Detection and Handling of Overlapping Speech for Speaker Diarization

Martin Zelenák and Javier Hernando

TALP Research Center,
Department of Signal Theory and Communications,
Universitat Politècnica de Catalunya
C. Jordi Girona 31, 08034 Barcelona, Spain
martin.zelenak@upc.edu

Abstract. This thesis concerns the detection of overlapping speech segments and its further application for the improvement of speaker diarization performance. We propose the use of three spatial cross-correlation-based parameters for overlap detection on distant microphone channel data. Spatial features from different microphone pairs are fused by means of principal component analysis or by an approach involving a multilayer perceptron. In addition, we investigate the possibility of employing long-term prosodic information. The most suitable subset of candidate prosodic features is determined by a two-step mRMR feature selection algorithm. For segments including detected overlapping speech the speaker diarization system picks a second speaker label, and such segments are also discarded from the model training. The proposed overlap labeling technique is integrated in the Viterbi-decoding part of the diarization algorithm.

Keywords: overlapping speech detection, speaker overlap, speaker diarization, spatial features, cross-correlation, prosody

1 Introduction

It is a well known fact that people sometimes tend to speak at the same time, i. e., simultaneously. It is a normal part of human conversation behavior. For that reason, audio recordings of meetings commonly include regions of overlapping speech. A lot of spoken language technologies suffer from this conversation phenomenon, one of them is speaker diarization.

Given a speech recording, *speaker diarization* aims to answer the question: *"Who spoke when?"*, generally without any prior knowledge. The application of diarization systems is often a very useful preprocessing step for other audio technologies such as ASR. Meetings are considered the most difficult application domain due to high spontaneity of speech, variable microphone signal quality, and room reverberation.

According to several studies [1, 2], a portion of the performance degradation on real meeting data can be directly associated with the occurrence of speaker overlaps. Nevertheless, the number of published proposals is rather limited and dealing with overlapping speech still remains a challenging problem.

The discussed thesis[1] addresses the issues related to the occurrence of simultaneous speech in meeting recordings. The motivation is to improve speaker diarization performance. However, the investigation of overlapping speech may also be useful beyond diarization, e.g., for speech, or speaker recognition. This paper briefly summarizes the outcomes of the thesis work and highlights some of the achieved results.

2 Motivation and Thesis Objectives

A common drawback of conventional diarization systems is that they are only able to assign one speaker label per segment. In cases when a segment contains simultaneous speech, this implicitly leads to missed speech errors. In addition, another possible effect of overlaps on diarization performance discussed in [1] is that speaker models could be corrupted if simultaneous speech is included into the training data.

The first main goal of the thesis was the development of a robust overlap detection system. The requirement was to work with distant channel data without any constraints about microphone configuration or the recording room. Our interest was to research and propose new features which may be useful for this task. We aimed at exploring the possibilities of employing spatial-based information for the detection of simultaneous speech since (smart) meeting rooms are normally equipped with microphone arrays. The availability of multi-channel data provides the option to estimate features that are in some way related to spatial location. Further option was to investigate the potential of higher-level information. "Higher" in this case refers to speech information which is above the level of short-term spectral or cepstral features, such as prosody.

Another main goal of this thesis was to apply the detected overlapping speech in the UPC speaker diarization system in order to reduce diarization error. This should be achieved by both recovering missed speaker time, as well as by improving the clustering. We sought to implement a novel technique for the assignment of extra speaker labels in speaker overlap segments. Different overlap detection systems were examined according to the quality of their hypotheses for diarization improvement.

Finally, since our general intention was contribute to the research in human language processing, we participated in the organization of the Albayzin evaluation campaign. Being in charge of the speaker diarization section, we presented the final results in [3] and [4].

3 Overlapping Speech Detection

Our baseline overlap detection system was firstly defined in [5], it utilizes a number of spectral-based features, such as MFCCs, LPC residual energy [6], spectral

[1] Thesis manuscript is accessible at: `http://www.tdx.cat/handle/10803/72431`

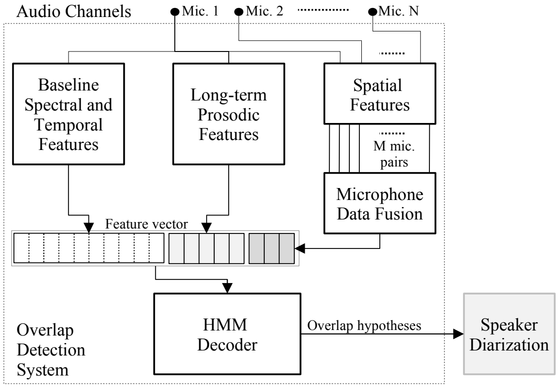

Fig. 1. Overlap detection system diagram.

flatness [7], and deltas. Following classes are considered to classify the signal and produce an output hypothesis by means of Viterbi decoding: non-speech (e. g., silence, noise), single-speaker speech, and overlapping speech. For each class a three-state hidden Markov model (HMM) is defined. Since the amount of training data is not balanced among classes, we use a higher number of Gaussian components for single-speaker speech than for overlapping speech and non-speech. Transition probabilities between different HMMs are set manually. In order to increase the precision, the transition from single-speaker speech to overlapping speech can be penalized with an overlap insertion penalty (OIP), and the transitions from overlapping speech to non-speech and vice versa are forbidden. The diagram showing the system architecture is given in Fig. 1.

3.1 Novel Spatial-based Features

The generalized cross-correlation function with phase transform weighting (GCC-PHAT) is a commonly used measure of the similarity between signals performing well in reverberant environments [8]. GCC-PHAT exhibits a prominent peak at the elapsed time corresponding to the dominant sound source in the room, minimizing the peaks of the non-dominant sources and reverberation at the same time.

Our hypothesis was that in case of multiple, possibly moving, concurrent speakers, the time delay of arrival (TDOA) estimates produced by the GCC-PHAT will jump from one speaker to another at a very high rate as one source dominates due to the non-stationarity of the voice. TDOA can be expressed as follows,

$$\hat{\tau}_{mn} = \arg\max_{\tau} R_{mn}(\tau), \tag{1}$$

where $R_{mn}(\tau)$ is the GCC-PHAT function for a pair of microphones m and n. The maximum value of the cross-correlation sequence should also be lower than in the single speaker situation, since multiple speakers introduce random peaks, which attenuate the main peak. We proposed three cross-correlation-based spatial features for every microphone pair, which are intended to provide some degree of information on speaker overlaps [5, 9].

The first is the *coherence value*, defined in (2), which is the principal peak value of the GCC-PHAT. In ideal conditions it should be high for single-source situations, while the presence of noise, reverberation, and concurrent acoustic sources attenuate this value.

$$C_{mn} = \max(R_{mn}(\tau)) \tag{2}$$

The second feature, coherence *dispersion ratio*, is derived from the coherence value. It is defined as follows,

$$D_{mn} = \frac{C_{mn}^2}{\sum_{t=-w_{mn}}^{w_{mn}} R_{mn}^2(t + \hat{\tau}_{mn})}. \tag{3}$$

This value is computed as the ratio between the square of the main peak value and the square quadratic sum of the cross-correlation values under a time delay window w_{mn}. The size of the window w_{mn} varies for different microphone pairs and it is set to the TDOA standard deviation of each pair. In this way, the dispersion ratio measures the relation between the energy of the main peak and the energy that is scattered in its neighborhood. Similar to the coherence feature (2), the dispersion ratio is close to 1 in the case of a single speaker and ideal conditions, while it has a lower value in reverberant conditions or concurrent acoustic sources situations.

Finally, the *delta of TDOA* obtained by (1) for every microphone pair also carries information on overlaps. The derivative of the TDOA is high in situations where the speaker is moving, multiple non-concurrent speakers change turns at talk, or multiple speakers talk simultaneously.

3.2 Microphone Data Fusion

Practical issues with the use of spatial features are the high and variable dimensionality of vectors. Especially the latter—sites may have different number of microphones—makes it difficult to train a general model. One of the strategies for dimensionality reduction and normalization is the application of a principal component analysis (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 kind per each site, and then we use just the first principal component.

We also considered an alternative approach to reduce the spatial vector dimensionality based on a multi-layer perceptron (MLP). The input of the MLP is composed by 6 input neurons, 3 for spatial features and 3 for normalization values (mean of coherence, variance of coherence, variance of TDOA) for every pair. The output is a binary score classifying between overlap and non-overlap, which is commensurable across microphone pairs. For a given frame the average score was taken and merged with corresponding spectral feature vectors.

3.3 Prosodic Features

Prosody, in general, is characterized by rhythm, intonation, stress, and juncture in speech. A few studies were published that researched the relationship between prosodic cues and the interaction of conversation participants. For instance, it was suggested that stretches of low pitch can trigger backchannel feedback from listener (*yeah, uh-huh, right*) [10]. It was also shown that speakers raise their voices when starting their utterance during somebody else's talk, compared to starting in silence [11].

For the detection of overlapping speech a number of prosody-based features are considered [12, 13]. They can be assigned to one of the following categories: pitch, intensity and (four) formant frequencies. In addition, for each of these categories we estimate besides the actual value for every given time point also long-term statistics such as mean, median, min., max., std. deviation and the difference between the min. and max. value. Missing values, such as F0 estimates for unvoiced speech, or parameters in non-speech regions, are substituted with default values. Non-speech regions are not considered for the computation of the statistical parameters.

The use of both instantaneous prosodic features and their long-term statistics is a source of redundancy in our prosodic feature set. Due to this fact, we have added a two-stage feature selection procedure to our feature extractor. The stages are as follows. In the first stage, we applied a mRMR algorithm [14] on held-out development data to score individually the candidate features against the target class (overlapping vs. single-speaker speech) and sorted them according to their minimum redundancy and maximal relevance.

The second feature selection stage involves conventional hill climbing wrapper approach, i. e., iteratively adding candidate features to the feature set, creating a model and evaluating the system on the development data.

3.4 Overlap Detection Experiments

Experiments were conducted on the AMI corpus, on far-field microphone array channels. We defined a single- and a multi-site scenario. The first included recordings only from Idiap site and the latter also from Edinburgh and TNO site. The average amount of overlap 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.

Performance is measured with recall (ratio between true detected and reference overlap time), precision (ratio between true and all detected overlap time), and with the sum of missed and false overlap time divided by the reference overlap time, referred to as detection error. To make a reasonably fair evaluation, results are measured at four operation points defined by the OIP value ($OIP = \{0, -10, -50, -100\}$).

The detection performance for different feature combination setups is given in Fig. 2. For single-site data it can be seen that the overlap detection also using spatial or prosodic parameters outperformed the baseline system (*Spct*).

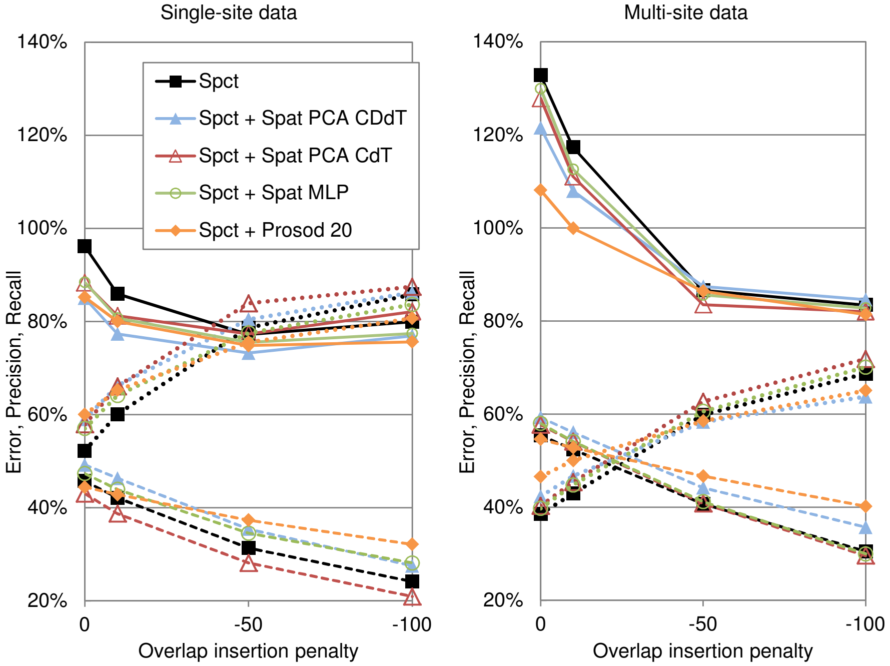

Fig. 2. Overlap detection performance for AMI (left) single- and (right) multi-site data using spectral features (Spct), and in combination with spatial (+ Spat) or prosodic (+ Prosod 20) features. Detection error, precision, and recall are delineated with solid, dotted, and dashed line, respectively.

We can conclude that the PCA-fused spatial feature including dispersion ratio (*Spct + Spat PCA CDdT*) is well suited for the single-site condition. The system using MLP score (*Spct + Spat MLP*) has a good detection performance in this scenario, but for high OIP its precision drops bellow the one of the baseline system.

For the multi-site scenario, however, the mentioned PCA-fused dispersion ratio (marked with D) seems to lack robustness. The possible reason for the worse performance of feature setups involving this parameter is its dependency on the spatial distribution of microphones, which might be an issue in case of using multiple recording rooms. Moreover, the limited ability to compensate for the variability of this scenario can most likely be attributed to the simplicity of the PCA technique.

In the multi-site case the better performing setup included only spatial coherence and delta TDOA (*Spct + Spat PCA CdT*), but the distinction in performance between setups including and not including dispersion ratio becomes evident only at higher OIPs. The MLP technique combines all three spatial parameters more effectively in this scenario and outperforms, or at least equals, the baseline system. In general, the less precise multi-site models need a higher amount of overlap penalization to arrive to the lowest detection errors. The addition of prosodic features decreased the overlap detection error in both scenarios either due to higher precision for low penalties or due to improved recall in high penalty region.

4 Speaker Diarization

Our speaker diarization system, detailed in [15], follows the commonly used agglomerative clustering approach and relies on 20 MFCCs extracted from 30 ms frames. The algorithm starts with an uniform initial segmentation where the number of initial clusters is determined automatically. Clusters are modeled with Gaussian mixture models (GMMs) and cluster pair merging in each iteration is driven by Bayesian information criterion (BIC). The complexity of GMMs is also determined automatically based on the amount of data corresponding to a particular cluster.

The system can be improved by multi-channel approach based on conventional techniques. We applied speech signal techniques such as Wiener filtering and beamforming for signal enhancement, and we also combined the time-delay-of-arrival (TDOA) information as a second stream in the diarization [16].

The performance of the speaker diarization was evaluated by means of the diarization error rate (DER). Defined by NIST, the DER is a time-weighted metric composed of the sum of missed speaker time, false alarms and speaker error time.

4.1 Handling Overlapping Speech

Overlap handling in diarization comprises the labeling and/or exclusion of simultaneous speech. The first 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. Overlap exclusion blocks overlap frames from being included into cluster initialization and HMM training, but does not prevent decoding them. The aim of this technique is to get lower speaker detection error rates with more precise clusters. The concept is depicted in Fig. 3.

Both techniques work independently from each other, or better said sequentially, since the exclusion works throughout the diarization process whereas the labeling is performed at the end of the iteration process. These two techniques do not necessarily have to share the same overlap hypothesis, they can be optimized independently and can possibly use two different hypotheses, i. e., one for each technique. This method was firstly suggested in [5].

4.2 Speaker Diarization Experiments

Baseline DERs and relative improvements by applying overlap exclusion and labeling in experiments on AMI data are given in Table 1. The most successful overlap detection setup on single-site data was the combination of spectral

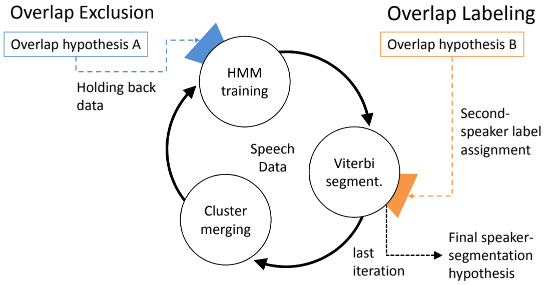

Fig. 3. Overlap handling concept in speaker diarization algorithm.

and the three PCA-transformed spatial parameters (*Spct+Spat PCA CDdT*), yielding improvement of 11.2% relative. A good result was also obtained with the combination of spectral and prosodic features. However, the result was not much higher compared to the *Spct* case.

In the multi-site scenario the relative improvements were higher, the best observed result of 17.0% DER reduction was with the combination of spatial coherence and delta TDOA (*Spct+Spat PCA CdT*). Another successful system with 13.9% improvement was the one using MLP for the spatial parameter fusion.

Table 1. Speaker diarization with exclusion and labeling of simultaneous speech detected by different systems, baseline DER and relative improvements over the diarization baseline (in %).

Overlap det.	Single-site +Ovlp. Excl.	Multi-site and Labl.
Baseline	38.3	37.3
Spct	+6.9	+6.7
Spct+Spat (PCA) CdT	+5.7	+17.0
Spct+Spat (PCA) CDdT	+11.2	+8.0
Spct+Spat MLP	+5.7	+13.9
Spct+Prosod 20	+7.2	+11.1

Table 2 shows the comparison of our Viterbi-integrated labeling technique to two simple labeling schemes in terms of relative DER improvement. The first of these techniques a posteriori attributes the overlapping speaker label according to the nearest neighboring (NNeigh.) speaker, as in [17]. The other competing technique assigns the overlapping label to the most talkative speaker (MTalk.) [18]. In case the most talkative speaker has already been picked by the diarization system, the second most talkative speaker is selected. In general, the differences between DER of the three labeling techniques are small, but it can be seen that the results of the technique proposed in this thesis are competitive, in single-site scenario in particular.

Table 2. Relative improvements of speaker diarization by different labeling strategies.

	Single-site			Multi-site		
Overlap det. +Labl.	Vit.	NNeigh.	MTalk.	Vit.	NNeigh.	MTalk.
Spct	**+5.3**	+5.1	+4.8	**+2.7**	+3.2	+2.3
Spct+Spat	**+6.2**	+5.9	+5.6	**+3.4**	+3.9	+3.0

5 Discussion

This work deals with the issues of overlapping speech in the context of speaker diarization on distant microphone channels. In order to locate the regions where multiple speakers are speaking simultaneously, an overlap detection system was built. We have found that spatial information can be utilized to perform this detection and proposed three novel cross-correlation-based features. The problem of high and variable dimensionality of spatial feature space was addressed with the application of a per-site-specific PCA, or an MLP neural network. Furthermore, we have also introduced features based on prosody and their long-term statistics.

Objectively, the overlap detection performance has still a lot of potential for improvement. The task proved to be extremely challenging for an automated system, but in a lot of cases it is difficult even for humans to decide what can and what cannot be considered overlapping speech.

By handling of the detected simultaneous speech segments, we managed to improve the baseline speaker diarization system. With the objective to build more precise speaker models, the speech frames including overlapping speech were excluded from the training process. In addition, we reduced diarization's missed speech by assigning second speaker labels for speaker overlap segments. Analyses beyond the scope of this paper showed that the performance of overlap exclusion exhibits a relatively non-stable nature [9]. Our proposed labeling technique delivers competitive results compared to alternative simple strategies, especially in the single-site scenario.

References

1. Otterson, S., Ostendorf, M.: Efficient use of overlap information in speaker diarization. In: Proc. ASRU '07, Kyoto, Japan (2007) 683–686
2. Huijbregts, M., Wooters, C.: The blame game: Performance analysis of speaker diarization system components. In: Proc. Interspeech '07, Antwerp, Belgium (2007)
3. Zelenák, M., Schulz, H., Hernando, J.: Albayzin 2010 Evaluation Campaign: Speaker Diarization. In: Proc. FALA 2010, Vigo, Spain (2010) 301–304
4. Zelenák, M., Schulz, H., Hernando, J.: Speaker Diarization of Broadcast News in Albayzin 2010 Evaluation Campaign. EURASIP Journal on Audio, Speech, and Music Processing **2012**(19) (2012)
5. Zelenák, M., Segura, C., Hernando, J.: Overlap Detection for Speaker Diarization by Fusing Spectral and Spatial Features. In: Proc. Interspeech '10, Makuhari, Japan (2010) 2302–2305

6. Boakye, K., Trueba-Hornero, B., Vinyals, O., Friedland, G.: Overlapped speech detection for improved speaker diarization in multiparty meetings. In: Proc. ICASSP '08, Las Vegas, NV, USA (2008) 4353–4356
7. Yantorno, R.E., Krishnamachari, K.R., Lovekin, J.M., Benincasa, D.S., Wenndt, S.J.: The spectral autocorrelation peak valley ratio (SAPVR) – a usable speech measure employed as a Co-Channel detection system. In: Proc. IEEE International Workshop on Intelligent Signal Processing (WISP), Budapest, Hungary (2001)
8. Brandstein, M., Silverman, H.: A robust method for speech signal time-delay estimation in reverberant rooms. In: Proc. ICASSP '97, Munich, Germany (1997) 375–378
9. Zelenák, M., Segura, C., Luque, J., Hernando, J.: Simultaneous Speech Detection with Spatial Features for Speaker Diarization. IEEE Transactions on Audio, Speech, and Language Processing, Special Issue on New Frontiers in Rich Transcription **20**(2) (Feb. 2012) 436–446
10. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics **32/2000** (2000) 1177–1207
11. Shriberg, E., Stolcke, A., Baron, D.: Can prosody aid the automatic processing of multi-party meetings? Evidence from predicting punctuation, disfluencies, and overlapping speech. In: Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding, Red Bank, NJ, USA (2001) 13–16
12. Zelenák, M., Hernando, J.: The Detection of Overlapping Speech with Prosodic Features for Speaker Diarization. In: Proc. Interspeech '11, Florence, Italy (2011) 1041–1044
13. Zelenák, M., Hernando, J.: Speaker Overlap Detection with Prosodic Features for Speaker Diarization. IET Signal Processing (2011) in review.
14. Peng, H., Long, F., Ding, C.: Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(8) (Aug. 2005) 1226–1238
15. Luque, J., Anguera, X., Temko, A., Hernando, J.: Speaker diarization for conference room: The UPC RT07s evaluation system. In Stiefelhagen, R., Bowers, R., Fiscus, J., eds.: Multimodal Technologies for Perception of Humans. Volume 4625/2008 of LNCS. Springer Berlin / Heidelberg (2008) 543–553
16. Anguera, X., Wooters, C., Hernando, J.: Acoustic beamforming for speaker diarization of meetings. IEEE Transactions on Audio, Speech, and Language Processing **15**(7) (2007) 2011–2022
17. Huijbregts, M., van Leeuwen, D., de Jong, F.: Speech Overlap Detection in a Two-Pass Speaker Diarization System. In: Proc. Interspeech '09, Brighton, UK (2009) 1063–1066
18. van Leeuwen, D.A., Konečný, M.: Progress in the AMIDA Speaker Diarization System for Meeting Data. Multimodal Technologies for Perception of Humans **4625/2008** (2008) 475–483