Convolutive Non-Negative Sparse Coding and New Features for Speech Overlap Handling in Speaker Diarization

Jürgen T. Geiger[1]*, Ravichander Vipperla*[2]*, Simon Bozonnet*[2]
Nicholas Evans[2]*, Björn Schuller*[1] *and Gerhard Rigoll*[1]

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
[2]Multimedia Communications Department, EURECOM, Sophia Antipolis, France
{geiger,schuller,rigoll}@tum.de, {vipperla,bozonnet,evans}@eurecom.fr

Abstract

The effective handling of overlapping speech is at the limits of the current state of the art in speaker diarization. This paper presents our latest work in overlap detection. We report the combination of features derived through convolutive non-negative sparse coding and new energy, spectral and voicing-related features within a conventional HMM system. Overlap detection results are fully integrated into our top-down diarization system through the application of overlap exclusion and overlap labeling. Experiments on a subset of the AMI corpus show that the new system delivers significant reductions in missed speech and speaker error. Through overlap exclusion and labelling the overall diarization error rate is shown to improve by 6.4 % relative.

Index Terms: speech overlap detection, convolutive non-negative sparse coding, speaker diarization

1. Introduction

The detection and handling of overlapping speech is still a major challenge in speaker diarization [1] and, indeed, in any field of automatic language processing [2]. Speaker diarization systems aim to determine “who speaks when” but, while overlapping speech is typical in uncontrolled, spontaneous conversations, current state-of-the-art speaker diarization systems are generally capable of detecting only a single active speaker. Consequently, intervals with multiple active speakers can contribute directly to diarization error. In addition, overlapping speech can lead to speaker model impurities which indirectly contribute to diarization error through degraded clustering performance.

Several different systems have been proposed to handle overlapping speech in the context of speaker diarization. Huijbregts et al. [3] use speech data around speaker turns to create show-specific models of overlapping speech. With similar models of non-overlapping speech these are subsequently used to detect intervals of overlap and to identify or label contributing speakers. Boakye et al. [4] propose an HMM-based approach to overlap detection which uses models of overlapping and non-overlapping speech trained on external data, while Zelenak et al. [5] explored the use of prosodic features with a similar HMM-based classifier.

Our previous work reports the use of convolutive non-negative sparse coding (CNSC) for detecting overlapping speech [6]. CNSC is used to project intervals of mixed speech onto speaker-specific bases; base activations are used to detect overlap. This approach was extended and improved in [7] using enhanced features and optimized CNSC parameters. While the approach successfully combines the advantages of mixed pattern decomposition due to non-negative constraints and powerful representation and noise robustness due to sparse coding, overlap/non-overlap classification is unrealistically erratic since it lacks any form of duration modelling.

The contributions of this paper are three-fold: first, we report the use of CNSC base activations within an HMM framework, which inherently includes duration modelling. Second, we introduce new energy, spectral and voicing related features which are well-suited to overlap detection. Third, we describe the integration of overlap detection into a full speaker diarization system and demonstrate improved performance through overlap exclusion and labeling.

The remainder of this paper is structured as follows: the CNSC-based approach to overlap detection is described in Section 2; new energy, spectral and voicing related features are introduced in Section 3; overlap detection and labeling experiments are reported in Section 4; conclusions and perspectives are reported in Section 5.

2. Convolutive Non-Negative Sparse Coding for Overlap Detection

This section describes the general approach to CNSC base learning and overlap detection.

2.1. Convolutive Non-Negative Sparse Coding

Non-negative sparse coding (NSC) [8] is an approach to represent non-negative, multi-variate data as a linear combination of lower rank bases. Only additive combinations are allowed due to the imposition of non-negative constraints.

With NSC, a non-negative matrix $D \in \mathbb{R}^{\geq 0}_{M \times N}$ is represented as:

$$D \approx WH \qquad (1)$$

where $W \in \mathbb{R}^{\geq 0}_{M \times R}$ and $H \in \mathbb{R}^{\geq 0}_{R \times N}$ are the bases and base activations respectively. These are learned such that the regularised least square error between the original matrix D and the recomposition (WH) is minimised according to:

$$(\hat{W}, \hat{H}) = \arg \min_{W,H} \|D - WH\|_F^2 + \lambda \sum_{ij} H_{ij}, \qquad (2)$$

where λ is a regularization parameter which controls the sparsity of the resulting representation. Our work involves a con-

This research was supported by the BFHZ and by the ALIAS project (AAL-2009-2-049) co-funded by the EC, the French ANR and the German BMBF.

volutive variant, referred to as convolutive NSC (CNSC) [9], where the decomposition takes the form:

$$D \approx \sum_{p=0}^{P-1} W_p \overset{p\rightarrow}{H}, \tag{3}$$

where P is the convolution range. The column shift operators $\overset{p\rightarrow}{\cdot}$ and $\overset{p\leftarrow}{\cdot}$ shift p columns of H to the right and left respectively. The learning of bases and activations together according to Eq. (2) is a non-convex optimization problem and is solved by iterative update rules presented by other authors in [10].

2.2. CNSC-based Features

To compute CNSC features for overlap detection, bases W are learned for each speaker in an audio document using spectral magnitude features extracted from segments of preferably pure (non-overlapping) speech. The base patterns for each speaker are then concatenated together to create a global basis W^G which spans the spectral patterns of all speakers. Spectral magnitude features across the whole audio document are then decomposed at the frame level according to Eq. (2) with W^G kept fixed and only H being updated to minimise the optimisation criterion.

The activations H and bases W can be used to separate and reconstruct each speaker's activity and hence to detect segments of overlapping speech. Since, however, bases W are normalised and thus activations of W reflect speaker energy, we use the activations H on their own to detect overlap. The energy for speaker s during frame j is estimated according to:

$$E_j(s) = \sum_{i \in I_s} H_{ij} \tag{4}$$

where I_s represents the speaker-specific rows in H, or the activations for speaker s. Due to erratic overlap/non-overlap classifications, speaker activation energies calculated as per Eq. (4) are smoothed with a moving average filter and used to compute two frame-level features for overlap detection.

The first feature is the energy ratio ER and is estimated for frame j as follows:

$$ER_j = \frac{E_j(\hat{s}_2)}{E_j(\hat{s}_1)} \tag{5}$$

where $\hat{s}_i$ denotes the speaker with the i-th highest energy. The energy ratio reflects the difference in activation energy for the two speakers who are deemed to be most active in the given frame. For overlapping segments we expect the ratio to be nearer to unity while for non-overlapping segments the ratio should be nearer to zero. Since overlapping speech segments typically have more energy (they comprise speech from multiple speakers) we also estimate the total energy E_j by summing Eq. (4) across all speakers $s \in S$:

$$E_j = \sum_{s \in S} E_j(s) \tag{6}$$

To normalise the total energy across different recordings, the mean over all the speech frames in the respective recording is subtracted from E_j, resulting in the normalised total energy ET:

$$ET_j = E_j - \frac{f}{|J_{sp}|} \sum_{j \in J_{sp}} E_j, \tag{7}$$

where f is a regularization factor tuned on held-out development data and J_{sp} denotes all speech frames in the recording, determined by the speech activity detection (SAD) component of our diarization system. Whereas our previous work investigated the simple thresholding of the normalized energy ratio ER and total energy ET to detect overlap, this paper reports their use as additional features in an HMM-based overlap detection system.

Feature	**win. size**	**KL score**
Energy & spectral (27)		
MFCC 1-12	**60**	**0.01-0.06**
loudness (auditory model based)	**60**	**0.29**
zero crossing rate	25	0.04
energy in band 250 - 650 Hz	**25**	**0.98**
energy in band 1 kHz - 4 kHz	**25**	**1.15**
25 % spectral roll-off point	25	0.03
50 % spectral roll-off point	25	0.02
75 % spectral roll-off point	25	0.02
90 % spectral roll-off point	25	0.01
spectral flux	**25**	**0.43**
spectral entropy	25	0.02
spectral variance	25	0.00
spectral skewness	25	0.02
spectral kurtosis	**25**	**0.06**
psychoacoustic sharpness	25	0.00
spectral harmonicity	**25**	**0.09**
Voicing related (6)		
F_0 (subharmonic summation (SHS) followed by Viterbi smoothing)	60	0.03
probability of voicing	**60**	**0.18**
jitter	**60**	**0.08**
shimmer (local)	**60**	**0.11**
jitter (delta: "jitter of jitter")	60	0.02
logarithmic Harmonic-to-Noise Ratio (logHNR)	60	0.01
CNSC-based (2)		
energy ratio	**40**	**0.05**
CNSC energy	**40**	**0.28**

Table 1: Candidate features with window sizes and score of the KL divergence based feature selection on the training set. Selected features are indicated in bold.

3. Additional Features and Feature Selection

In addition to the two CNSC-based features described above we also consider new energy, spectral and voicing related features which are well-suited to overlap detection. They are a subset of the AVEC2011 audio feature set [11] and are extracted using the open-source openSMILE toolkit [12]. The resulting 35 candidate features, including CNSC and baseline MFCC features, are listed in Table 1. All features are computed every 20 ms with indicated window sizes.

We use a Kullback-Leibler (KL) divergence-based feature selection approach similar to that reported by Zhou *et al.* [13] to identify features most pertinent to overlap detection. The discriminant value of each feature f is computed according to:

$$d_f = D(p_f \; || \; q_f), \tag{8}$$

where $D(\cdot \; || \; \cdot)$ is the KL divergence, p_f is the distribution of feature f for overlap frames, and q_f is the distribution over all frames. The KL divergence $D(p \; || \; q)$ of two probability distributions p and q is computed as

$$D(p \,||\, q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} \, \mathrm{d}x. \tag{9}$$

Under the assumption of Gaussian distributed features with mean μ and variance σ^2, Eq. (9) can be computed as:

$$D(p \,||\, q) = \log \frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2}. \tag{10}$$

KL divergence scores for all features are also displayed in Table 1 and show that a small selection are particularly well-suited to overlap detection. Scores for loudness, the two spectral energy features, spectral flux, kurtosis, harmonicity, probability of voicing, jitter, shimmer and the two CNSC features (illustrated in boldface in Table 1) are all higher than those for MFCC features. The energy-related features give the highest scores, which is somewhat expected, since the signal energy should be a good indicator of overlap. Jitter is a measure of fluctuations in fundamental frequency while shimmer is a measure of amplitude variability and it is thus of no surprise that they are also good indicators of overlap. Accordingly, all of these features are used together with standard MFCCs as additional inputs to an HMM overlap classifier. The feature set is augmented with first order regression coefficients and is normalized, using the statistics of the training data only, to have zero mean and unity variance.

4. Experiments

We report an assessment of our new overlap detection system using a subset of the AMI meeting corpus.

4.1. HMM Overlap Detection System

Experiments were conducted using an HMM classifier similar to that reported in [4]. There are three models corresponding to non-speech, non-overlapping speech and overlapping speech. Each model has three states and observations are modeled with a multivariate Gaussian Mixture Model (GMM) with diagonal covariance matrices. Due to unbalanced training data each mixture in the speech model has 256 components, while those in both the nonspeech and overlap models have 64 components. Models are trained with an iterative mixture splitting technique with successive re-estimation. Transitions from non-speech to overlapping speech are forbidden, as are self-transitions, e.g. from overlapping speech to overlapping speech. The log-likelihood transition penalty from speech to overlapping speech (also referred to as the overlap insertion penalty OIP) is tuned to control the trade off in precision and recall performance.

4.2. Overlap Handling

Overlap handling is achieved with two setups which correspond to different OIPs applied during HMM decoding. First, detected intervals of overlapping speech are excluded from the diarization clustering process to reduce speaker model impurities. For this approach, a high overlap detection recall is desired in order to detect and discard as much overlapping speech as possible.

Second, overlap labeling is applied by labelling a second speaker in the diarization output for all intervals of detected overlapping speech. While this can reduce missed speaker time, it can also introduce false alarms and thus high precision is desirable. One of two approaches is used to determine the second speaker where either the GMM likelihoods (LLKs) or the CNSC energies according to Eq. (4) are summed up over the detected overlap segment. The speaker with the highest summed score (or the second highest, if the speaker with the highest score is already that detected by the baseline system) is then added as a second speaker.

Test set			
EN2003a	EN2009b	ES2008a	ES2015d
IN1008	IN1012	IS1002c	IS1003b
IS1008b	TS3009c		

Table 2: Meetings from the AMI evaluation dataset used for the tests

4.3. Experimental Setup

A selection of 40 meeting recordings is used for training whereas all evaluation work is conducted with the the same ten files used in previous work by other authors [14]. The list of meetings in the evaluation/test set is displayed in Table 2. In all cases we considered only the single-channel, far-field microphone recordings. On average the amount of overlapping speech is in the order of 20%.

CNSC bases are first learned for each speaker in the standard diarization system output (which is regarded as pure speech). The algorithm described in Section 2 is applied to magnitude spectra computed for 40 ms windows with a window shift of 20 ms. CNSC speaker activations are calculated with speaker bases of size $R = 35$, a convolutional range of $P = 4$ and a sparseness parameter of $\lambda = 0.05$. The factor f in Eq. (7) was tuned on held-out development data and set to $f = 1.2$.

Overlap detection performance is assessed using averaged, frame-level precision and recall statistics. In addition, we report the overlap detection error (E), which is defined as the sum of false alarm and missed overlap times divided by the reference overlap time. This measure is a good indicator for the possible improvement in DER through overlap handling. For overlap exclusion the OIP is set to zero, while for overlap labeling the OIP is set to -100. In the following we thus report overlap detection performance for each set of features using both OIP values. The speaker diarization system used for all experiments reported below is the top-down LIA-EURECOM system reported in [15]. Finally, so that all results are independent of speech activity detection, we used reference speech/nonspeech segmentations in all cases.

4.4. Overlap Detection Results

Table 3 shows overlap detection results for each of the different system setups. Results reported in [14] for an HMM-based system with MFCC and other features are illustrated in the first line for high recall (left) and high precision (right) setups. They are slightly better than those for our previous CNSC-based system [7]. The last three lines in Table 3 show results for our new system using only baseline MFCC features, the same system with additional AVEC2011 features and then with additional CNSC features. In all cases, for the high recall setup $OIP = 100$ whereas for the high precision setup $OIP = 0$. The use of AVEC2011 features leads to a substantial improvement in precision and error over our own MFCC baseline but a drop in recall. The inclusion of CNSC features brings further significant improvements to recall performance which is then comparable to previous work [14, 7] but with significantly better precision and also the lowest error.

System			P	R	E	P	R	E
MFCC [14]			0.55	0.40	-	0.64	0.24	-
CNSC [7]			0.55	0.31	-	0.64	0.23	-
MFCC	AVEC11	CNSC	OIP					
			0			-100		
✓	-	-	0.45	0.50	1.18	0.54	0.26	0.96
✓	✓	-	0.61	0.24	0.91	0.86	0.13	0.89
✓	✓	✓	0.66	0.31	0.85	0.82	0.23	0.82

Table 3: Overlap detection results on the test set, comparing previously published results with our HMM system with various features. For each system, two different precision (P) vs. recall (R) operating points with their respective overlap detection error (E) are shown, depending on the OIP.

System	Miss	FA	SpkE	DER	Imp.
Baseline [15]	15.0	0.0	18.2	33.2	
+Exclusion	15.0	0.0	17.7	32.7	+1.5
+Labeling LLK	11.6	0.6	20.1	32.3	+2.7
+Labeling CNSC	11.6	0.6	19.6	31.9	+4.0
+Exc. + Lab. LLK	11.6	0.6	19.4	31.6	+4.8
+Exc. + Lab. CNSC	11.6	0.6	18.9	31.1	+6.4

Table 4: Influence of overlap handling (applying either overlap exclusion or overlap labeling or both) for our test set, showing the missed speaker error (Miss), false alarm error (FA), speaker error (SpkE), diarization error rate (DER) and relative improvement in DER over the baseline. Overlap labeling is performed using either LLK scores or CNSC energy scores.

4.5. Diarization Results

The best setup using MFCCs, AVEC2011 and CNSC features was then used to integrate overlap handling into our full diarization system [15]. The high recall setup was used for overlap exclusion whereas the high precision setup was used for overlap labeling.

Results are presented in Table 4. For the baseline system overlapping speech is shown to contribute 15 % to missed speech whereas there are no false alarms due to the use of reference speech/non-speech transcriptions. With a speaker error of 18.2 % a baseline DER of 33.2 % falls marginally to 32.7 % (1.5 % relative improvement) when overlap exclusion is used to reduce clustering impurities. On its own (without exclusion) overlap labeling has a slightly larger impact on performance. The DER improves by 2.7 % relative when labelling is performed using LLK scores and by 4.0 % relative for CNSC scores. With small increase in false alarms, the average missed speech rate falls to 11.6 % whereas there is a small increase in speaker error due to erroneous labeling. When used in addition to overlap exclusion, LLK and CNSC based overlap labeling approaches give relative improvements of 4.8 % and 6.4 % respectively.

5. Conclusions

This paper reports our successful efforts to advance the state of the art in speech overlap handling for speaker diarization. It shows how CNSC and new energy, spectral and voicing related features can be coupled with MFCC features and integrated into an HMM-framework to improve diarization performance through overlap exclusion and labelling.

The two tasks require different operating points in overlap detection. Whereas overlap exclusion requires high recall, labeling requires high precision. Compared to our own MFCC baseline system energy, spectral and voicing related features bring improvements in precision whereas CNSC features bring improvements in recall. Since recall rates remain low and missed speaker rates remain high, there is still significant potential to improve speaker diarization performance. Future work should concentrate on improved recall performance and we believe that CNSC-based approaches warrant further attention.

6. References

[1] M. Huijbregts and C. Wooters, “The Blame Game: Performance Analysis of Speaker Diarization System Components,” in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 1857–1860.

[2] E. Shriberg, “Spontaneous Speech: How People Really Talk and Why Engineers Should Care,” in *Proc. Eurospeech*, Lisbon, Portugal, 2005, pp. 1781–1784.

[3] M. Huijbregts, D. Van Leeuwen, and F. De Jong, “Speech Overlap Detection in a Two-Pass Speaker Diarization System,” in *Proc. Interspeech*, Brighton, UK, 2009, pp. 1063–1066.

[4] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, “Overlapped Speech Detection for Improved Diarization in Multi-Party Meetings,” in *Proc. ICASSP*, Las Vegas, NV, USA, 2008, pp. 4353–4356.

[5] M. Zelenak and J. Hernando, “The Detection of Overlapping Speech with Prosodic Features for Speaker Diarization,” in *Proc. Interspeech*, Florence, Italy, 2011, pp. 1041–1044.

[6] 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 *Proc. ICASSP*, Kyoto, Japan, 2012, pp. 4181–4184.

[7] J. Geiger, R. Vipperla, N. Evans, B. Schuller, and G. Rigoll, “Speech Overlap Detection and Attribution Using Convolutive Non-Negative Sparse Coding: New Improvements and Insights,” in *Proc. EUSIPCO*, Bucharest, Romania, 2012.

[8] P. O. Hoyer, “Non-negative Matrix Factorization with Sparseness Constraints,” *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[9] P. Smaragdis, “Convolutive Speech Bases and Their Application to Supervised Speech Separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[10] W. Wang, “Convolutive Non-Negative Sparse Coding,” in *Proc. IEEE International Joint Conference on Neural Networks*, Hong Kong, China, 2008, pp. 3681–3684.

[11] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, “AVEC 2011–The First International Audio/Visual Emotion Challenge,” in *1st Int. Audio/Visual Emotion Challenge and Workshop*, Memphis, TN, USA, 2011, pp. 415–424.

[12] F. Eyben, M. Wöllmer, and B. Schuller, “openSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor,” in *Proc. ACM Multimedia (MM)*, Florence, Italy, 2010, pp. 1459–1462.

[13] X. Zhou, X. Zhuang, M. Liu, H. Tang, M. Hasegawa-Johnson, and T. Huang, “HMM-based Acoustic Event Detection with AdaBoost Feature Selection,” *Multimodal Technologies for Perception of Humans*, pp. 345–353, 2008.

[14] K. Boakye, O. Vinyals, and G. Friedland, “Improved Overlapped Speech Handling for Speaker Diarization,” in *Proc. Interspeech*, Florence, Italy, 2011, pp. 941–944.

[15] S. Bozonnet, N. Evans, and C. Fredouille, “The LIA-Eurecom RT09 Speaker Diarization System: Enhancements in Speaker Modelling and Cluster Purification,” in *Proc. ICASSP*, Dallas, TX, USA, 2010, pp. 4958–4961.