FALA 2010
VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop

On the Improvement of Speaker Diarization by Detecting Overlapped Speech

Martin Zelenák, Javier Hernando

Universitat Politècnica de Catalunya, Barcelona, Spain
{martin.zelenak,javier.hernando}@upc.edu

Abstract

Simultaneous speech in meeting environment is responsible for a certain amount of errors caused by standard speaker diarization systems. We are presenting an overlap detection system for far-field data based on spectral and spatial features, where the spatial features obtained on different microphone pairs are fused by means of principal component analysis. Detected overlap segments are applied for speaker diarization in order to increase the purity of speaker clusters and to recover missed speech by assigning multiple speaker labels. Investigation on the relationship between overlap detection properties and diarization improvement revealed very distinct behaviour of overlap exclusion and overlap labeling.
Index Terms: speaker overlap detection, speaker diarization

1. Introduction

Spontaneous human conversation very often includes certain amount of simultaneous speech. This naturally occurring phenomenon is typical for meeting environment, where listening people for example interrupt the leading speaker in order to grab floor or give backchannel to encourage his talk. Some speaker overlaps can also be the result of elevated emotions (laughing, arguing). Shribergh [1] observed that the amount of overlapped speech is not necessarily dependant on the amount of people involved in the conversation, ergo, few people can produce significant overlap too. Overlapped speech poses a problem for many automatic human language technologies, including speaker diarization, which, given a recording, strives to answer the question *"Who spoke when?"*. In general, no prior knowledge about the speakers is provided. Conventional diarization systems are able to assign only one speaker label per segment, which, obviously, leads to missed speech for overlapped speakers. Furthermore, including overlapped speech into the model building can be a potential source of speaker error, since the models could be corrupt.

In previous works, the common way to detect speaker overlap in meetings was to segment each of the individual speaker channels with an ergodic hidden Markov model (HMM). Overlapped speech was either one of the decoding classes [2] or was marked in a post-processing algorithm [3]. The solution suggested in [4] eliminated the necessity to train any model.

Some published algorithms focused on distant microphone channels exclusively. Knowing the number of speakers beforehand and assuming their location will not change during a multiparty conversation, the authors in [5] proposed to use microphone pair time delays (TDE) to segment audio according to speakers. They showed the possibility to detect two simultaneous speakers by modeling short-term speaker turns for every pair of the assumed speakers with an HMM.

Explicit modeling of all pairs of speakers after an initial diarization was also explored in [6]. Even though the authors claim to be able to detect overlap, it did not lead to a reduction of the diarization error. Improvement of speaker diarization by handling overlaps detected with an HMM-based detection system was firstly presented in [7] on a subset of the AMI corpus.

In this paper we are presenting an overlapped speech detection system for distant channel microphones, which successfully combines spectral and spatial features. To deal with the high and variable dimensionality of spatial feature space we are suggesting the application of principal component analysis (PCA), which fuses feature vectors across different microphone pairs into one spatial feature set. A similar approach was also chosen for diarization purposes in [8].

Our motivation for detecting overlapped speech is to improve a baseline diarization system with two techniques. In the first, also referred to as overlap exclusion, overlaps shall be discarded from training cluster models, hoping to achieve a more precise segmentation. The second technique makes it possible to assign two speaker labels in segments with simultaneous speech. We examined the behaviour of these two techniques in the context of changing overlap detection properties more in detail and it results to be substantially different for labeling as for exclusion. Experiments were conducted on single- and multi-site recordings from the AMI Meeting corpus.

The remainder of this paper is organized as follows. Speaker overlap detection and speaker diarization system with overlap handling improvements are outlined in Sections 2 and 3, respectively. Experimental results are discussed in Section 4 and conclusions are given in Section 5.

2. Speaker overlap detection

2.1. Spectral features

The overlap detection system uses several spectral-based features which were identified as to be conveying some overlap information. Cepstrum is successfully applied for a handful of speech related tasks and constitutes a good basis of a feature set, for that reason 12 MFCCs were extracted every 10 ms over a window of 30 ms.

Linear predictive coding (LPC) analyzes the speech signal by estimating the formants of a speaker. It is assumed that LPC of a reasonably chosen order can model the spectrum of a single speaker quite well, but will fail for a region with multiple speakers [9]. In the latter case, more energy will be left in the residual signal, which represents prediction error. In this system, residual energy of a 12th-order LPC (LPCRE) was computed over a 25 ms window. The feature set was furthermore extended with first order delta coefficients and all features were mean-variance normalized according to statistics obtained from training data.

This work has been funded by the Spanish project SAPIRE (TEC2007-65470). The first author is supported by a grant from the Catalan autonomous government.

Another spectral-based feature is the spectral flatness (SF) extracted over a window of 30 ms. This feature was applied for discrimination between speech and non-speech [10], but can eventually convey information about the number of speakers speaking. It is defined as the ratio between geometric and arithmetic mean of a certain number (100 in our case) of spectral magnitudes

$$SFM_{dB} = 10\log_{10}\frac{\sqrt[N]{\prod_{i=0}^{N-1} mag(i)}}{\sum_{i=0}^{N-1} mag(i)}. \quad (1)$$

2.2. Spatial features

Several spatial features based on cross-correlation were introduced to improve spectral overlap detection on distant channel data. The first spatial feature is the value of the principal cross-correlation peak, which is a measure of *coherence* between signals. For a pair of microphones *i* and *j* it is defined as

$$C_{ij} = \max(R_{ij}(\tau)), \quad (2)$$

where $R_{ij}(\tau)$ is the Generalized Cross Correlation with Phase Transform weighting (GCC-PHAT) [11] that is often used in order to improve robustness in reverberant environments. Ideally, the coherence value should be high for single-source situations and low if noises, reverberation or concurrent acoustic sources are present. In the general case, the main peak is attenuated when multiple sources introduce random peaks.

In situations dealing with multiple, possibly moving, concurrent speakers it was observed that time delay estimates (TDEs) produced by the GCC-PHAT jump from one speaker to another at a very high rate as one source dominates due to the non-stationarity of the voice. Thus, the first order derivative of the time delay estimate *delta TDE* is expected to carry certain degree of information on overlaps. TDE is defined as

$$\hat{\tau}_{ij} = \underset{\tau}{\operatorname{argmax}} R_{ij}(\tau). \quad (3)$$

Derived from the coherence value, we are also proposing to extract the coherence *dispersion ratio*, as shown in eq. 4. This value is computed as the relation of the square of main peak value and the sum of secondary peaks square values corresponding to other acoustic sources that may be present in the scenario,

$$D_{ij} = \frac{C_{ij}^2}{\sum_{t=-w_{ij}}^{w_{ij}} R_{ij}^2(\hat{\tau}_{ij}+t)}, \quad (4)$$

where the size of the window w_{ij} is adjusted in accordance with the TDE standard deviation of microphone pair *(i,j)*.

The dimensionality of a spatial-feature vector can be very high since we extract three features for every microphone pair. Furthermore, the number of microphones differs from site to site, making it difficult to train a general model. In order to deal with these problems, we are proposing to unify and reduce the number of microphone pairs with a PCA transformation. For each discussed spatial feature and for every site we estimated a transformation matrix and then applied just the first component. Consequently, we obtained three transformed features (coherence, dispersion, delta TDE) for each frame.

2.3. System architecture

A schematic block diagram of the overlap detection system with link to speaker diarization is given in Fig. 1. The system considers three acoustic classes representing non-speech, single-speaker speech and overlapped speech. For each class a three-state HMM is defined where every state is modeled with a GMM with diagonal covariance. Since the amount of training data is not balanced for all classes, we are using 256 Gaussian components for single-speaker speech and 32 or 64 components for overlapped speech and non-speech. GMMs are created by iterative Gaussian splitting and subsequent re-estimation. Spectral and spatial features are modeled with separate GMMs, with output probability calculated with feature stream weights of 0.75 and 0.25, respectively. In the case of spatial GMMs, the means and variances are shared across the three states.

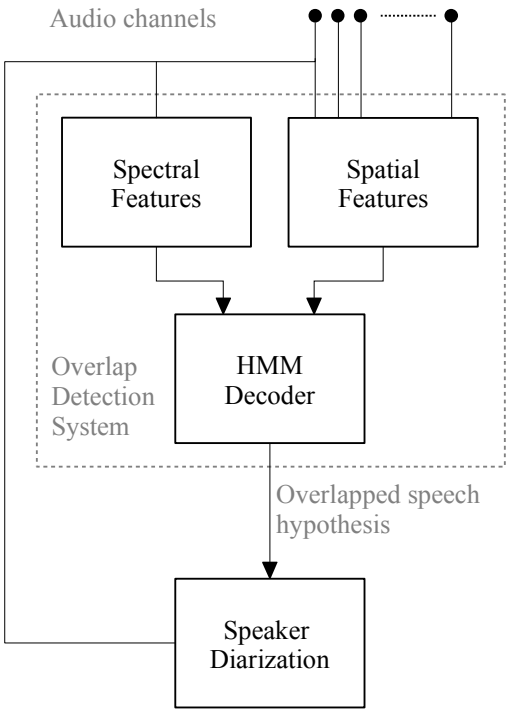


Figure 1: *Overlap detection system block diagram*

The detection hypothesis is obtained by Viterbi decoding and applying a word network. For precision purposes the transition from single-speaker speech to overlapped speech can be penalized with an overlap insertion penalty (OIP) and direct transitions between non-speech and overlapped speech are completely forbidden.

3. Speaker diarization system

Our speaker diarization system, detailed in [12], follows the commonly used agglomerative clustering approach. In the beginning, speech is broken into rather short uniform segments and the successive clustering stage groups acoustically similar segments and assigns them to speaker clusters. The number of initial clusters is determined automatically from audio length with minimal and maximal value constraints. Clusters are modeled with GMMs and cluster pair merging in each iteration is driven by Bayesian information criterion (BIC). The system operates with 20 MFCCs extracted from 30 ms frames. The performance of diarization is evaluated by means of diarization error rate (DER), which is the sum of missed seech rate, false alarm rate and speaker error rate.

Overlap handling in diarization comprises the exclusion and/or labeling of simultaneous speech. The first technique

blocks overlap frames from being included into cluster initialization and GMM training, but does not prevent decoding them. The aim of this technique is to get lower speaker detection error rates with more precise clusters. Overlap labeling technique seeks to select the two most likely clusters in Viterbi decoding instead of only one. In this way the missed speaker time should be decreased.

In order to evaluate just the impact of overlapped speech on speaker segmentation, detected overlaps are masked with reference speech/non-speech segments before given to diarization system. The diarization system is using reference speech segments as well.

4. Experiments

4.1. Database and experimental setup

The database used for our experiments was the AMI Meeting corpus, which comprises 100 hours of meeting recordings. We were working with far-field microphone array channels sampled at 16 kHz. We defined single- and multi-site scenarios. The first included recordings only from Idiap site and the latter also from Edinburgh and TNO sites. The recordings were then divided into training set (22 for both single- and multi-site scenario), development set (3 and 9) and evaluation set (11 and 10). The average amount of overlapped speech in these scenarios was 14.40% and 15.10%, respectively. Training and evaluation of the overlap detection system are performed with forced-alignment annotations obtained by the SRI's DECIPHER recognizer.

In the presented experiments, we are comparing the results obtained with two feature setups for the detection of speaker overlap. The first one is a baseline spectral system (*Spct*) and the second is a system based on the combination of spectral and PCA-transformed spatial features (*Spct+Spat*). Overlap detection performance is measured with Recall—ratio between true detected and reference overlap time, Precision—ratio between true and all detected overlaps, and with Error—the sum of missed and false overlaps divided by reference overlap time. Results depend very much on the value of the overlap insertion penalty, which controls the amount of overlaps the system will posit. It can be perceived as a compensation for an undertrained model. Initially, four values of OIP were defined based on different detection characteristics on development data, accounting for hypotheses with the highest recall, the highest F-ratio, the lowest error rate and an acceptably high precision.

It is assumed that hypotheses exhibiting high recall will be suitable for overlap exclusion, because as many overlaps as possible will be discarded from model building. On the contrary, high precision and low error will be needed for successful overlap labeling, since all of the false overlaps will be propagated to DER, but only a perfect labeling would transform all true overlaps into reduction of missed speaker time.

Obviously, the nature of the two techniques is very different. Therefore, it is not useful to use necessarily the same overlap hypohesis for both, but rather two independent overlap hypotheses, one for each technique.

4.2. Single-site experimental results

The DER relative improvements of handling overlapped speech over the diarization baseline for single-site recordings are given in the right column of Table 1. We can see that the overall better *Spct+Spat* overlap detection hypothesis used for exclusion resulted in higher DER improvement than the *Spct* hypothesis. In the case of labeling, the hypothesis with lower error (*Spct+Spat*), though lower precision as well, gained more improvement as the other (*Spct*). The highest improvement of 11.6% was achieved by applying overlaps detected by the combined spectral and spatial system.

Table 1: *Speaker diarization with excluding and/or labeling overlapped segments on single-site evaluation data. Overlap detection recall, precision, error and corresponding DER rel. improvement over the baseline (all values in %)*

Baseline DER					38.3
Overlap det.:	Rcl.	Prc.	Err.	DER rel. imp. [%]	
Spct	45.7	52.2	96.1	+Excl.	+3.9
	27.0	83.9	78.2	+Labl.	+4.9
	”	”	”	+Both	**+6.9**
Spct+Spat	49.2	59.0	85.0	+Excl.	+5.2
	35.4	80.5	73.2	+Labl.	+5.5
	”	”	”	+Both	**+11.6**

Table 2: *Speaker diarization with excluding and/or labeling overlapped segments on multi-site evaluation data. Overlap detection recall, precision, error and corresponding DER rel. improvement over the baseline (all values in %)*

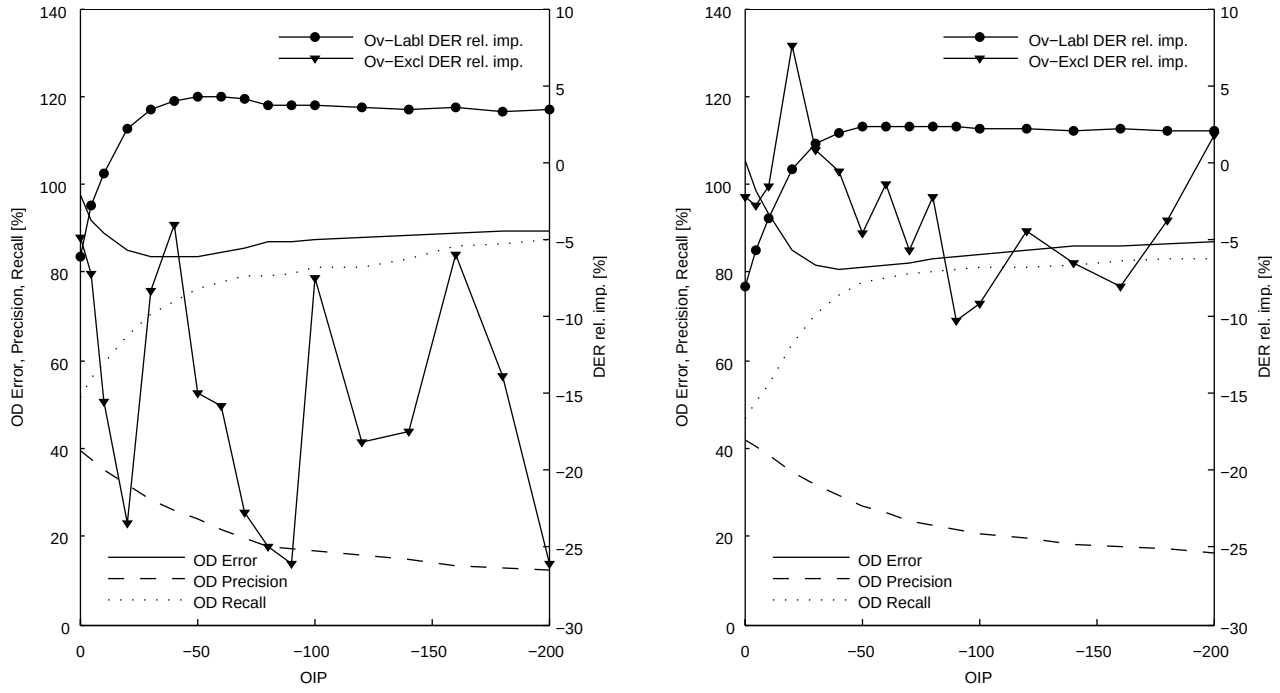
Baseline DER					37.3
Overlap det.:	Rcl.	Prc.	Err.	DER rel. imp. [%]	
Spct	49.5	41.8	119.5	+Excl.	+7.5
	25.4	74.9	83.1	+Labl.	+2.1
	”	”	”	+Both	**+10.2**
Spct+Spat	59.3	42.3	121.5	+Excl.	+6.9
	30.4	70.5	82.3	+Labl.	+1.7
	”	”	”	+Both	**+9.5**

4.3. Multi-site experimental results

The overlap detection results in multi-site scenario, given in Table 2, are considerably worse than in the case of single-site data. Despite this worse overlap detection performance, excluding overlapped segments led to surprisingly higher relative DER improvements. More expected are the modest improvements by labeling. The best improvement of up to 10.2% is obtained with (*Spct*) hypotheses. The lower performance of spatial setup could be eventually explained by the fact that the spatial features are, in general, not commensurable accross different microphone pairs, since they are tied to physical characteristics of a particular pair. The PCA-based transformation of features from multiple sites probably lacked some robustness in this case.

4.4. Overlap detection and diarization improvement

In order to investigate more on the relationship between overlap detection performance and the obtained improvements in diarization, we performed a further set of experiments on single- and multi-site development data. A large number of overlap hypotheses produced with several overlap insertion penalties was employed for exclusion and labeling.



(a) Single-Site

(b) Multi-Site

Figure 2: *Overlap detection performance of spectral system and corresponding speaker diarization relative DER improvement over the baseline by excluding and labeling detected overlap segments for (a) single- and (b) multi-site development data.*

The detection metrics, i.e., recall, precision and error, and the corresponding DER improvements are given in Fig. 2 (a) and (b). Note that the peak of labeling performance lies within the region of lowest detection error and is also somewhat shifted to the right towards higher precisions. This observation is in compliance with our former assumption and the fact that the complement of the error tells us how much we can theoretically gain by assigning second labels. Overlap exclusion exhibits a less predictable behaviour in terms of DER improvements, making it difficult to derive any kind of conclusion at this point. Still, the results from Tables 1 and 2 show that the improvements can be significant.

5. Conclusions

We have presented an overlap detection system based on spectral and spatial features. Detected overlaps were used in speaker diarization for increasing the purity of speaker models and to recover missed speech by assigning multiple speaker labels. Experiments on evaluation single- and multi-site data showed improvements over baseline diarization system. Further investigation using held-out data revealed the changeable nature of improvements by overlap exclusion, but also confirmed the labeling performance's dependence on overlap precision and error.

6. References

[1] Shriberg, E., "Spontaneous Speech: How People Really Talk and Why Engineers Should Care," in *Proc. Interspeech '05*, Lisbon, Portugal, 2005, pp. 1781–1784.

[2] Wrigley, S.N. et al., "Speech and Crosstalk Detection in Multichannel Audio," *IEEE Transactions on Speech and Audio Processing*, vol 13. pp. 84–91, 2005.

[3] Pfau, E., Ellis, D.P.W and Stolcke, A., "Multispeaker Speech Activity Detector for the ICSI Meeting Recorder," in *Proc. ASRU '01*, Madonna di Campiglio, Italy, 2001, pp. 107–110.

[4] Laskowski, K., Jin, Q. and Schultz, T. "Crosscorrelation-based Multispeaker Speech Activity Detection," in *Interspeech '04*, Jeju Island, Korea, 2004, pp. 973–976.

[5] Lathoud, G. and McCowan, L., "Location Based Speaker Segmentation," in *Proc. ICME '03*, Baltimore, USA, 2003, pp. III-621–4 vol.3.

[6] van Leeuwen, D.A. and Huijbregts, M., "The AMI Speaker Diarization System for NIST RT06s Meeting Data," in *Machine Learning for Multimodal Interaction*, LNCS, vol. 4299/2006, Springer Berlin/Heidelberg, 2006, pp. 371–384.

[7] Boakye, K., Trueba-Hornero, B., Vinyals, O., and Friedland, G., "Overlapped Speech Detection for Improved Speaker Diarization in Multiparty Meetings," in *Proc. ICASSP '08*, Las Vegas, USA, 2008, pp. 4353–4356.

[8] Otterson, S., "Improved Location Features for Meeting Speaker Diarization," in *Proc. Interspeech '07*, Antwerp, Belgium, 2007, pp. 1849–1852.

[9] Sundaram, N. et al., "Usable speech detection using linear predictive analysis - a model based approach," in *Proc. of ISPACS*, Awaji Island, Japan, 2003, pp. 231–235.

[10] Yantorno, R., "The Spectral Autocorrelation Peak Valley Ratio (SAPVR) – A Usable Speech Measure Emplyed as a Co-Channel Detection System," in *Proc. of IEEE Workshop on Intelligent Signal Processing*, 2001.

[11] Brandstein, M. S. and Silverman, H. F., "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. ICASSP '97*, Munich, Germany, 1997, pp. 375–378.

[12] Luque, J. et al., "Speaker Diarization for Conference Room: The UPC RT07s Evaluation System," in *Multimodal Technologies for Perception of Humans*, LNCS, vol. 4625/2008, Springer Berlin/Heidelberg, 2008, pp. 543–553.