SPEECH OVERLAP DETECTION USING CONVOLUTIVE NON-NEGATIVE SPARSE CODING: NEW IMPROVEMENTS AND INSIGHTS

Jürgen T. Geiger[1], *Ravichander Vipperla*[2], *Nicholas Evans*[2], *Björn Schuller*[1], *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,evans}@eurecom.fr

ABSTRACT

This paper presents recent advances in the application of convolutive non-negative sparse coding (CNSC) to the problem of overlap detection in the context of conference meetings and speaker diarization. CNSC is used to project a mixed speaker signal onto separate speaker bases and hence to detect intervals of competing speech. We present new energy ratio and total energy features which give significant improvements over our previous work. The system is assessed using a subset of the AMI meeting corpus. We report results which are comparable to the state of the art which support the potential of a new approach to overlap detection. An analysis of system performance highlights the importance of further work to addresses weaknesses in detecting particularly short segments of overlapping speech.

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

1. INTRODUCTION

Overlapping speech is known to degrade the performance of speaker diarization systems [1]. Unfortuately its occurence is typical in uncontrolled, spontaneous scenarios such as that of conference meetings which have been the focus of the NIST Rich Transcription (RT) evaluations since 2004[1]. Accordingly there is an increasing effort within the community to develop new algorithms to detect and appropriately handle overlapping speech. New algorithms are needed, first to detect segments of overlap so that they can be removed from data used in clustering and modelling and, second so that segments of overlapping speech can be attributed to relevant speakers. Even if there is contradictory evidence that the removal of overlapping speech from data used in clustering and modelling gives any significant reduction in the diarization error rate (DER), detection is nonetheless a pre-cursor to attribution which can significantly improve performance [2, 3]. Overlap detection is an unsolved problem and the focus in this paper.

There is little prior work in the open literature. Boakye et al. [3, 4] investigated the use of various different features in a hidden Markov model (HMM) framework for overlap detection together with a post-processing step for attribution. They show significant improvements in the diarization error rate (DER) on a subset of the AMI corpus [5]. Huijbregts et al. [6] report a detection approach which uses a model of overlapping speech, trained on data localised around speaker turns. They show minor improvements in the DER on a more challenging NIST RT data.

Our own approach to overlap detection [7] is based on convolutive, non-negative matrix factorisation (CNMF) [8] with sparse coding constraints. The resulting convolutive non-negative sparse coding (CNSC) approach combines the advantages of mixed pattern decomposition due to non-negative constraints and powerful representation and noise robustness due to sparse coding. The acoustic signal is projected onto a set of speaker bases and the resulting base activations are used to detect overlapping speech. We achieved results comparable to the state-of-the-art systems in both overlap detection and attribution on RT corpora.

Recent work by Zelenak et al. [9, 10] reports an HMM system using spatial features/localisation and prosodic features in addition to conventional acoustic features. Significant improvements in precision and recall are reported. Our particular interest, however, involves a single distant microphone where no localisation features are available. The restriction to a single microphone makes the problem more challenging but solutions more versatile.

The performance of each approach described above is at best modest and further work is needed to improve overlap detection performance before attention can be turned toward the development of effective attribution algorithms. The work presented in this paper aims to identify the weaknesses in our own overlap detection system as a guide to future work. Since it contains a higher degree of spontaneous speech and more frequent intervals of overlap this work was carried out using AMI data. The contributions are two-fold. First, we report a

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.

[1] `http://www.itl.nist.gov/iad/mig/tests/rt/`

new measure to detect ovelapping segments using CNSC activations. Second, we report an analysis of overlap detection performance which highlights weaknesses in detecting particularly short, but significant segments of overlapping speech.

The paper is organised as follows. In Section 2 we describe the CNSC algorithm which is the basis of our approach to overlap detection described in Section 3. Experiments are reported in Section 4 with a detailed analysis of system performance. Our conclusions and ideas for further work are given in Section 5.

2. CONVOLUTIVE NON-NEGATIVE SPARSE CODING

Non-negative sparse coding (NSC) [11, 12] is an approach to represent non-negative, multi-variate data as a linear combination of lower rank bases. Only additive combinations are allowed in the representation 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:

$$D \approx WH \tag{1}$$

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

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

where, λ is a regularization parameter which controls the sparsity of the resulting representation.

This formulation, however, fails to capture the correlation between adjacent frames in the data matrix D that is inherent in speech signals. A convolutive variant, referred to as convolutive NSC (CNSC) [8] addresses this issue. The CNSC decomposition takes the form:

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

where P is the convolution range. The operators $\overset{p\rightarrow}{\cdot}$ and $\overset{p\leftarrow}{\cdot}$ are column shift operators which shift p columns of the matrix 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 iteratively updating W and H until convergence using the following update rules [13]:

$$W_p = W_p \odot \frac{D \overset{p\rightarrow}{H}{}^{T}}{\hat{D} \overset{p\rightarrow}{H}{}^{T}} \tag{4}$$

$$H(p) = H \odot \frac{w_p^T \overset{p\leftarrow}{D}}{w_p^T \overset{p\leftarrow}{\hat{D}} + \lambda U} \tag{5}$$

$$H = \frac{1}{P} \sum_{p=0}^{P-1} H(p), \tag{6}$$

where U is an $R \times N$ unit matrix, $\odot$ is the Hadamard product and where the division of matrices is performed element-wise. After each update of W, its columns are normalised to unit vectors. This is an essential step in sparse coding since it ensures that W does not grow in an uncontrolled manner and encourages sparse representation.

3. CNSC-BASED OVERLAP DETECTION

We show here how the CNSC algorithm can be readily applied to detect overlapping speech. CNSC bases are learnt for individual speakers such that an interval of overlapping speech can be decomposed into its underlying speaker components, thereby providing a natural solution to overlap detection. We first describe the CNSC-based decomposition of speech signals and then introduce a new frame-level approach to overlap detection.

3.1. Base learning and decomposition

CNSC bases W are learnt for each speaker in an audio document using spectral magnitude features extracted from segments of pure (non-overlapping) speech. The base patterns for each speaker are then concatenated together to create a global basis W^G that spans the spectral patterns of all speakers. Spectral magnitude features across the whole audio document, including overlapping segments, 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 in H for any given frame and any given speaker therefore serve as an indication of that speaker's activity. While the activations H and corresponding basis W can be used to reconstruct or separate each speaker's contribution to overlapping segments, we use the activations H directly to detect each speaker's activity and hence segments of overlapping speech.

3.2. Activation energy

Since the bases W are normalised, the sum of the activations for any given speaker is strongly correlated to the signal energy from that particular speaker and therefore serves as an indicator of that speaker's activity during any given frame.

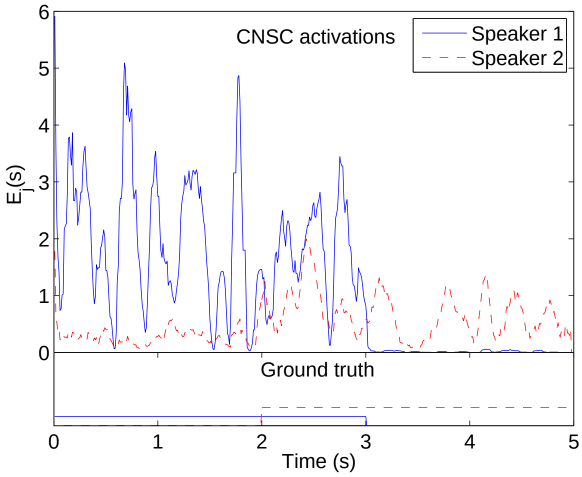

Fig. 1: An illustration of the correlation between ground-truth speaker activity (bottom) and CNSC activation energies (top) for two speakers in a conversation containing an interval of overlapping speech.

The energy for speaker s during frame j is estimated according to:

$$E_j(s) = \sum_{i \in I_s} H_{ij} \tag{7}$$

where I_s represents the speaker-specific rows in H, or the activations for speaker s.

Figure 1 (top) illustrates the CNSC activation energy against time for two speakers during a short interval from an example meeting recording where the speaker energy is calculated according to Eq. 7. Ground-truth reference speaker activities are plotted below using the same colour profile for corresponding speakers. The latter are plotted on different scales solely for clarity. It is seen that the CNSC activation energies have a clear potential as an indicator of speaker activity and as can be seen from the figure, both the speakers have high activation energy in the overlapping segment between 2 and 3 seconds.

3.3. Overlap detection

Speaker activation energies calculated as per Eq. 7 are smoothed with a moving average filter and used to implement a frame-based overlap detector. It is based on an energy ratio ER for frame j estimated as follows:

$$ER_j = \frac{E_j(\hat{s}_2)}{E_j(\hat{s}_1)} \tag{8}$$

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 ET_j by summing Eq. 7 across all speakers and filter out frames with low total energy. All frames with an energy ratio ER_j and total energy ET_j greater than empirically optimised thresholds δ_{ER} and δ_{ET} are deemed to contain overlapping speech.

In our previous work [7], we had used variance of speaker activation enegy differences in a frame as a measure for detecting overlaps. However the energy ratio measure gives much better results when used in conjunction with the total energy threshold introduced in this work.

4. EXPERIMENTS

We report here an assessment of our new overlap detection system using a subset of the AMI meeting corpus [5].

4.1. Oracle segmentation

In a practical speaker diarization scenario there is no speaker specific training data other than that contained within the audio recording itself. Consequently, the diarization system hypothesis must itself be used to estimate regions of clean speech for each speaker. Due to diarization errors, this speech material is not entirely pure, but is the only data available with which to learn speaker-specific base matrices for CNSC overlap detection. Any derived results are therefore dependent on the performance on the underlying speaker diarization system and thus the extraction of generalised results is troublesome.

In such scenarios it is typical to use oracle references to marginalise the impact of systems elements that are not under direct observation and thus to minimise their influence on observed results. This approach is adopted here; we use the reference transcription to identify intervals of pure speech for each speaker. Accordingly, results presented in this paper are independent of errors in an automatically derived speaker segmentation or diarization output and thus the assessment focuses on CNSC alone. While such an approach does not necessarily give a reliable estimate of performance under practical conditions, we note that our previous work [7] showed little difference in overlap detection performance using reference segmentations to those obtained with a real speaker diarization system.

4.2. CNSC optimisation

We used a subset of six meeting recordings for development and the same ten files for evaluation as used in previous work by other authors [3]. In all cases we used only the single-channel far-field microphone recordings. The list of used meetings is displayed in Table 1. Both development and evaluation sets contain approximately 20% overlapped speech.

Development set			
TS3009d	IS1009d	EN2009c	ES2014c
IN1016	IB4002		
Test set			
EN2003a	EN2009b	ES2008a	ES2015d
IN1008	IN1012	IS1002c	IS1003b
IS1008b	TS3009c		

Table 1: Meetings from the AMI evaluation dataset used for development and testing

The CNSC algorithm was optimised on a small, artifical 2-speaker test set where overlapping speech was manually created and controlled in order to better understand system behaviour and the influence of different parameterisations. Parameterisations reported here were subsequently re-optimised on the AMI development data and thus different to those reported previously [7]. The algorithm is applied to magnitude spectra computed on 40 ms windows (cf. 20 ms previously) with a window shift of 20 ms. CNSC speaker activiations are calculated with speaker bases of size $R = 35$ (c.f. 50 previously), a convolutional range of $P = 4$, and a sparseness parameter of $\lambda = 0.05$. The use of larger window sizes captures more dicriminative speaker features whereas the use of smaller bases leads to more effective modeling and avoids overfitting.

4.3. Metrics and assessment

Overlap detection performance is assessed using precision and recall statistics calculated at the frame level. Improvements in speaker diarization require overlap detection with high precision, whereas recall is usually of lower importance [3]. However, given that overlap detection can be applied in different processing steps of a typical speaker diarization system (namely overlap exclusion during clustering and overlap attribution during segmentation), different operating points with different precision and recall values may be beneficial. Therefore, in addition to precise figures, we also show the dynamic influence of the energy threshold δ_{ET} on the trade-off between precision and recall performance. A higher threshold will identify less overlap yielding lower recall but higher precision.

4.4. Results

The energy ratio threshold was tuned on the development set and set to $\delta_{ER} = 0.5$ across all audio recordings whereas we observed significantly better results when δ_{ET} is set dynamically for each audio recording and according to a fraction t_r of the mean energy over the entire recording. Figure 2 shows overlap detection performance in terms of precision and recall as a function of t_r (solid blue profile) and shows considerably better performance than our previous system (dashed black profile) [7]. The new energy ratio and total energy features thus give a marked improvement in system performance.

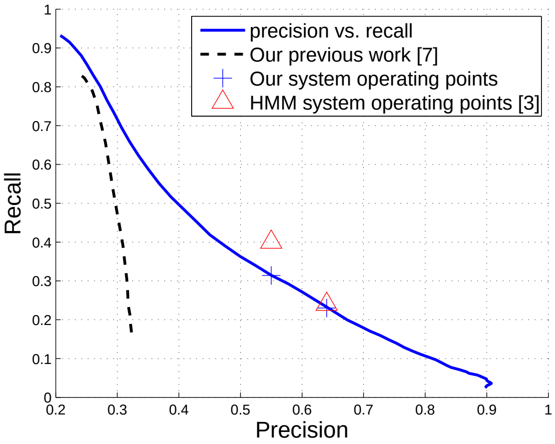

Fig. 2: Overlap detection performance in terms of precision and recall on the evaluation dataset.

Boakye et al. [3] reported experiments with the same evaluation set and report precision/recall values of 0.55/0.40 and 0.64/0.24. Our system achieves similar values of 0.55/0.31 and 0.64/0.23. The two sets of results are also illustrated in Figure 2 with red triangles and blue crosses respectively. Our system achieves comparable performance without a duration model that is implicitly inherent in the HMM based approaches.

In order to better understand the performance and weakness of our new overlap detection system we analyzed performance as a function of overlap segment duration. For this work we arbitrarily chose the first operating point with precision/recall of 0.55/0.31. Figure 3 shows four histogram plots for the test set which illustrate overlap detection performance in terms of detected and missed overlap (top right and bottom left) and recall (bottom right). For comparison reference overlap histogram is also presented (top left). The distribution of overlap segment durations show the total contribution (in seconds) to the corpus for each bin (not the number of segments with the respective length, as would be the case in a conventional histogram).

The plots show that the largest contribution to overlap comes from shorter segments with durations between 0.5 and 1.5 seconds. However, there are a suprising number of longer overlap segments with durations in excess of 4 seconds. The missed overlap and recall histograms shows that short segments, which occur most frequently, are the least well detected. Furthermore, the number of overlapping segments that have a duration greater than the standard 0.25-second collar used in the standard diarization error rate (DER) shows that a significant penality will be incurred if segments of between 0.25 and 2 seconds in duration are not detected reliably. Fu-

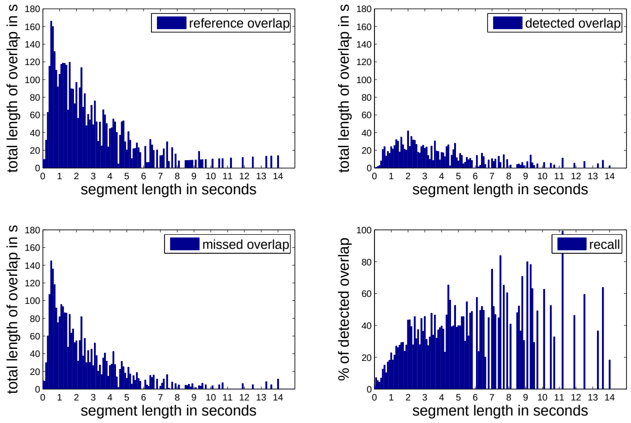

Fig. 3: Weighted length histograms for reference overlap segments, detected overlap, missed overlap and recall.

ture work should therefore focus on improving detection of shorter segments and is likely to prove a significant challenge.

5. CONCLUSIONS

This paper reports our recent advances and system enhancements in applying convolutive non-negative sparse coding (CNSC) to the detection of overlapping speech in the context of conference meetings and speaker diarization. CNSC can be used to separate a potentially overlapping speech signal into single-speaker signals. We show how the resulting CNSC base activations can be applied to detect overlapping speech segments.

The new CNSC approach gives overlap detection results which are comparable to a state-of-the-art HMM overlap detection approach, when evaluated on the AMI meeting corpus. Compared to an HMM approach, we use a rather simple classifier which is not dependent on large amounts of training data. Optimized parameterisations and new energy ratio and total energy thresholds give significantly better performance than our previous work and supports the potential for CNSC-based overlap detection. A new analysis of overlap detection performance highlights the need for continued work to improve overlap detection particularly for shorter segments of between 0.25 and 2 seconds in duration. A large part of these short overlap segments are backchannel utterances, where one speaker speaks in the middle of a longer utterance of another speaker. However, very often it is not the case that there is a real acoustic overlap between these two speakers. Therefore, these segments can not be detected by overlap detection systems which rely only acoustic features.

Our current work aims to integrate CNSC activations into an HMM overlap detection framework to exploit the benefit of duration modelling. This work is expected to improve overlap detection performance for overlapping segments of especially short and especially long duration. In addition, we experiment with the inclusion of several different energy-related features, since the introduction of the total energy threshold gave such a big improvement in system performance. Future work includes the full integration of overlap detection into a regular speaker diarization framework. In addition to continued work to develop detection performance this will require new work to optimise overlap attribution algorithms.

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] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *Proc. ASRU*, Kyoto, Japan, December 2007, pp. 683–686.

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

[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, Nevada, USA, 2008, pp. 4353–4356.

[5] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, et al., "The AMI meeting corpus: A pre-announcement," *Machine Learning for Multimodal Interaction*, pp. 28–39, 2006.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] M. Zelenak, C. Segura, J. Luque, and J. Hernando, "Simultaneous speech detection with spatial features for speaker diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 436–446, 2012.

[11] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, 2000, pp. 556–562.

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

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