EFFICIENT USE OF OVERLAP INFORMATION IN SPEAKER DIARIZATION

Scott Otterson and Mari Ostendorf

University of Washington, Dept. EE. Seattle, Washington, USA

ABSTRACT

Speaker overlap in meetings is thought to be a significant contributor to error in speaker diarization, but it is not clear if overlaps are problematic for speaker clustering and/or if errors could be addressed by assigning multiple labels in overlap regions. In this paper, we look at these issues experimentally, assuming perfect detection of overlaps, to assess the relative importance of these problems and the potential impact of overlap detection. With our best features, we find that detecting overlaps could potentially improve diarization accuracy by 15% relative, using a simple strategy of assigning speaker labels in overlap regions according to the labels of the neighboring segments. In addition, the use of cross-correlation features with MFCC's reduces the performance gap due to overlaps, so that there is little gain from removing overlapped regions before clustering.

Index Terms— speaker identification, diarization, localization, overlap

1. INTRODUCTION

The goal of speaker diarization is to extract segments of speech and to associate them with the correct speaker. This task is particularly challenging for meeting recordings, in which a variable (meeting-dependent) number of microphones are placed at unknown distances from the speakers in an unknown (ad hoc) configuration. In the NIST meeting recognition evaluation [1], this is known as the multiple-distant-microphone (MDM) test condition.

A difficulty in meeting diarization, particularly for the MDM condition, is that speakers often talk at the same time. In typical NIST meetings, 78% of word tokens occur within silence bounded regions containing speaker overlap [1]. Using a different definition of overlap over similar data, the authors of [2] found that between 8-17% of words, and 31-54% of sentence-like "spurts" contained overlap. Typically, systems ignore the problem of overlapping speech, which poses a challenge to any type of speech processing. In early work on meeting speech, error analyses suggested that overlaps caused an 11% increase in word error rate (WER) and a 17% increase in diarization error [1]. On recent diarization systems, and with more precise word times, overlaps have been estimated to cause a 3.5% diarization error increase [3].

In this paper, we conduct another analysis of the impact of speaker overlaps on diarization error, looking at the potential gain in performance due to perfect overlap detection. In addition, we distinguish between potential causes of errors at different stages of the diarization process and look at the role of location features such as microphone pair correlation lags. We find that detecting overlaps can indeed improve performance substantially, but is most important in the labeling (vs. clustering) stage of processing. In the sections to follow, we briefly review related work on overlaps, raise two main experimental questions, outline the experimental paradigm, and present the results.

2. OVERLAP DETECTION AND PROCESSING

One way of dealing with the problem of overlaps is to preprocess the overlapped speech signal with a source separation algorithm [4, 5, 6, 7], the potential advantage being that speaker-specific characteristics could be isolated, and thus contribute to overall diarization accuracy during the speaker clustering process. However, these methods have various limitations. For example, the currently popular independent component analysis (ICA) methods have difficulties with one or more of the conditions present in conversational speech. Many have problems in the presence of reverberation [8]. Others require time windows on the order of 5s for convergence [9], much longer than the typical overlap present in conversational speech (we have calculated that the median overlap length, averaged over 54 meetings in [10], is 250ms). Nearly all source separation algorithms assume that the number of speakers is known [11].

Another approach to overlaps is to detect them and then exclude them from subsequent processing. In the monaural channel vocoder work of [12], spikes in speech amplitude kurtosis were found to bracket 83-92% of single talker speech, where single talker speech was defined as speech with greater than 10dB talker to interferer ratio (TIR). Closely related overlap detection algorithms in [13, 14, 15] were applied to speaker identification. On synthetically overlapped TIMIT data, up to 75% of "usable" speech was detected, where "usable" segments were defined as those in which the level of one talker was 20 dB or more greater than that of the other. The authors of [15] used the spectral autocorrelation peak valley ratio (SAPVR) as a proxy to TIR. After removing speaker overlap with a SAPVR threshold, they found that a speaker

identification system got results equivalent to removing speech with a TIR less than 20dB. However, speaker identifications (ID) were assigned based on a known TIR, a quantity unavailable in practice. SAPVR features yielded poor overlap detection performance for meeting speech personal microphone voice activity detection, but good performance was obtained using kurtosis and cross-correlations [16].

Other monaural work includes [17], which used pitch prediction and cepstral features to detect overlapped speech, and [18], which used robust Hough transform pitch detection, LPC model fit and entropy measures. On a subset of the TIMIT corpus [19], the combined feature classifier caught 74.7% of overlapped frames with a false detect rate of 12.3% (F=0.81).

In [20], clustered "Eigen locations" derived from ROOT-MUSIC beamforming outputs were able to find location peaks in projected space due to simultaneous talkers. Unfortunately, the ROOT-MUSIC is inappropriate for meetings due to its high sensitivity to microphone position errors.

Finally, in [21], speaker segments obtained from a single-speaker diarization system were used to train hidden Markov model (HMM) speaker states corresponding to every possible overlap pair between detected speakers. Meeting data was then re-segmented using the combined single-speaker and overlap HMM. While the authors do not cite accuracy numbers, they state that, while the system was capable of detecting overlap, it correctly identified the overlapped speakers only a third of the time, and that the approach did not reduce overall diarization error.

3. QUESTIONS IN OVERLAP HANDLING

These different results leave open several questions about the impact of overlap regions on diarization performance and processing strategies. While analyses have shown that overlaps contribute to diarization error, the results are mixed as to the relative importance. Other work suggests that, while detecting meeting speech overlaps may be within reach, it may still be difficult to design an algorithm that can tell who is talking. Most current systems (including our baseline) assign only one speaker label to any region of speech. Is this alone what leads to errors, or is the overlapped speech corrupting the single-speaker models learned in the process of diarization? It may be that very high accuracy overlap detection is needed to successfully leverage overlap information, or it may be that there is little gain from such detection.

In order to better understand the potential impact of overlap detection, we factor out the effect of overlap detection errors by using an "oracle" overlap detector and evaluating alternatives for using this information in the diarization process.

In a typical diarization system, a stream of speech is initially broken into short time segments, either at detected speaker change points or at uniform intervals. Then these segments are grouped and assigned to a speaker by agglomerative clustering. Overlaps cause errors in at least two ways. First, clustering assigns segments to only one speaker during an overlap; other speakers during the overlap will be missed. Second, clusterer speaker models can be corrupted when overlapped speech is included in their training data.

In diarization experiments, we ask the following questions:

1. Would diarization be improved if overlaps were detected and removed before speaker clustering?
2. After the initial single-speaker diarization has been completed, does the assignment of two speaker labels given knowledge of location of overlap regions lead to significant improvements in diarization scores?

4. EXPERIMENTAL PARADIGM

Experiments were conducted on data from the NIST Rich Meeting Transcription Project. In these meetings, held at five locations in conventional conference rooms, 3 to 18 participants were recorded with 1 to 16 distant, omni-directional microphones. Sound was acquired at 16bits and 16KHz on multi-track digital recording systems. No information about microphone or speaker location is available. The data used included the NIST 2004 development (dev) test and evaluation (eval) sets, and the eval sets for 2005 and 2006. Meetings with fewer than two microphones were omitted, since location features cannot be computed in the single-microphone case, and their use was of interest in this study. The final data set contained 31 meetings.

The diarization system used for evaluating our overlap handling algorithm is based on the system described in [22], which yielded the best results in the NIST 2006 competition (the authors made this feature extraction and clustering software available to us). As in most meeting diarization (and speaker recognition) systems designed for this task, the system uses mel-warped cepstral coefficients (MFCC's), but in addition it includes "location features," the set of correlation lags computed between microphones placed at unknown locations. The system also does a delay-sum beamforming of the distant microphone channels; the MFCC speaker ID features were generated from this signal. The diarization system uses a standard agglomerative clustering scheme with a Bayesian information criterion (BIC) stopping threshold and a hidden Markov model (HMM) to enforce minimum length constraints. Tuning parameters (number of mixtures, HMM states, etc.) were fixed to those found to be optimal for the NIST 2007 evaluation.

For this work, an improved version of the correlation features is used, based on a speech-specific Hilbert envelope for computing correlations together with a low-dimensional vector of features based on a principal components analysis transform of a vector of microphone pair correlations, as described in [23].

Table 1. **Effect of input overlap processing on single-talker diarization accuracy**

Features	Overlaps in Input?	Diarization Error (%)
MFCC	Y	18.5
MFCC	N	17.9
MFCC+XC	Y	11.9
MFCC+XC	N	11.6

Speaker diarization results were measured against references with word times determined by forced alignments, using the NIST diarization scoring software[1]. For this tool, the diarization error is the sum of time over all reference speakers for which speech is either missed or falsely detected – including during overlaps and silences – divided by the total speech time of the scored region, counting overlap times for each speaker.

5. RESULTS

Table 1 shows the effect of overlaps on single talker speech accuracy, that is, the subset of speech where there are no overlaps. For this series of experiments, references derived from forced alignments were used to exclude overlaps from the final diarization outputs, so that the diarization error was calculated only over single talker speech. We then compared the cases where the input to the diarization clustering includes vs. excludes the overlap regions. To explore the effect of overlaps on different features, clustering experiments were conducted using: i) MFCC's alone, and ii) MFCC's in one observation stream of the HMM and cross correlation features (XC) in a second stream, as in [23].

In the first two rows of Table 1, we see that an ideal overlap detector, which excludes overlaps from MFCC-only clustering input data, would improve single-speaker-only diarization performance by 0.6%. Adding XC features reduces the error by about 35% relative in both cases, shrinking the gap in performance.

In Table 2, we show the effect of different overlap output processing approaches on the full diarization score, i.e. including overlap regions in the diarization score. As in the previous experiments, we use oracle overlap regions, i.e., assuming perfect detection, in this case to obtain different input and output processing conditions. The first row gives the diarization baseline performance with MFCC features, with no special overlap handling. In the next two rows, we show that replacing the speaker label associated with an overlap with the labels associated with the two speakers detected closest to the overlap (referred to here as "nearest-2") yields a 2% absolute improvement (10% relative) if overlaps are not excluded at the clusterer input. If overlaps are excluded at the clustering input, a 2.5% absolute improvement results.

[1] www.nist.gov/speech/tests/rt/rt2006/spring/code/md-eval-v21.pl

Table 2. **Effect of input/output overlap processing on full diarization accuracy**

Features	Overlaps in Input?	Output Post-process	Diarization Error (%)
MFCC	Y	none	21.6
MFCC	Y	nearest-2	19.6
MFCC	N	nearest-2	19.1
MFCC	N	perfect	18.0
MFCC+XC	Y	none	15.1
MFCC+XC	Y	nearest-2	12.9
MFCC+XC	N	nearest-2	12.9
MFCC+XC	N	perfect	12.2
MFCC+XC	Y	perfect	12.5

In the fourth row, we show the result of "perfect" overlap post processing, where oracle overlap regions are assigned to the true speaker labels. Here, "perfect" means "the best you can do." If the clusterer had underestimated the number of speakers so that some overlap segments contained speakers which were not detected anywhere in the meeting, then no segment for that speaker was inserted into the overlap region. This causes a diarization error, but it matches the condition of the nearest-2 strategy, which can only select from speakers detected by the clusterer. For MFCC features, this "perfect" overlap scheme yields a 1.1% absolute improvement over the nearest-2 approach; the nearest-2 approach obtained 70% of the improvement ideally possible with input and post-processing overlap detection.

The remaining rows of Table 2 show that, with MFCC+XC features, the nearest-2 approach comes much closer to the ideal. Comparing the baseline performance in the fifth row with the next two, we see that the simple nearest-2 post processing step improves performance by 2.2% absolute (15% relative), regardless of whether or not overlaps were excluded at the clusterer input. From the last two rows, we see that nearest-2 approach obtains 75% of the ideal improvement.

6. CONCLUSIONS

In this paper, we have demonstrated that a simple nearest-2 post-processing step will yield most of diarization performance improvement possible from overlap detection, given the speaker clusters detected by the diarization system under study. We show that, for MFCC+XC features, removing overlap segments from the input of the diarization clusterer yields no further improvement when cross-correlation features are included in clustering. The results suggest that one way in which cross correlation features help diarization is to improve the overlap-robustness of single-speaker model building.

Future work would, of course, include the development of an accurate automatic overlap detector, and further explo-

ration of overlap post processing methods in this context.

7. REFERENCES

[1] J. S. Garofolo, C. D. Laprun, and J. G. Fiscus, "The rich transcription 2004 spring meeting recognition evaluation," in *NIST Meeting Recognition Workshop*, 2004.

[2] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Eurospeech*, 2001.

[3] D. A. van Leeuwen and M. Konecny, "Progress in the AMIDA speaker diarization system for meeting data," in *NIST RT07 Workshop*, 2007.

[4] F. Abrard, Y. Deville, and P. White, "From blind source separation to blind source cancellation in the underdetermined case: A new approach based on time-frequency analysis," in *Independent Component Analysis and Blind Signal Separation, Intl. Conf. on*, 2001.

[5] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Independent component analysis and blind signal separation, Intl. Conf. on*, pp. 215–220, 2000.

[6] H. Attias and C. E. Schreiner, "Blind source separation and deconvolution by dynamic component analysis," *Neural Networks for Signal Processing*, vol. VII, pp. 456–465, 1997.

[7] Y. Cao, S. Sridharan, and M. Moody, "Multichannel speech separation by eigendecomposition and its application to co-talker interference removal," *IEEE Trans. on Speech and Audio Proc.*, vol. 5, no. 3, pp. 209–219, 1997.

[8] S. Araki, S. Makino, R. Mukai, T. Nishikawa, and H. Saruwatari, "Fundamental limitation of frequency domain blind source separation for convolved mixture of speech," in *Independent Components Analysis, Intl. Conf. on*, 2001.

[9] T.-W. Lee, A. Ziehe, R. Orglmeister, and T. Sejnowski, "Combining time-delayed decorrelation and ICA: Towards solving the cocktail party problem," in *Proc. ICASSP*, vol. 2, pp. 1249–1253, 1998.

[10] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," in *Proc., Human Language Technology Conf.*, 2001.

[11] A. W. van der Kouwe, D. Wang, and G. J. Brown, "A comparison of auditory and blind separation techniques for speech segregaton," *IEEE Trans. on Speech and Audio Proc.*, vol. 9, pp. 189–195, 2000.

[12] K. R. Krishnamachari, R. E. Yantorno, J. M. Lovekin, D. S. Benincasa, and S. J. Wendt, "Use of local kurtosis measure for spotting usable speech segments in co-channel speech," in *Proc. ICASSP*, vol. 1, pp. 649 –652, 2001.

[13] J. Lovekin, K. R. Krishnanmachari, and R. E. Yantorno, "Adjacent pitch period comparison (APPC) as a usability measure of speech segments under co-channel conditions," in *ISPACS*, 2001.

[14] N. Chandra and R. E. Yantorno, "Usable speech detection using the modified spectral autocorrelation peak to valley ratio using the LPC residual," in *Intl. Conf., Signal and Image Proc.*, 2002.

[15] R. E. Yantorno, K. R. Krishnamachari, D. S. Benincasa, J. M. Lovekin, and S. J. Wenndt, "The spectral autocorrelation peak valley ratio (SAPVR) - a usable speech measure employed as a co-channel detection system," in *IEEE Intl. Workshop on Intelligent Signal Processing*, 2001.

[16] S. N. Wrigley, G. J. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multi-channel audio," *IEEE Trans., Speech and Audio Proc.*, vol. 13, no. 1, pp. 84–91, 2005.

[17] M. A. Lewis and R. P. Ramachandran, "Cochannel speaker count labelling based on the use of cepstral and pitch predicton derived features," *Pattern Recongnition*, vol. 34, pp. 449–507, 2001.

[18] S. Otterson, S. Furui, and M. Ostendorf, "Speaker overlap detection with Hough transform pitch features," Tech. Rep. 2004-0012, Univesity of Washington, 2004.

[19] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM.," Tech. Rep. NISTIR 4930, National Institute of Standards and Technology, 1993.

[20] E. D. D. Claudio, R. Parisi, and G. Orlandi, "Multi-source localization in reverberant environments by ROOT-MUSIC and clustering," in *Proc. ICASSP*, vol. 2, pp. 921–924, 2000.

[21] D. A. van Leeuwen and M. Huijbregts, "The AMI speaker diarization system for NIST RT06s meeting data," in *NIST RT06 workshop*, 2006.

[22] X. Anguera, C. Wooters, and J. M. Pardo, "Robust speaker diarization for meetings: ICSI RT06s evaluation system," in *Proc. ICSLP*, 2006.

[23] S. Otterson, "Improved location features for meeting speaker diarization," in *Interspeech*, 2007.