IMPACT OF OVERLAPPING SPEECH DETECTION ON SPEAKER DIARIZATION FOR BROADCAST NEWS AND DEBATES

Delphine Charlet[1]*, Claude Barras*[2] *and Jean-Sylvain Liénard*[2]

[1]Orange Labs, France Telecom, Lannion, France
[2]LIMSI-CNRS, Univ. Paris-Sud, 91403, Orsay, France
[1]delphine.charlet@orange.com, [2]barras@limsi.fr, jslienard@gmail.com

ABSTRACT

The overlapping speech detection systems developped by Orange and LIMSI for the ETAPE evaluation campaign on French broadcast news and debates are described. Using either cepstral features or a multi-pitch analysis, a F1-measure for overlapping speech detection up to 59.2% is reported on the TV data of the ETAPE evaluation set, where 6.7% of the speech was measured as overlapping, ranging from 1.2% in the news to 10.7% in the debates. Overlapping speech segments were excluded during the speaker diarization stage, and these segments were further labelled with the two nearest speaker labels, taking into account the temporal distance. We describe the effects of this strategy for various overlapping speech systems and we show that it improves the diarization error rate in all situations and up to 26.1% relative in our best configuration.

Index Terms— speaker diarization, overlapping speech

1. INTRODUCTION

Automatic speech recognition and speaker diarization on broadcast data long focused on contents where speech overlaps were rare, or excluded speech overlap segments from their evaluation. On the other hand, studies on more spontaneous speech from multi-party conversations, especially telephone conversations and meetings report that 6 to 14% of words are overlapped [1, 2], and overlapping speech was identified as a major cause of error for speaker diarization [3]. On broadcast data, the assumption that speech overlap is negligible is no longer valid, when it comes to deal with political interviews [4] or talk-shows [5].

To the best of our knowledge, all the published studies about overlapped speech in speaker diarization focused on meeting data or on telephone conversations. This work focuses on overlapping speech detection in French TV broadcasts and its impact on speaker diarization in the context of the ETAPE evaluation campaign[1]. The next sections describes the ETAPE evaluation and its data, the proposed overlapping speech detection approaches and their performance, their integration for speaker diarization and concludes with a comparison with existing work.

Experiments were performed in the context of the French ETAPE evaluation campaign. LIMSI work was partly realized as part of the Quaero Program and the QCompere project, respectively funded by OSEO (French State agency for innovation) and ANR (French national research agency). Orange work was partly realized as part of the PERCOL project funded by ANR.

[1]http://www.afcp-parole.org/etape.html

2. ETAPE CAMPAIGN AND DATA

The ETAPE evaluation campaign took place in Spring 2012 [5]. It evaluated segmentation, transcription and information extraction in the audio channel of French TV and radio broadcasts. This work addresses the segmentation tasks, i.e. the multiple speaker detection and speaker turn segmentation.

The multiple speaker detection task (SES-2) aims at detecting the start and end times of segments containing speech from more than one speaker. Manual annotation of overlaps along with the reference transcription was provided by ELDA, and the temporal extent of the overlaps was refined through a forced alignment between the reference and an automatic transcription[2]. Due to the exploratory nature of the task, several metrics were proposed but no official metric was chosen. We report recall and precision of multiple speech detection expressed in duration, with non-speech regions excluded from the scoring, and the resulting F1-measure. Moreover, we restrict to the TV data since the forced alignment was not performed for the radio training data subset.

Performance in the speaker diarization task (SRL) is measured by the Diarization Error Rate (DER) as the sum of Miss Detection Rate, False Alarm Rate and Speaker Error Rate, where Speaker Error Rate is obtained after optimal mapping between automatic clusters and reference speakers[3]. Usually, evaluation of speaker diarization is performed after excluding regions of overlapped speech. Here, we evaluate the system including overlapped speech regions. If an overlapped speech region is assigned to only one speaker, this region is considered as Missed speech for the second speaker. If an non-overlapped speech region is falsely detected as overlapped speech region and assigned to 2 speakers, this region will be counted as false alarm speech for the second speaker.

The ETAPE TV subset consists in 29 hours of data (18 hours training, 5.5 hours development and 5.5 hours test) from three French TV channels (LCP, BFM and TV8) with news (BFM Story, LCP Top Questions), debates (LCP Pile et Face, LCP Ça vous regarde, LCP Entre les lignes) and reportages from a local TV with unprofesionnal speakers (TV8 La place du village). For the whole dataset, the ratio of overlaps amounts to 5.9%, ranging from 1.8% in the news to 3% in the reportage and 8.7% in the debates (cf. Table 1). The mean duration of overlap segments in the training set is 1.07 sec. and their median duration is 0.72 sec (see Figure 1 for the normalized histogram of overlap durations). The cumulated duration of overlaps on the same Figure shows that overlaps longer

[2]Thanks to Olivier Galibert from LNE for providing these alignments to the ETAPE participants.

[3]DER reported in this paper were computed using the conventionnal NIST evaluation tools with the default value of collar of 250ms

than 1 sec. cover about 70% of the cumulated duration of overlapping speech, even if they represent only 35% of the occurences. The corpus is biased towards male speakers, with a lot of journalists and politicians: there are 160 male vs. 43 female speakers in the training set, accounting for about 90% of the total duration, and overlaps almost exclusively involve male speakers.

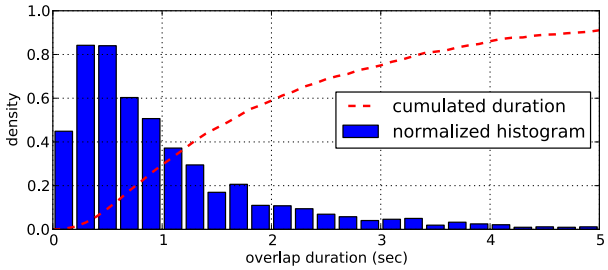

Fig. 1. Distribution of overlap durations in the training set

3. OVERLAPPING SPEECH DETECTION

This section describes the overlapping speech detection systems developed by LIMSI and Orange for the ETAPE evaluation campaign and their performance in the SES2 task.

3.1. LIMSI system combining cepstral and multi-pich features

For the LIMSI system, three GMM $\{\lambda_i\}_{i=0...2}$ with 256 Gaussians, respectively for non-speech, non-overlapping speech and overlapping speech, are trained using forced alignement between automatic and reference transcriptions of the ETAPE training data provided by LNE. Cepstral features (12 PLP and log-energy along with their first- and second-order derivatives) are computed over a windows of 30 ms with a 10 ms step. The frame-level likelihood-ratio (LLR) between the multiple-speaker model $f(x|\lambda_2)$ and the two other hypothesis is smoothed over a Hamming window H and the resulting value l_t compared to a decision threshold optimized on the development set:

$$l_t = \sum_{j=-d/2}^{d/2-1} H(j).\log \frac{f(x_{t+j}|\lambda_2)}{f(x_{t+j}|\lambda_0) + f(x_{t+j}|\lambda_1)}$$

In our developments, this approach was found to perform better than a Viterbi decoding with the $\{\lambda_i\}$ models as proposed in [6] (either with a minimal duration or a transition penalty).

Given the harmonic nature of voiced speech, it can be expected that approaches and features which are relevant for speech separation and multi-pitch detection [7] are also of interest for overlapping speech detection. We performed our experiments with the PSH algorithm designed by Liénard et al. [8]. It is based on a frequential approach and uses several spectral combs. Spectral combs are used as pattern matching tools for detecting the harmonic structures of voiced segments of speech. The dot product between a comb and the amplitude spectrum produces a pitch function exhibiting local maxima at frequencies where F0 is most probable. But in the multipitch cases, numerous spurious peaks appear. To strongly attenuate them, two families of combs are used: negative teeth comb and missing teeth comb which treat selectively harmonics errors and subharmonics errors. The algorithm performs a frame-to-frame analysis over a 50 ms window with 10 ms step without any post-processing, thus remaining computationnaly light compared to other multi-pitch estimators. Let $p_t \in \{0, 1, 2\}$ be the number of hypothesized F_0 output by the multi-pitch detector for the frame x_t. This value is then smoothed through a Hamming window H of size d, resulting in the frame-level harmonic feature h_t:

$$h_t = \sum_{j=-d/2}^{d/2-1} H(j).p_{t+j}$$

Finally, a frame-level linear combination of the l_t and h_t values was submitted as LIMSI primary system to the ETAPE SES2 task, with combination weights and decision thresholds optimized on the development set.

3.2. Cepstral Features by Orange

Three GMMs $\{\lambda_i\}_{i=0...2}$ with 256 Gaussians, resp. for male non-overlapped speech, female non-overlapped speech and overlapped speech, were trained using forced-alignement between automatic and reference transcriptions of the ETAPE training data provided by LNE. Cepstral features (12 MFCC and log-energy with their first and second order derivatives) are computed over a windows of 32 ms with a 16 ms frame rate. A 2-class HMM (overlapped/non-overlapped (with male and female GMM models) is then built. Viterbi decoding, only applied after a first external speech/non-speech segmentation step, is associated with a minimal state duration (2s in non-overlapped speech and 0.5s in overlapped speech). Finally, a post-processing filtering discards all the detected overlapped-speech segments whose length is less than 1s. This was the system integrated to in the diarization process submitted to the ETAPE evaluation campaign.

Since the campaign, the post-processing filtering based on the length of the segments has been replaced with a filtering based on log-likelihood ratio value at the segment level. For a detected overlapped speech segment X of N frames $(x_1, .., x_N)$, whith the same definition as above, the confidence measure is:

$$S(X) = \frac{1}{log(N)} \sum_{t=1}^{N} log \frac{f(x_t|\lambda_2)}{f(x_t|\lambda_1)}$$

The confidence measure $S(X)$ is compared to a threshold to validate the detection or not. The length normalisation by $log(N)$ instead of the usual N is meant to favor the long detections. Indeed, we have observed that the long detections of overlapped speech were more likely to be correct than the short ones.

3.3. Development and evaluation results

For LIMSI system, the performance of the multi-pitch system was significantly worse than the cepstral system on the development set (F1 measure of 45.3% vs. 54.5%), however it only relies on a frame-level ternary feature compared to the 13 real-valued features used in the cepstral system. The optimal size for the Hamming smoothing window was found to be slightly different (2.5 sec. for l_t vs. 2 sec. for h_t). The combination of both systems further improved the F1 performance to 55.8% on the development set and was chosen for the LIMSI primary submission to the ETAPE SES2 task.

Table 2 presents the results on the evaluation TV subset. The segment length filtering dramatically improves F1 value of Orange cepstral system from 43.3% to 55.2% mainly due to an increase in precision, and the alternative LLR filtering further improves it to 59.8%. LIMSI system presents a F1-measure slightly lower at 58.2%

Show type	Train	Development	Evaluation	All
News	5.6 / 297.9 (1.9 %)	1.1 / 67.6 (1.7 %)	0.8 / 66.4 (1.2 %)	7.6 / 431.9 (1.8 %)
Debates	41.5 / 486.3 (8.5 %)	9.6 / 128.7 (7.5 %)	13.6 / 130.1 (10.4 %)	64.7 / 745.1 (8.7 %)
Reportage	-	0.7 / 37.6 (1.8 %)	1.7 / 41.9 (4.0 %)	2.4 / 79.4 (3.0 %)
Total	47.1 / 784.2 (6.0 %)	11.4 / 233.9 (4.9 %)	16.0 / 238.3 (6.7 %)	74.6 / 1256.4 (5.9 %)

Table 1. Duration (in minutes) and ratio of overlapping speech relative to total speech in train, development and evaluation subsets, depending on the genre of the show

with a lower recall (52.7% vs. 64.3%) but a better precision than the best Orange system (64.9% vs. 56.0%). As could be expected, longer overlaps are easier to detect, and this behaviour is illustrated for the best performing O_2 system on Figure 2 where the overlap segments in the reference which are shorter than a minimal duration are ignored for the scoring.

System	P	R	F1
Orange cepstral	29.9	78.5	43.3
Orange cepstral+length filtering (O_1)	45.5	70.2	55.2
Orange cepstral+LLR filtering (O_2)	56.0	64.3	59.8
LIMSI system (L_1)	64.9	52.7	58.2

Table 2. Precision (P), recall (R) and F1-measure of overlap detection on the ETAPE TV evaluation subset for the different systems (official submissions to the ETAPE evaluation in bold).

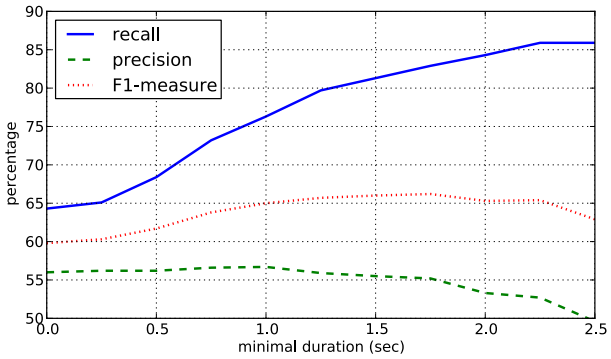

Fig. 2. Performance of overlap detection as a function of minimal segment length in the reference for O_2 system.

4. DIARIZATION EXPERIMENTS

4.1. Principle

In previous studies about speaker diarization in meetings [3], detailed error analysis showed that overlapped speech was a major cause of error, and [9] proposed two methods for handling overlap in speaker diarization:

- overlap exclusion: exclude overlapped speech segments from the diarization process;
- overlap labeling: label the overlapped speech segments with the speaker labels of the 2 nearest speakers (in time).

On oracle experiments, with perfect overlap detection, the overlap labeling strategies with the 2 nearest speakers proved to be very effective, and closed to a perfect (oracle) labeling strategies. This overlap handling scheme has been commonly adopted in the studies about the impact of overlapped speech in speaker diarization (e.g. [10]). Here, we perform overlap exclusion and propose a slightly modified labeling strategy, to cope with the errors of overlap detection:

- overlap labeling: always label the segment with the nearest speaker (in time), and label with the second nearest speaker only if its temporal distance to the segment is below a given threshold T_s.

This variant is meant to cope with error of false detection of overlapped speech in the middle of a speaker turn.

The diarization system used in these experiments is the one developped by Orange based on the principles of [11]: the first step consists in building an agglomerative clustering of speech segments based on Bayesian Information Criterion (where each cluster is modeled by a single Gaussian with a full covariance matrix). When each cluster contains enough data to model the voice more precisely, the clusters are modeled with Gaussians mixture, and the agglomerative clustering is pursued with a distance between clusters based on a cross-likelihood criterion. At each iteration of the clustering based on cross-likelihood, a Viterbi decoding is also performed to resegment the speech data into speaker turns, given the new clusters.

4.2. Evaluations

In the Figure 3, we plot the diarization error rate (DER) obtained with different overlapping speech detection systems, when including overlapping speech in the evaluation, and as a function of the threshold T_s. For comparison, the baseline system processes the documents without detection of overlapped speech. The other systems apply the overlap exclusion step, and the proposed new labeling strategy, using one of the proposed overlapping speech detection systems O_1, O_2 or L_1. The system with O_1 was the one submitted by Orange to the SRL task of the ETAPE challenge. Finally, oracle experiments (i.e. automatic speaker diarization with a perfect overlapping speech detection) are also reported. $T_s = 0$ corresponds to the performances obtained when a second speaker label is never attributed to the detected overlapping speech segments. On the contrary, for $T_s = \infty$ a second speaker label is always attributed to these segments.

First, we can observe that, whichever overlapping speech detection system is used and for any T_s, it always outperforms the baseline system without overlapping speech detection. The performances obtained with $T_s = 0$ (only one speaker is assigned to the overlapping speech segment) are always far better than the baseline system. This improvement is due to the purification of the clusters, which are only fed with detected non-overlapping speech. It can be seen that the best automatic system reaches the same performance as the oracle system, when only purification is performed. Thus, even though the performance of the overlapping speech detector is average (F1=59.8%), it is good enough for the exclusion

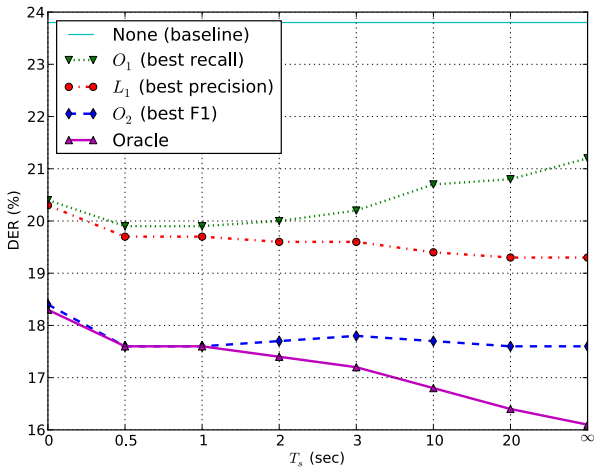

Fig. 3. Diarization error rate (DER) as a function of the threshold T_s controlling the attribution of the second speaker.

Overlaps detector	M.D.	F.A.	SER	DER
None (baseline)	6.1	0.7	17.0	23.8
O_1 (best recall)	1.6	4.9	14.6	21.2
L_1 (best precision)	3.1	2.0	14.2	19.3
O_2 (best F1)	2.0	3.0	12.6	17.6
Oracle	2.1	1.3	12.6	16.0

Table 3. Decomposition of the Diarization Error Rate into false alarm (F.A), missed detection (M.D.) and speaker error rate (SER).

step, to get all the benefits of the purified clusters. Then, the impact of the overlapping speech detections on the labeling strategies can be seen on the rest of the curves. For the system with high recall and low precision on overlapped speech detection, it is better not to always assign a second speaker: indeed, the increase of errors due to false alarm of overlapped speech is bigger than the reduction of errors due to the attribution of segment of overlapped speech. For the other systems with higher precision on overlapped speech detection, always attributing a second speaker appears to be a valid strategy. Table 3 shows the detailed components of the DER for the "always 2-nearest speakers" strategy, with the different overlapping speech detectors. The speaker error obtained with O_2 is the same as the one obtained with the oracle detector, and the main differences lies in the false alarm rate on speaker, which is bigger because of the false alarm of overlapping speech detection. On the other hand, the LIMSI approach has a higher precision than the other ones and leads to a smaller fale alarm rate, but does not fully benefits from the clusters purification, thus leading to a higher speaker error rate.

The DER per type of shows (news, debates or reportage) are presented in Table 4, for the baseline system, and for the best system O_2 (along with the relative improvement rate), and the overlapped speech ratio per type of shows. The more overlapping speech there is in the data, the better the improvement due to overlapping speech handling is. For news shows with very little amount of overlapped speech, the imprecision due to the overlapped speech detector degrades the overall results, while for debates shows, the decrease of DER reaches 33.2%.

Type of show	Overlapping speech (%)	Baseline DER	DER with O_2 detector
News	1.2	11.9	12.6 (+5.6%)
Debates	10.4	24.7	16.5 (-33.2%)
Reportage	4.0	41.6	29.8 (-28.4%)
All	6.7	23.8	17.6 (-26.1%)

Table 4. Relative DER improvement per type of shows for the best diarization system integrating O_2 vs. the baseline system.

5. RELATION TO PRIOR WORK

Many studies have been published on overlapping speech detection for speaker diarization of meetings. Some of them perform source separation or source localization relying on multiple channel recordings [12, 13, 14, 15] which are not available for broadcast data. [16] tested various features for overlapping speech detection in a HMM-based segmenter. On far-field recordings of the AMI meeting corpus with 18% of overlapped speech, they get 38% F-score in overlaps detection. Features such as silence distribution [10] or prosodic features [17] also gives a F-score on overlap detection around 40% in meetings.[18] proposed a convolutive non-negative sparse coding approach to speech overlap detection ; they get a 16.1% recall and 28.6% precision of overlapping speech detection on NIST RT meetings. On telephone converstations,[19] used entropy features estimated in the time domain for detecting overlapping speech, but their approach is only suitable in a two-speakers situation. Finally, relevant research for overlapping speech detection is also developed in the context of single-channel speech separation [20, 21].

When it comes to the integration of overlapping speech detection in speaker diarization system, the classical approach consists in applying exclusion and labeling, when labeling is either performed with speaker posterior probabilities [6, 16, 22] or 2-nearest speaker labeling [9]. Relative improvement of DER such as 4.2% [23], 6.5% [18], 7.2% [17], 12.4% [22] and 18.7% [10] have been reported for meetings data, when the major part of the improvement is due to the exclusion step.

Thus, our work on broadcast data give consistent results with prior studies on meetings data. But the results obtained on broadcast are significantly better than those reported on meetings, either for F-mesure on overlap detection or for relative DER improvement.

6. CONCLUSIONS

In this paper, we have studied the impact of overlapping speech detection in speaker diarization for broadcast news and debates. Whereas many studies have been done in the context of meetings diarization, this is the first time that this question is treated and evaluated in the broadcast context. The basic strategy of overlap handling proposed in [9] has been applied, with different overlapped speech detectors. The influence of the second-speaker labeling step has been studied with a modified labeling strategy. The experiments were conducted on a corpus of 5.5 hours of 7 different TV shows with a varying level of overlapped speech, from the ETAPE evaluation campaign. Two overlapping speech detection systems were developed by Orange and LIMSI, relying on standard cepstral features or on a multi-pitch analysis. The best configuration presents a F1-measure of about 60%. The diarization experiments show that this level of performance is sufficient to provide all the benefits of the exclusion step due to the purification of the clusters, and enable also improvement at the labeling step. The DER decreases from 23.8% with no overlap handling to 17.6% with automatic overlap detection.

7. REFERENCES

[1] E. Shriberg, A. Stolcke, and D. Baron, "Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation," in *Proceedings of the 7th European Conference of Eurospeech*, Aalborg, September 2001, pp. 1359–1362.

[2] O. Cetin and E. Shriberg, "Errors in meetings: Effects before, during, and after the overlap," in *ICASSP 2006*, Toulouse, France, May 2006.

[3] M. Huijbregts and C. Wooters, "The blame game: Performance analysis of speaker diarization system components," in *proc. Interspeech*, Antwerp, Belgium, September 2007.

[4] Gilles Adda, Martine Adda-Decker, Claude Barras, Philippe Boula de Mareüil, Benoît Habert, and Patrick Paroubek, "Speech Overlap and Interplay with Disfluencies in Political Interviews," in *International workshop on Paralinguistic Speech - between models and data, ParaLing 2007*, Sarbrücken, August 2007, pp. 41–46.

[5] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, O. Galibert, et al., "The ETAPE corpus for the evaluation of speech-based TV content processing in the french language," in *International Conference on Language Resources, Evaluation and Corpora*, 2012.

[6] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4353–4356.

[7] A. De Cheveigné, "Multiple F0 estimation," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, DeLiang Wang and Guy J. Brown, Eds., pp. 65–70. Wiley/IEEE Press, 2006.

[8] J-S. Liénard, C. Barras, and F. Signol, "Using sets of combs to control pitch estimation errors," *Proc. of Meetings on Acoustics*, vol. 4, no. 1, 2008.

[9] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *proc. ASRU*, Kyoto, Japan, December 2007.

[10] S.H. Yella and F. Valente, "Speaker diarization of overlapping speech based on silence distribution in meetings recordings," in *proc. Interpseech*, Portland, USA, September 2012.

[11] C. Barras, X. Zhu, S. Meignier, and J-L. Gauvain, "Multi-stage speaker diarization of broadcast news," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1505–1512, 2006.

[12] Thilo Pfau, Daniel P.W. Ellis, and Andreas Stolcke, "Multispeaker Speech Activity Detection for the ICSI Meeting Recorder," in *Proceedings IEEE Automatic Speech Recognition and Understanding Workshop - ASRU*, Trento, Italy, December 2001.

[13] S. J. Wrigley, G. J. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multi-channel audio," *IEEE Trans. on Speech and Audio Processing*, vol. 13, pp. 84–91, 2005.

[14] Kornel Laskowski and Tanja Schultz, "Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings," in *ICASSP 2006*, Toulouse, France, May 2006, pp. 993–996.

[15] Jose Pardo, Xavier Anguera, and Chuck Wooter, "Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences," in *Interspeech 2006 - ICSL*, Pittsburgh, USA, September 2006, pp. 2194–2197.

[16] K. Boakye, O. Vinyals, and G. Friedland, "Two'sa crowd: Improving speaker diarization by automatically identifying and excluding overlapped speech," in *Proc. Interspeech 2008*, 2008, pp. 32–35.

[17] M. Zelenak and J. Hernando, "The detection of overlapping speech with prosodic features for speaker diarization," in *Proc. Interspeech 2011*, 2011, pp. 32–35.

[18] R. Vipperla, J. Geiger, S. Bozonnet, D. Wang, N. Evans, B. Schuller, and G. Rigoll, "Speech overlap detection and attribution using convolutive non-negative sparse coding," in *ICASSP-12*, 2012, pp. 4181–4184.

[19] O. Ben-Harush, H. Guterman, and I. Lapidot, "Frame level entropy based overlapped speech detection as a pre-processing stage for speaker diarization," in *Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop on*, 2009, pp. 1–6.

[20] P. Mowlaee, M. G Christensen, Z. H Tan, and S. H Jensen, "A MAP criterion for detecting the number of speakers at frame level in model-based single-channel speech separation," in *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*, 2010, pp. 538–541.

[21] R. Saeidi, P. Mowlaee, T. Kinnunen, Z. H Tan, M. G Christensen, S. H Jensen, and P. Fränti, "Improving monaural speaker identification by double-talk detection," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[22] K. Boakye, O. Vinyals, and G. Friedland, "Improved overlapped speech handling for speaker diarization," in *12th Annual Conference of the International Speech Communication Association*, Florence, Italy, 2011, pp. 941–944.

[23] M. Huijbregts, D. van Leeuwen, and F. de Jong, "Speech overlap detection in a two-pass speaker diarization system," in *Proc. Interpseech*, 2009.