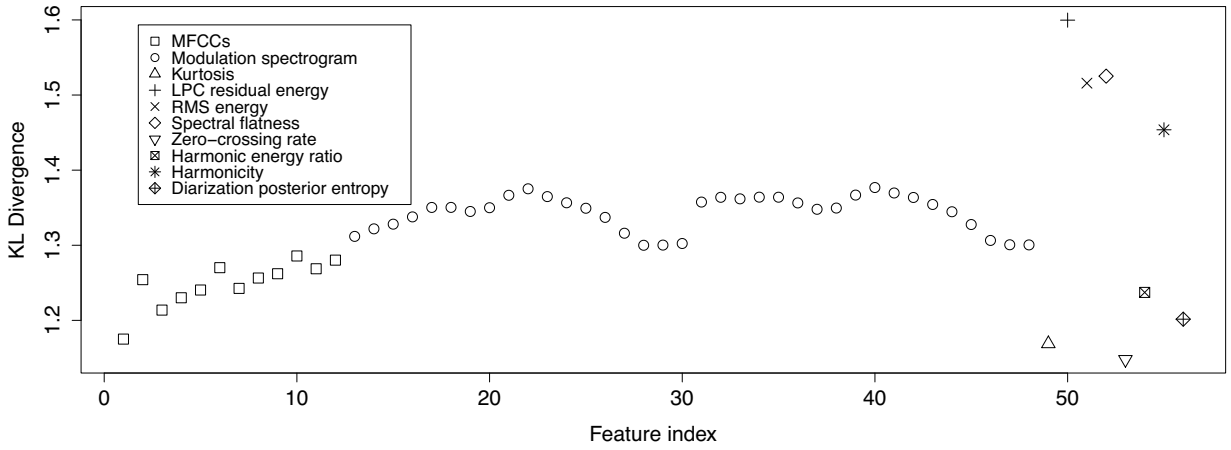


Figure 3: *KL divergence measure for candidate features computed over training data.*

used. Also, we observe that, for exclusion with new tuning, the improvement slightly surpasses that of the reference overlap segments. As exclusion improvement is not bounded in the same way as segment labeling, this suggests—though very weakly—that additional gains could be made. Lastly, to illustrate the differing operating points of exclusion and labeling, we note that the precision/recall performance of exclusion is 0.55/0.40 and labeling is 0.64/0.24 for the new tuning system.

Automatic segmentation

Though the use of reference segmentation is important for the analysis of the impact of overlap handling, automatic segmentation represents a more realistic operational scenario and is examined here as well. For this experiment, speech activity regions were determined using the standard speech/nonspeech detector for the ICSI diarization system. Results are presented in Table 3 in the same format as the previous experiment.

Baseline	34.99		
Overlap det.:	+Exclusion	+Labeling	+Both
Previous	33.10/+5.4	34.21/+2.2	**33.07/+5.5**
New (old tuning)	33.71/+3.7	34.41/+1.7	**33.45/+4.4**
New (new tuning)	30.91/+11.7	34.41/+1.7	**30.66/+12.4**
Oracle	31.77/+9.2	29.73/+15.0	**26.91/+15.3**

Table 3: *Diarization error rate and relative improvement over baseline (in %) using automatic speech/nonspeech regions*

The trends observed in these results are similar to those above. The automatic segmentation produces a small degradation in performance at all levels, in part due to the introduction of false alarm errors from the speech activity detection. Again, we observe that the new system with new tuning substantially outperforms the previous system, and that overlapped speech exclusion is the main source of improvement.

6. Conclusion

We have presented our ongoing work in addressing the issue of overlapped speech in speaker diarization through the use of overlap segmentation, overlapped speech exclusion, and overlap segment labeling. We have demonstrated the utility of a feature analysis technique—discriminant capability analysis—by applying it to a set of candidate features for overlap segmentation and identifying and selecting those most salient. In addition, through independent optimization of overlap exclusion and labeling, we have obtained significant increases in relative DER improvement, in the case of exclusion rivaling the performance of segmentation using reference overlap segments. Finally, we have shown that these gains persist in the case of automatic rather than reference speech/nonspeech segmentation.

7. Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. This research was also supported by a CISCO research grant.

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] K. Boakye, O. Vinyals, and G. Friedland, "Two's a crowd: Improving speaker diarization by automatically identfying and excluding overlapped speech," in *Proc. Interspeech 2008*, 2008, pp. 32–35, Brisbane, Australia.

[3] M. Zelenák, C. Segura, and J. Hernando, "Overlap detection for speaker diarization by fusing spectral and spatial features," in *Proc. Interspeech 2010*, 2010, pp. 2302–2305, Makuhari,Japan.

[4] 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.

[5] Kasturi R. Krishnamachari, Robert E. Yantorno, and Jereme M. Lovekin, "Use of local kurtosis measure for spotting usable speech segments in co-channel speech," in *Proc. of ICASSP 2001*, 2001, pp. 649–652, Salt Lake City, UT.

[6] Sassan Ahmadi and Andreas S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 333–338, May 1999.

[7] J. Wise, J. Caprio, and T. Parks, "Maximum likelihood pitch estimation," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 5, pp. 418–423, oct 1976.

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

[9] X. Zhou, X. Zhuang, M. Liu, H. Tang, M. Hasegawa-Johnson, and T. Huang, "HMM-based acoustic event detection with AdaBoost feature selection," *Multimodal Technologies for Perception of Humans*, pp. 345–353, 2009.