The Detection of Overlapping Speech with Prosodic Features for Speaker Diarization.

Conference Paper · January 2011

Source: DBLP

2 authors:

Martin Zelenák
Universitat Politècnica de Catalunya
7 PUBLICATIONS **91** CITATIONS
SEE PROFILE

Javier Hernando
Universitat Politècnica de Catalunya
251 PUBLICATIONS **2,167** CITATIONS
SEE PROFILE

The Detection of Overlapping Speech with Prosodic Features for Speaker Diarization

Martin Zelenák, Javier Hernando

Universitat Politècnica de Catalunya, Barcelona, Spain

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

Abstract

Overlapping speech is responsible for a certain amount of errors produced by standard speaker diarization systems in meeting environment. We are investigating a set of prosody-based long-term features as a potential complement to our overlap detection system relying on short-term spectral parameters. The most relevant features are selected in a two-step process. They are firstly evaluated and sorted according to mRMR criterion and then the optimal number is determined by iterative wrapper approach. We show that the addition of prosodic features decreased overlap detection error. Detected overlap segments are used in speaker diarization to recover missed speech by assigning multiple speaker labels and to increase the purity of speaker clusters.

Index Terms: overlapping speech detection, prosody, feature selection, speaker diarization

1. Introduction

Human conversation often includes certain amount of overlapping speech. Several works identified these specific conversation events as a challenge for many automatic human language technologies [1, 2]. One of these technologies is speaker diarization, which, given a recording, strives to answer the question *“Who spoke when?”* without any prior knowledge about the speakers. The problem is that conventional diarization systems assign only one speaker label per segment and, consequently, miss speech from overlapping speakers. Furthermore, it is reasonable to assume that overlapping speech included into the training data of a single-speaker model can lead to some level of corruption of the models.

Prosody describes the rhythm, intonation and stress of speech. It can reflect various things about the speaker or the utterance, e. g., the emotional state. There has been significant effort to use this kind of higher-level speech information for various tasks like speaker verification and identification. Recently, prosodic features were also successfully applied for speaker diarization [3, 4].

A few studies were published which researched the relationship between prosodic cues and the interaction of conversation participants, e. g., one speaker jumping into the talk of another. The work by Ward and Tsukahara [5] suggests that stretches of low pitch can trigger back-channel feedback from listener (*yeah, uh-huh, right*). Shriberg *et al.* [6] showed that speakers raise their voices when starting their utterance during somebody else’s talk, compared to starting in silence. Somewhat related work was presented in [7], where a specific feature based on pitch prediction was used for speaker count labeling, but experiments were performed on artificially overlapped speech.

In [8] the authors presented a system, which exploits cross-correlation-based spatial features for overlapping speech detection in a multi-microphone environment. In this paper, we shift our focus back to single distant microphone scenario and propose the use of several long-term prosody-based features for the detection of overlapping speech. We believe that they may act complementary to the short-term spectral features. The set of the most appropriate prosodic features is determined from the candidate set in a two-step process. In the first step, the features are sorted according to minimal-redundancy-maximal-relevance (mRMR) criterion and then, in the second step, a standard wrapper selection method is applied.

Our speaker diarization system assigns multiple speaker labels for the obtained overlap regions in order to decrease the missed speech error. Overlap segments can also be used to indicate data which should not be used for model building with the aim of a purer clustering. The experiments were conducted on the AMI Meeting corpus.

This paper is organized as follows. Overlap detection system is described in Section 2. The candidate prosodic features and feature selection process is discussed in Section 3. Speaker diarization system and its improvements are briefly outlined in Section 4. Experimental results and conclusions are given in Section 5 and 6, respectively.

2. Overlapping speech detection

The baseline overlap detection system, which was presented in [8], relies on a number of spectral-based features. First parameter kind is the cepstrum, thus 12 MFCCs were extracted every 10 ms over a window of 30 ms. Next, assuming that linear predictive coding (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], we computed the residual energy of a 12th-order LPC (LPCRE) over a 25 ms window. Residual energy, which corresponds to the prediction error, should be higher in overlapping speaker situations. Another 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 also information about the number of speakers speaking. This set of spectral parameters is extended with their first order derivatives and all features were mean-variance normalized according to statistics obtained from training data.

The system considers three acoustic classes representing non-speech, single-speaker speech and overlapping speech. For each class an HMM is defined. For a more accurate modeling of transitions between classes the HMM has three states, which also works as a minimum duration constraint. Every

This work has been funded by the Spanish project SARAI (TEC2010-21040-C02-01).

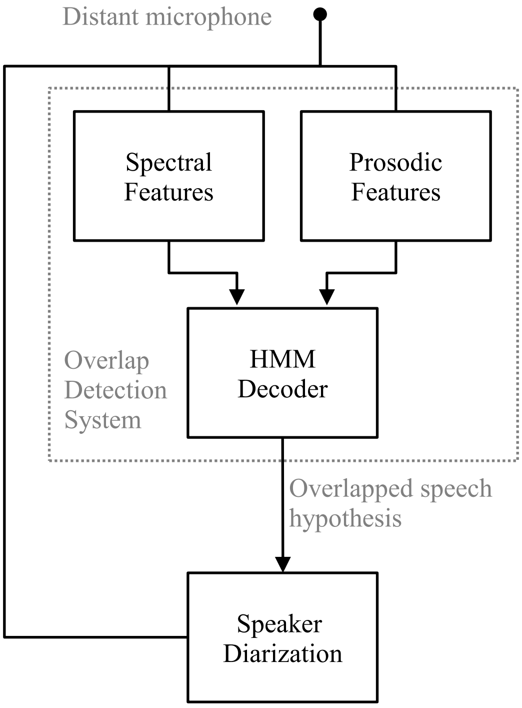

Figure 1: *Overlap detection system block diagram*

state is modeled with a GMM using diagonal covariance. Since the amount of training data is not balanced among classes, we use 256 Gaussian components for single-speaker speech and 64 components for overlapping speech and non-speech. GMMs are created by iterative Gaussian-splitting technique and subsequent re-estimation. A diagram of the overlap detection system with link to speaker diarization is given in Figure 1.

Detection hypothesis is obtained by Viterbi (maximum-likelihood) decoding and applying a word network. The transition probabilities between different HMMs are not trained. They 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 certain transitions are completely forbidden.

Overlap detection performance is measured with Recall—ratio between true detected and reference overlap time, Precision—ratio between true and all detected overlap time, and with Error—the sum of missed and false overlap time divided by reference overlap time. Results depend very much on the value of the OIP, which controls the amount of overlaps the system will hypothesize. It can be perceived as a compensation for an undertrained model. Initially, four values of OIP were selected based on results on development data, accounting for hypotheses with the highest recall (OIP = 0, no penalization), the highest F-ratio (OIP = -10), the low detection error rate (OIP = -50) and an acceptably high precision (OIP = -100).

3. Prosodic features and feature selection

The prosodic features that we are computing can be assigned to following categories: pitch, intensity and (four) formant frequencies. For each of these feature categories we estimate besides the actual value for every given time point also long-term statistical characteristics such as mean, median, minimum, maximum, standard deviation and the difference between the min and max value. Long-term statistics are extracted from 500 ms windows with 10 ms step for synchronization reasons with spectral features. Prosodic features were extracted with the help of Praat [1].

The feature selection process can be divided into two stages. In the first, we applied a mRMR algorithm [11] on held-out development data to score individually the candidate features against the target class (overlapping speech vs. single-speaker speech) and sorted them according to their minimum redundancy and maximal relevance. The ordered first 25 out of total 42 candidate features are given in Table 1.

Table 1: *Candidate prosodic features sorted according to the mRMR criterion, f0—pitch, int—intensity, f1-4—formants*

1.	f0_max	10.	f3_max	19.	f0_std
2.	f4_max	11.	int_diff	20.	int_min
3.	f4	12.	f3_min	21.	f4_std
4.	f0_min	13.	f0	22.	f1
5.	int	14.	f2	23.	f2_max
6.	f2_min	15.	f2_std	24.	int_std
7.	f4_min	16.	f0_med	25.	f3_med
8.	f1_min	17.	f4_med		
9.	f2_med	18.	f1_max		

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. The overlap detection performance for the baseline spectral system and five prosodic subsets are given in Figure 2. It can be seen that the systems with prosodic features achieve lower error especially for low penalization values when compared to the spectral-only system.

Unfortunately, it is not clear from the graphic what number of prosodic features is the optimal value. In order to solve this problem, we suggest to calculate the area under the curves in Figure 2 and use it as a decision factor. The amount of area reflects the overlap detection error of a particular system. For a fair comparison, every curve is extended with the same fictional starting point (Recall = 100%, False Alarm = 100%) and ending point (Recall = 0%, False Alarm = 0%). The values of this "overlap detection error" area for different number of prosodic features are given in Figure 3. Based on these results it was determined to select the first 20 prosodic features from Table 1.

The fusion strategy is very similar to the one in [8]. The emission probabilities are weighted by 0.9 and 0.1 for spectral and prosodic feature stream, respectively. These weights were defined in a similar way as the optimal number of prosodic features on the development data.

4. 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

[1] Praat: doing phonetics by computer [Computer program]. Version 5.2.04, retrieved from http://www.praat.org/

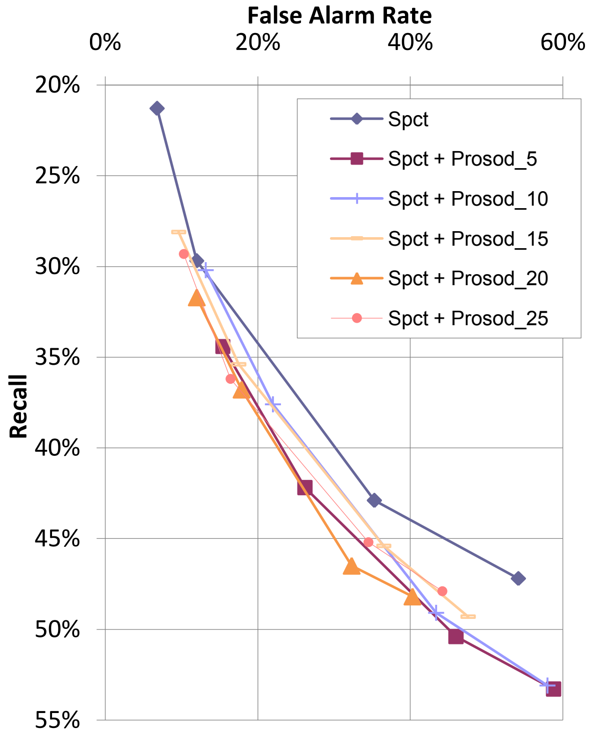

Figure 2: *Overlap detection performance on development data for spectral features (*Spct*) and combinations of spectral and various number of prosodic features (*Spct + Prosod 5–25*). Performance is measured at four OIP values (0,-10,-50,-100).*

driven by Bayesian information criterion (BIC). The system operates with 20 MFCCs extracted from 30 ms frames.

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 [13].

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.

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 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.

In order to evaluate just the impact of overlapping 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.

5. Experiments

5.1. Database and experimental setup

The experiments were conducted on the AMI Meeting corpus, which consists of 100 hours of meeting recordings. We were working with far-field microphone array channels sampled at 16 kHz. We used recordings from the Idiap site and divided them into training set (22 recordings), development set (3 recordings) and evaluation set (11 recordings). The average amount of overlapping speech was 14.40%. Training and evaluation of the overlap detection system were performed with forced-alignment annotations obtained by the SRI's DECIPHER recognizer. We did not apply any forgiveness collar around segment boundaries in scoring to make sure that overlap segments are considered, because the median overlap duration in this corpus is rather short (0.46 s).

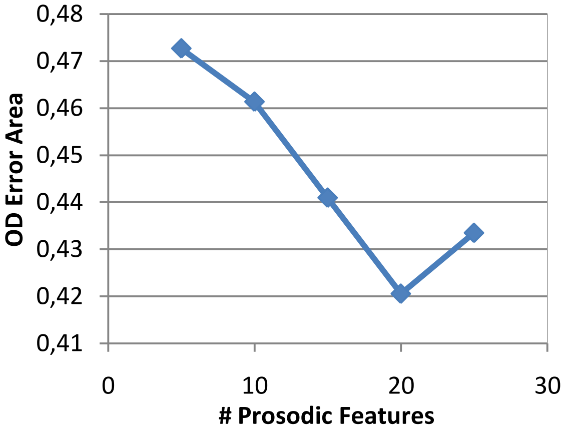

Figure 3: *The amounts of area under the ROC-like curves of Figure 2 for different number of selected prosodic features. The area value reflects overall overlap detection error.*

5.2. Overlap detection results

The comparison of overlap detection performance in terms of recall, precision and detection error for the pure spectral and combined—spectral and prosodic—system on evaluation data is given in Figure 4. The combined features outperform the spectral in terms of error for all OIPs with the lowest value of 75% at OIP -50. On the other hand, the situation is not so unequivocal with precision. The precision does not rise so steeply with increasing OIP in the new system. From our experience, this behavior could be possibly related with the higher amount of model parameters which need to be trained in the combined system.

5.3. Speaker diarization results

Based on previous results on development data, we use the overlap hypothesis at OIP -100 for overlap labeling in the diarization system and OIP 0 hypothesis for overlap exclusion. Table 2 shows the DER improvements of baseline diarization system when handling overlap detected either with the spectral overlap detection system, or with the combined system. The difference is not dramatic, but still, it can be seen a slight increase of improvement when overlap segments are detected also with prosodic features.

The results of a similar set of experiments where the baseline diarization was improved with both beamforming and TDOA feature stream are in Table 3. Here, as expected, the absolute DER values are better, compared with Table 2. The relative DER improvements by overlap labeling are higher for both feature sets, because the clustering is improved and the second label assignment is consequently more precise. Interesting is

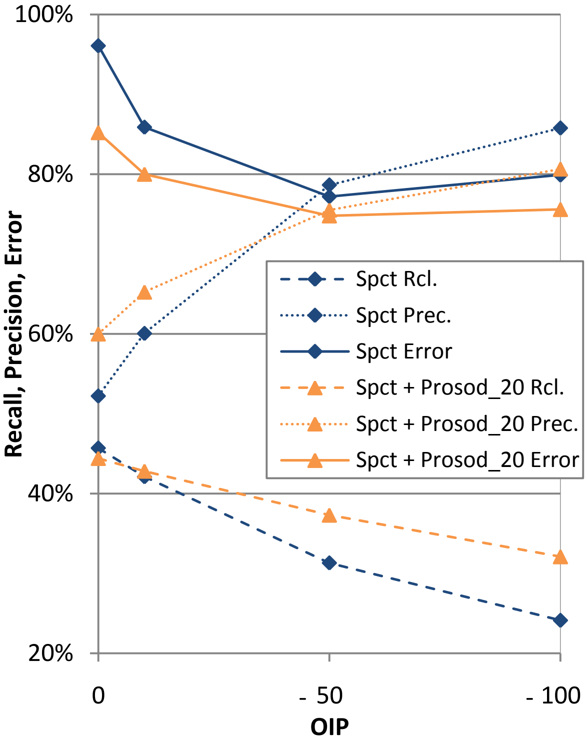

Figure 4: *Overlap detection performance for evaluation data using spectral features only (*Spct*), and the combination of spectral and 20 prosodic features (*Spct + Prosod 20*) in terms of detection error (solid lines), precision (dotted line) and recall (dashed line).*

Table 2: *Comparison of using overlapping speech detected with spectral (*Spct*) or combined spectral-prosodic system (*Spct+Prosod_20*) for labeling and exclusion in speaker diarization. DER and rel. improvements over the baseline (in %)*

Baseline		**38.3**
Overlap. det.:	+Labeling	+Labl. +Excl.
Spct	**36.5 / +4.7**	**35.6 / +6.9**
Spct+Prosod_20	**36.2 / +5.5**	**35.5 / +7.2**

that overlap exclusion did not lead to further improvement of the DERs in this case. There is probably some sort of improvement redundancy between the overlap exclusion technique and beamforming with TDOAs in speaker diarization.

6. Conclusions

We have proposed the use of prosodic features for the detection of simultaneous speech on distant channel data. Final subset from all candidate features was selected according to mRMR criterion and successive hill-climbing wrapper selection method. The obtained results after fusing short-term spectral and long-term prosodic features indicate that prosody conveys some complementary information for the detection of speaker overlap. Handling detected overlap segments in speaker diarization so that a second speaker label is assigned did improve both diarization baseline and diarization extended with beamforming and TDOA feature stream.

Table 3: *Speaker diarization improved with beamforming and TDOAs with labeling and exclusion of overlapping speech. Comparison of using overlaps detected with spectral (*Spct*) or combined spectral-prosodic (*Spct+Prosod_20*) system. DER and rel. improvements over the new baseline (in %)*

Baseline + Beam. + TDOAs		**35.7**
Overlap. det.:	+Labeling	+Labl. +Excl.
Spct	**33.8 / +5.3**	**34.0 / +4.9**
Spct+Prosod_20	**33.4 / +6.5**	**33.9 / +5.0**

7. References

[1] E. Shriberg, “Spontaneous speech: How people really talk and why engineers should care,” in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, Lisbon, Portugal, 2005, pp. 1781–1784.

[2] N. Morgan *et al.*, “The Meeting Project at ICSI,” in *Proc. 1st International Conference on Human Language Technology Research*, San Diego, USA, 2001, pp. 1–7.

[3] G. Friedland and O. Vinyals and Y. Huang and C. M´uller, “Prosodic and other Long-Term Features for Speaker Diarization,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 985–993, Jul. 2009.

[4] J. Žibert and F. Mihelič, “Fusion of Acoustic and Prosodic Features for Speaker Clustering,” *Lecture Notes in Computer Science*, vol. 5729/2009, pp. 210–217, 2009.

[5] N. Ward and W. Tsukahara, “Prosodic features which cue backchannel responses in English and Japanese,” *Journal of Pragmatics*, vol. 32/2000, pp. 1177–1207, 2000.

[6] E. Shriberg, A. Stolcke, and D. Baron, “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, USA, 2001, pp. 13–16.

[7] M. A. Lewis and R. P. Ramachandran, “Cochannel speaker count labelling based on the use of cepstral and pitch prediction derived features,” *Pattern Recognition*, vol. 34, no. 2, pp. 499–507, Feb. 2001.

[8] M. Zelenák, C. Segura, and J. Hernando, “Overlap detection for speaker diarization by fusing spectral and spatial features,” in *Proc. Interspeech ’10*, Makuhari, Japan, 2010, pp. 2302–2305.

[9] N. Sundaram, R. Yantorno, B. Smolenski, and A. Iyer, “Usable speech detection using linear predictive analysis - a model based approach,” in *Proceedings of ISPACS*, Awaji Island, Japan, 2003, pp. 231–235.

[10] R. Yantorno, “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] H. Peng, F. Long, and C. Ding, “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[12] J. Luque, X. Anguera, A. Temko, and J. Hernando, “Speaker diarization for conference room: The UPC RT07s evaluation system,” *Multimodal Technologies for Perception of Humans*, vol. 4625/2008, pp. 543–553, 2008.

[13] X. Anguera, C. Wooters, and J. Hernando, “Acoustic beamforming for speaker diarization of meetings,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.