



# Overlap Detection for Speaker Diarization by Fusing Spectral and Spatial Features

Martin Zelenák, Carlos Segura, Javier Hernando

Universitat Politècnica de Catalunya, Barcelona, Spain

{martin.zelenak, carlos.segura, javier.hernando}@upc.edu

## Abstract

A substantial portion of errors of the conventional speaker diarization systems on meeting data can be accounted to overlapped speech. This paper proposes the use of several spatial features to improve speech overlap detection on distant channel microphones. These spatial features are integrated into a spectral-based system by using principal component analysis and neural networks. Different overlap detection hypotheses are used to improve diarization performance with both overlap exclusion and overlap labeling. In experiments conducted on AMI Meeting Corpus we demonstrate a relative DER improvement of 11.6% and 14.6% for single- and multi-site data, respectively.

**Index Terms:** speaker overlap detection, speaker diarization, cross-correlation

## 1. Introduction

Simultaneous speech by two or more speakers is a naturally occurring event in human conversation. For instance in a meeting environment, people tend to give backchannel to the leading speaker, or try to grab floor from him, herewith creating speaker overlaps. This phenomenon is observable even with few people involved in the conversation [1]. Overlapped speech poses a problem for many automatic human language technologies, speaker diarization being one of them. The speaker diarization task aims to give answer to the question “*Who spoke when?*”, in general, without any prior information about the speakers. The drawback of conventional diarization systems concerning overlapped speech is that they are able to assign only one speaker label per segment, which, obviously, leads to missed speech for overlapped speakers. Furthermore, including simultaneous speech into the creation of cluster models can be a potential source of speaker error, since the models are less pure.

Several works on the detection of overlapped speech in meetings make use of personal (close-talking) microphones. The usual way is to segment each of the individual speaker channels with an ergodic hidden Markov model (HMM). In [2], overlap was marked in a post-processing step based on cross-correlation analysis, whereas in [3], overlapped speech was one of the decoding classes. A more simple solution, without the necessity of training a model, can be found in [4].

A few of presented algorithms employ distant microphones exclusively. For instance, the authors in [5] proposed to use microphone pair time delays to segment audio according to a fixed number of speakers. They assumed the location of speakers will not change and showed the possibility to detect two simultaneous speakers by modeling short-term turns for each speaker combination. The method suggested in [6] used a previous diarization segmentation to create a Gaussian mixture model

(GMM) for all pairs of detected speakers. These were then integrated with individual cluster states into a new HMM and the meeting data was resegmented again. Even though overlap was detected with this approach, it did not lead to decreasing the diarization error. Overlap detection system for single distant channel presented in [7] is an HMM-based segmenter that utilized various features, e.g., cepstrum, entropy, modulation spectrogram, etc. Detected overlaps were subsequently applied in diarization and in experiments on AMI corpus the authors were able to achieve up to 6.8% relative DER improvement on multi-site data.

In this paper we are proposing to use several cross-correlation-based spatial features for improving overlap detection on distant channel data and to integrate them into a spectral-based overlap detection system. The dimensionality of spatial feature space is very high and can also vary across different meeting rooms. To deal with these issues we are suggesting two strategies for fusion of spectral and spatial information. In the first, we reduce and unify the size of spatial feature vectors with principal component analysis (PCA), similar approach was also chosen for diarization purposes in [8]. In the second strategy, a multilayer perceptron artificial neural network (ANN) is initially trained with spatial information. Then, sets of spatial features for all microphone pairs are sequentially fed to the ANN to obtain a classification score. This score is later added to the spectral vector.

There are two ways how detected overlapped speech can aid diarization. Overlaps can be excluded from cluster-building stage with the purpose of decreasing speaker error and an extra speaker label can be given for these segments to recover some of the missed speech. These two techniques are quite different in their manner, hence we believe that diarization would benefit more if each technique would use its own tailored overlap detection hypothesis instead of using just a single hypothesis for both. Proposed methods were evaluated on AMI Meeting Corpus on single- and multi-site recordings.

This paper is organized as follows. Baseline overlap detection is described in Section 2. Spatial features and fusion are detailed in Section 3 and 4. Speaker diarization system and its improvements are briefly outlined in Section 5. Experimental results and conclusions are given in Section 6 and 7.

## 2. Baseline overlap detection

Baseline overlap detection utilizes a number of spectral-based features, namely 12 MFCCs extracted every 10 ms over a window of 30 ms, residual energy of a 12th-order LPC (LPCRE) computed over a 25 ms window, spectral flatness (SF) over 30 ms and first order differences. Spectral flatness was applied for discrimination between speech and non-speech [9], but can eventually convey information about the number of speakers

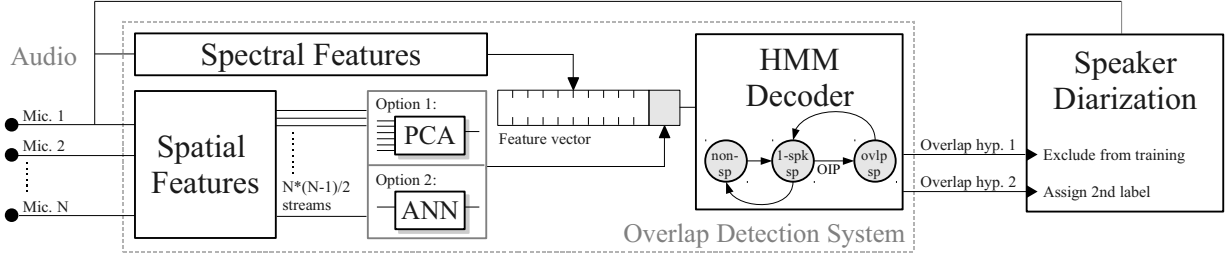


Figure 1: Overlap detection system diagram

speaking [7]. It is defined as the ratio between geometric and arithmetic mean of a certain number (100 in our case) of spectral magnitudes

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

Features were mean-variance normalized according to global statistical moments of the training data.

Similar to [7], our system considers three acoustic classes representing non-speech, single-speaker speech and overlapped speech. For a more accurate modeling of transitions between them a three-state HMM was defined for each class, where each state's feature-vector distribution is modeled with a diagonal-covariance GMM. Since the amount of training data for the three classes is not balanced, we are using 256 Gaussian components for single-speaker speech and 32 or 64 components for overlap and non-speech. GMMs are created by iterative Gaussian splitting technique and subsequent re-estimation.

Detection hypothesis is obtained by Viterbi decoding and applying a word network whose topology is depicted in Fig. 1 in the *HMM decoder* block. For precision purposes the transition from single-speaker speech to overlapped speech can be penalized with an overlap insertion penalty (OIP) and some transitions are completely forbidden.

### 3. Spatial features

The cross-correlation function is well-known as a measure of the similarity between signals for any given time displacement and ideally its maximum lies in correspondence to the delay between the pair of signals [10]. A commonly used technique to estimate the time delay that performs robustly in reverberant environments is the Generalized Cross Correlation with Phase Transform weighting (GCC-PHAT) [11]. For a pair of microphones  $ij$ , it can be expressed in terms of the inverse Fourier transform of the estimated cross-power spectrum,  $G_{ij}(f)$ , as follows,

$$R_{ij}(\tau) = \int_{-\infty}^{\infty} \frac{G_{ij}(f)}{|G_{ij}(f)|} e^{j2\pi f\tau} df \quad (2)$$

And the time delay estimation (TDE) is as follows:

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

The value of the GCC-PHAT peak provides a measure of the coherence between signals independent of the microphone gains or the signal powers, and depends on the distance between microphones, the distance between acoustic source and microphone pair and on the environmental noise and reverberation conditions.

In situations dealing with multiple, possibly moving, concurrent speakers, we have observed that the time delay estimates 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. The maximum value of the cross-correlation sequence is also lower than in the single speaker situation, since multiple speakers introduce random peaks, which in the general case attenuate the main peak.

Based on these observations we are proposing several cross-correlation-based spatial features for every microphone pair that provide some degree of information on speaker overlaps.

An easily observable feature is the *coherence value*, defined in eq. 4. This is the value of principal peak of the GCC, and in ideal condition should be high for single-source situations. Noises, reverberation and concurrent acoustic sources will lower this value.

$$C_{ij} = \max(R_{ij}(\tau)) \quad (4)$$

Derived from the coherence value, we are also proposing to extract the coherence *dispersion ratio*, as shown in eq. 5. 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 as follows,

$$D_{ij} = \frac{C_{ij}^2}{\sum_{t=-w_{ij}}^{w_{ij}} R_{ij}^2(\hat{\tau}_{ij} + t)}, \quad (5)$$

where the size of the window  $w_{ij}$  is adjusted to TDE standard deviation of the microphone pair  $ij$ , that is related with the possible delay range that can be measured by the pair. Note that the window length  $w_{ij}$  varies for different microphone pairs.

Finally, the *delta of TDE* for every microphone pair also carries information on overlaps. The first order derivative of the TDE is high in situations where the speaker is moving, multiple non-concurrent speakers change turns or multiple speakers talk simultaneously.

### 4. Fusion

One of the main problems that arise is the high dimensionality of spatial feature vectors. A recording involving 12 microphones yields to 66 pairs and 198 features. Also the number of microphones differs from site to site, making it difficult to train a general model. Our first strategy for dimensionality reduction and unification is the application of PCA, which transforms the original feature space into a new coordinate system with the greatest variance lying on the first component. We estimated a separate transformation matrix for every discussed spatial feature for each site and then used just the first principal component. Hence, in the given example with 12 microphones we

would end up with one transformed coherence, one dispersion and one delta TDE.

Another issue is that the proposed spatial features are, in general, not comparable across different microphone pairs, since they are intrinsically tied to physical characteristics of the pair like the inter-microphone distance. To normalize the spatial features and reduce their dimensionality we are considering a four-layer perceptron. The input of the ANN is composed by 6 input neurons, 3 for spatial features and 3 for normalization values (*mean of coherence*, *variance of coherence*, *variance of TDE*) for every pair. The output is a score classifying between overlap and non-overlap, which is comparable across microphone pairs. For a given frame the average score was taken.

Spatial information is modeled in the HMM with a separate Gaussian mixture which shares means and variances across states. The output probabilities are weighted in the ratio 0.75 and 0.25, for spectral and spatial feature stream, respectively. The weights were set empirically. A schematic architecture of our overlap detection system with a link to speaker diarization is shown in Fig. 1.

## 5. Speaker diarization system

Our speaker diarization system 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 system is described in detail in [12]. The performance of diarization is evaluated by means of diarization error rate (DER), which is the sum of missed speaker time, false alarms and speaker error.

The overlap extension to diarization system comprises the exclusion and labeling of simultaneous speech. The first technique means to discard overlap frames from cluster initialization and GMM training with the intention of obtaining more precise models. In the latter technique, Viterbi decoding selects for these segments besides the most likely cluster also the second most likely. In this way the missed speaker time will 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. The diarization system is using reference speech segments as well.

## 6. Experiments

### 6.1. Corpus

Experiments were conducted on AMI corpus, which consists of 100 hours of meeting recordings, on 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 site. Data was 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 of the overlap detection system and evaluation is performed with force-aligned annotations obtained by SRI's DECIPHER recognizer.

### 6.2. Overlap detection results

Overlap detection experiments have been done for three setups, for spectral system, for fusion of spectral feature with PCA-transformed spatial features (*Spect+XC-PCA*) and for spectral features with spatial ANN score (*Spect+XC-ANN*). Performance is measured with Recall—true detected overlaps, Precision—ratio between true overlaps and all detected overlaps, and with Error—the sum of missed and false overlaps. Results depend very much on the value of the overlap insertion penalty, which controls the amount of overlaps the system will detect. 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 highest recall, F-ratio, lowest error and acceptable high precision.

The detection performance on single-site recordings is given in Fig. 2a. In the lower penalty region both spatial setups outperform the spectral in error, recall and precision, whereas for higher penalization the differences are becoming smaller. The spatial PCA setup achieves the lowest error 73% corresponding to precision 80% and recall 35%. Results of the more difficult multi-site scenario are shown in Fig. 2b. Obviously, the overall performance is significantly worse than before. Spatial PCA setup seems again to be the best for low penalties, most clearly in terms of error. But with increasing penalization the errors are almost alike and in precision the places are switched with spectral and spatial ANN setup, which also achieves the best result with 83% error, 76% precision and 25% recall.

### 6.3. Speaker diarization results

The complement of the overlap detection error tells us how much the diarization can possibly gain by labeling detected overlap segments, 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. High precision is also important. A relationship between overlap exclusion performance and some of the detection metrics is not clear. In general, it can be presumed that a high recall hypothesis will be working better. Following these guidelines, we selected for each setup two or three detection hypotheses for exclusion and for labeling and then selected the best performing to do both.

The DER values and relative improvements for single-site data are given in Table 1. All setups are yielding improvements over the baseline diarization, with the best relative improvement by spatial PCA setup of 11.6%. This corresponds with the results of overlap detection in Fig. 2a, where spatial PCA was the overall best performing setup. Somehow surprising is that spatial ANN could not turn good improvements by exclusion and labeling separately also to higher combined performance.

The relative diarization improvements observed on multi-site data are even better than in the previous case, the results are presented in Table 2. In this scenario, the spatial PCA setup does not confirm its best performance from before, falling even slightly behind spectral setup. Presumably, the PCA transformation does not unify spatial information from various sites as good as the ANN, which achieves the best performance with improvement of up to 14.6%. In all evaluation, none of the best performing hypotheses for labeling performed also the best for exclusion and vice versa. This proves our initial assumption that diarization improvement will be better with two different overlap hypotheses.

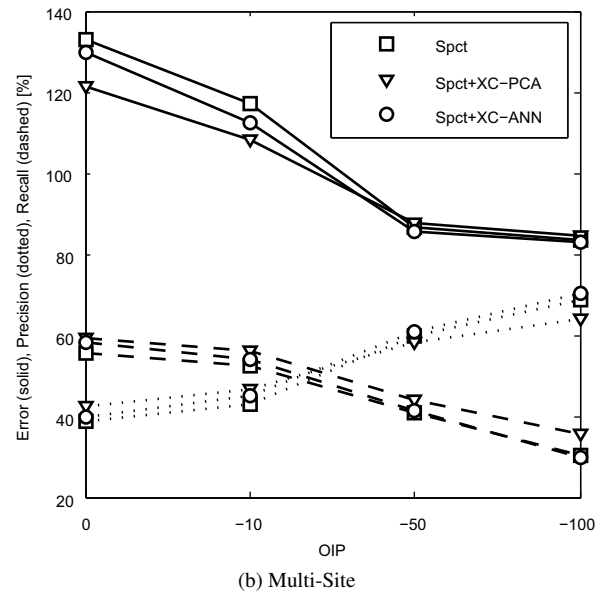
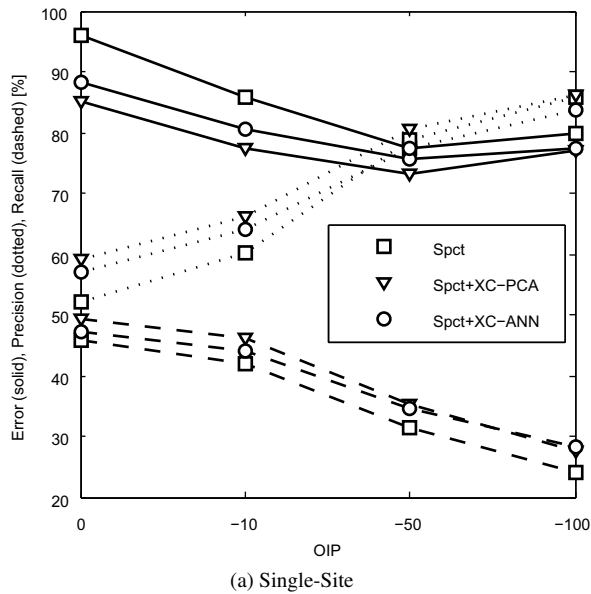


Figure 2: Overlap detection performance for (a) single- and (b) multi-site data using spectral features only (Spct), spectral with PCA-fused (Spct+XC-PCA) and ANN-score-fused spatial features (Spct+XC-ANN). Error—solid, Prec.—dotted, Recall—dashed line.

Table 1: Speaker diarization with overlapped speech on single-site data, DER and rel. improvement over the baseline (in %)

Baseline	38.3		
Overlap det.:	+Exclusion	+Labeling	+Both
Spct	36.8/+3.9	36.4/+4.9	<b>35.6/+6.9</b>
Spct+XC-PCA	36.3/+5.2	36.2/+5.5	<b>33.8/+11.6</b>
Spct+XC-ANN	37.1/+3.1	36.3/+5.1	<b>36.1/+5.7</b>

Table 2: Speaker diarization with overlapped speech on single-site data, DER and rel. improvement over the baseline (in %)

Baseline	37.3		
Overlap det.:	+Exclusion	+Labeling	+Both
Spct	34.5/+7.5	36.6/+2.1	<b>33.5/+10.2</b>
Spct+XC-PCA	34.7/+6.9	36.7/+1.7	<b>33.8/+9.5</b>
Spct+XC-ANN	32.8/+12.1	36.5/+2.3	<b>31.9/+14.6</b>

## 7. Conclusions

We have proposed three new cross-correlation-based spatial features for the detection of overlapped speech. Spatial and spectral information were fused applying either PCA or an ANN, and promising results were achieved in comparison with spectral baseline system. Speaker diarization was especially improved by handling overlap segments detected by spatial PCA setup on single-site and by spatial ANN setup on multi-site data. Furthermore, our suggestion to use independent detection hypotheses for overlap exclusion and labeling has been successful.

## 8. Acknowledgements

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. The authors would like to thank K. Boakye (ICSI) for providing ASR-aligned annotations.

## 9. 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] 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.
- [3] Wrigley, S.N. et al., "Speech and Crosstalk Detection in Multi-channel Audio," *IEEE Transactions on Speech and Audio Processing*, vol 13, pp. 84–91, 2005.
- [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., Vinyals, O., and Friedland, G., "Two's a Crowd: Improving Speaker Diarization by Automatically Identifying and Excluding Overlapped Speech," in *Proc. Interspeech '08*, Brisbane, Australia, 2008, pp. 32–35.
- [8] Otterson, S., "Improved Location Features for Meeting Speaker Diarization," in *Proc. Interspeech '07*, Antwerp, Belgium, 2007, pp. 1849–1852.
- [9] Yantom, R., "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.
- [10] Svaizer, P. et al., "Acoustic source location in a three-dimensional space using crosspower spectrum phase," in *Proc. ICASSP '97*, Munich, Germany, 1997, pp. 231–234.
- [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.