SPEECH OVERLAP DETECTION AND ATTRIBUTION USING CONVOLUTIVE NON-NEGATIVE SPARSE CODING

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

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

ABSTRACT

Overlapping speech is known to degrade speaker diarization performance with impacts on speaker clustering and segmentation. While previous work made important advances in detecting overlapping speech intervals and in attributing them to relevant speakers, the problem remains largely unsolved. This paper reports the first application of convolutive non-negative sparse coding (CNSC) to the overlap problem. CNSC aims to decompose a composite signal into its underlying contributory parts and is thus naturally suited to overlap detection and attribution. Experimental results on NIST RT data show that the CNSC approach gives comparable results to a state-of-the-art hidden Markov model based overlap detector. In a practical diarization system, CNSC based speaker attribution is shown to reduce the speaker error by over 40% relative in overlapping segments.
Index Terms: overlap detection, speaker attribution, speaker diarization, convolutive non-negative sparse coding

1. INTRODUCTION

Over recent years, state-of-the-art speaker diarization systems have advanced to the point where overlapping speech can be a dominant source of error [1, 2]. The occurrence of overlap 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].

In a speaker diarization context two problems need to be addressed. The first involves the detection of overlapping speech so that it can be removed from speaker clustering and model training. The second problem involves the attribution of overlapping speech to contributing speakers and naturally depends on reliable overlap detection. There is some evidence that a solution to the first problem alone is unlikely to be sufficient and that a solution to speaker attribution is potentially more valuable [3–5].

Only a small number of attempts to treat overlapping speech have been successful. Boakye et al. [4, 6] investigated the use of multiple features for overlap detection using a hidden Markov model (HMM) based system for detection and a post-processing step for attribution. Encouraging results are reported and oracle experiments confirm the full potential.

Our approach to overlap detection and attribution involves convolutive, non-negative matrix factorisation (CNMF). CNMF captures spectro-temporal patterns and has been successfully applied to speech denoising/separation applications [7]. While initial results with the basic CNMF algorithm were encouraging, the use of sparse coding constraints [8] gave more promising results. 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.

In a practical speaker diarization system, the CNSC-based overlap detection and speaker attribution involves learning speaker-specific base patterns using the diarization output and projecting the acoustic signal onto the set of speaker bases. Base activations provide an indication of speaker-specific activity and hence can be used to both detect and attribute intervals of overlapping speech. Due to the sparseness constraints, the distribution of speaker energy is enforced to only a small number of bases and thus provides better discrimination between active and inactive speakers. Experimental results demonstrate the merit of the proposed approach and support further work to develop the potential.

2. CONVOLUTIVE NON-NEGATIVE SPARSE CODING

Non-negative matrix factorisation (NMF) [9] is an approach for the linear decomposition of a non-negative matrix $D \in \mathbb{R}^{\geq 0}_{M \times N}$ with similar non-negative constraints on the decomposed matrices $W \in \mathbb{R}^{\geq 0}_{M \times R}$ and $H \in \mathbb{R}^{\geq 0}_{R \times N}$ so that:

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

The columns of W are base vectors and the columns of H are the base activations or weights needed to recompose an estimate of the original matrix. As described in [10], the decomposition is performed iteratively using elegant and computationally efficient multiplicative update rules to minimise the distance between the original matrix and its approximation:

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

where, $\|.\|_F$ is the Frobenius norm. The matrix representation of a speech signal D typically comprises windowed magnitude spectra which satisfy the non-negative constraint. The decomposition results in base vectors that correspond to prominent spectral features. NMF, however, does not capture the correlation between adjacent frames that are inherent in speech signals. A convolutive variant, referred to as convolutive NMF (CNMF) [7] addresses this issue. The CNMF decomposition takes the form:

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

This research was supported 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/

where P is the convolution range. The operators $\overset{p\rightarrow}{\cdot}$ and $\overset{p\leftarrow}{\cdot}$ are column shift operators that shift p columns of the matrix to the right and left respectively. Vacated columns are filled with zeros. A sequence of P vectors corresponding to the i^{th} column of W_p can be viewed as a base dimension that captures one of the prominent spectro-temporal patterns in the given signal.

The further application of sparse constraints [8, 11] with the following optimisation criterion leads to a sparse activation matrix H.

$$(\hat{W}, \hat{H}) = \arg\min_{W,H} \|D - WH\|_F^2 + \lambda \sum_{ij} H_{ij} \tag{4}$$

where H_{ij} denotes the elements of H. This decomposition is referred to here as convolutive non-negative sparse coding. In our implementation, we use the update rules proposed in [8] for computing W and H:

$$W_p = W_p \odot \frac{D \overset{p\rightarrow}{H}^T}{\hat{D} \overset{p\rightarrow}{H}^T} \tag{5}$$

$$H(p) = H \odot \frac{w_p^T \overset{p\leftarrow}{D}}{w_p^T \overset{p\leftarrow}{\hat{D}} + \lambda U} \tag{6}$$

$$H = \frac{1}{P} \sum_{p=0}^{P-1} H(p) \tag{7}$$

where $\odot$ is the Hadamard product and where the division of matrices is performed element-wise. U is an $R \times N$ unit matrix. W and H are updated iteratively until $\hat{D}$ converges. After each update of W, 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 forces the resulting activations to be sparse.

3. OVERLAP DETECTION

Here we describe our approach to apply CNSC to the detection of overlapping speech. Attribution, where we aim to determine contributing speakers, is covered in Section 4. We first consider performance where the ground-truth reference is used to learn speaker bases and then assess performance using an automatic segmentation output from a practical speaker diarization system.

3.1. Ground-truth references

CNSC is implemented according to the following procedure:

1. Using pure (non-overlapping) speech for each given speaker, learn base matrices W using spectral magnitude features.
2. Concatenate together the W's for all speakers to create a global set W^G that spans the spectral patterns of all speakers.
3. Decompose the magnitude spectrum of a mixed, and possibly overlapping speech signal (same speakers as in 1.) according to W^G and update only H to minimise the optimisation criterion.

The activations of H which correspond to the bases for any given speaker therefore serve as an indication of that particular speaker's activity. Since the bases W are normalised, the sum of the activations for a speaker is strongly correlated to the signal energy from that particular speaker. The speaker energy is determined according to:

$$E_j(s) = \sum_{i \in I_s} H_{ij} \tag{8}$$

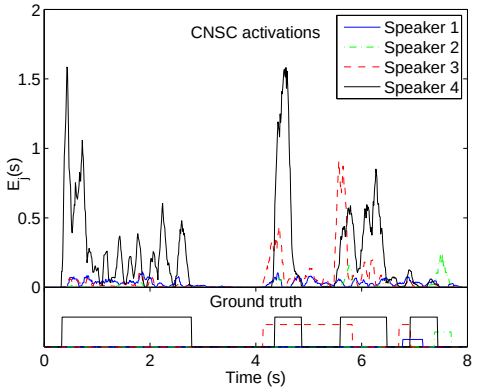

Fig. 1. An illustration of the correlation between ground-truth speaker activity (bottom) and CNSC activation energies (top) for 4 speakers in a conversation.

where I_s is the set of rows in H corresponding to the bases of speaker s and where j is the frame index.

Figure 1(top) illustrates the energy profiles against time for all speakers during a short interval from an example meeting recording. 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 and show that, for the most part, there is only one or two active speakers. Between 6.5s and 8s, however, there are four active speakers. Energy profiles correlate well with the ground-truth references and thus support the utility of CNSC activations as an indicator of speaker activity.

In our approach speaker energies calculated as per Eq. 8 are then smoothed with a moving average filter and used to implement a frame-based overlap detector. At this step, the output of a voice activity detection (VAD) component can optionally be used to identify and remove nonspeech frames. We performed experiments with reference VAD transcriptions and with or without the VAD of our baseline diarization system.

A couple of measures were considered to detect overlapping speech, namely, the ratio of two highest speaker energies in a frame and the variance of the energy difference. The latter gave better performance and was thus used for all experiments reported here. To compute this measure, the speaker energy is normalised according to the highest energy across all speakers. The difference in normalised energy is then calculated for all speaker pairs. The inverse variance of energy differences forms the measure.

$$m_j = \frac{1}{\text{var(energy difference)}} \tag{9}$$

This value is expected to be higher in overlapping speech segments than in non-overlapping segments since the variation in speaker energies should be respectively lower. After smoothing, overlapping speech frames are detected by comparing m_j to a threshold t_{det}.

We evaluated our approach to overlap detection using a set of 16 single distant microphone (SDM) conference meeting waveforms from the standard NIST RT evaluation dataset. To compute the speaker bases for each evaluation file, pure speech was first obtained for each known speaker according to the reference transcripts in an oracle-style experiment. This was done to avoid the impact of errors in an automatically derived speaker segmentation or diarization output and thus to focus the assessment on CNSC alone. We used 50 bases for each speaker with a convolutional range of 4 and a sparseness parameter of 0.2 chosen heuristically. Features are magnitude spectra computed on 20ms windows with a window shift of 10ms.

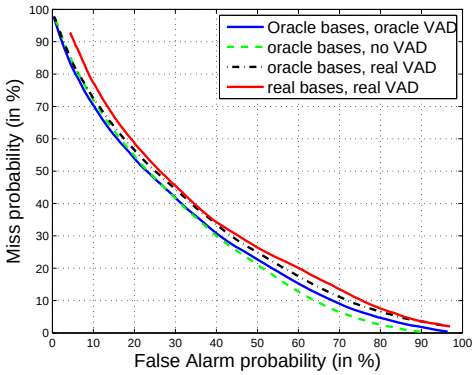

Fig. 2. Detection error tradeoff curve using the CNSC approach, using either oracle bases or real bases and either oracle VAD, no VAD or a real VAD component.

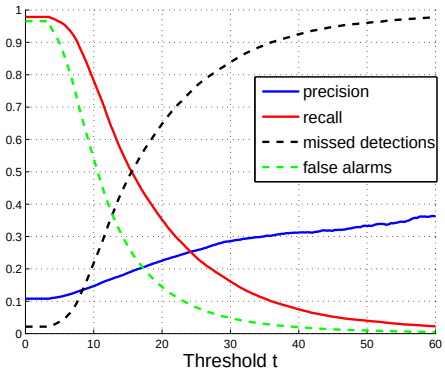

Fig. 3. Precision, recall, missed detection probability and false alarm probability (using real VAD output) for overlap detection

Overlap detection performance for varying t_{det} is illustrated in the detection error trade-off (DET) curve shown in Fig. 2 for a practical (real) VAD. The equal error rate (EER) is approximately 36%. Profiles and EERs obtained without (no) VAD and with a reference (oracle) VAD are almost identical.

For consistency with prior work, we also consider precision and recall metrics. Improvements in speaker diarization require overlap detection with high precision, whereas recall is of lower importance [4]. Fig. 3 shows the variation in precision, recall, false alarms and missed detections for different thresholds t_{det}. A value of $t_{det} = 30$ yields a precision of 28.6 % while the recall is 16.1 %. With an oracle VAD and the same t_{det} a precision of 32.7 % and a recall of 16.0 % are obtained, while with no VAD, the precision and recall are 23.7 % and 16.2 % respectively.

We compared the performance of the proposed system to a state-of-the-art overlap detector based on [6], where intervals of speech, nonspeech and overlap are modeled as 3 state HMMs and where each state is modeled by a 32 component Gaussian mixture model (GMM). Mel-frequency cepstral coefficients (MFCCs) with first differential and energy coefficients were used as features which are computed using a window size of 60ms and a frame shift of 20ms as in [6]. A set of 23 NIST RT evaluation files, disjoint from the test set, were used to train the models. Recall was traded off for higher precision by varying the speech to overlap transition penalty during Viterbi decoding. The best precision/recall rate for this system is 28.0% / 26.0% and the precision peaks at around 28.0%. The same system when run on AMI SDM corpus using a similar train, tune and test set as in [4], gives a precision/recall rate of 60.0% / 20.0%.The low precision rate obtained with both the systems on NIST RT corpora highlight the particularly challenging nature of this corpora for overlap detection.

System	precision (in %)	recall (in %)
oracle bases, oracle VAD	32.7	16.0
oracle bases, real VAD	28.6	16.1
oracle bases, no VAD	23.7	16.2
real bases, real VAD	21.2	24.8
HMM	28.0	26.0

Table 1. A summary of precision/recall values on overlap detection for all the experiments

3.2. Automatic segmentation

CNSC relies on the availability of pure speech to train speaker bases and, in the experiments reported above, this was done using reference transcripts to avoid the influence of errors in an automatically derived segmentation. We now aim to assess performance using the output of a practical diarization system, rather than the ground-truth reference, in an otherwise identical setup. This work was undertaken using the top-down speaker diarization system reported in [12].

Perhaps the most significant difference between the reference and the diarization output lies in the number of real and automatically detected speakers which will naturally lead to increased error. Overlap detection performance using the real diarization output results in an EER of 37.0%. The DET curve for this setup is also illustrated in Figure 2. At an operating point defined by $t_{det} = 30$, the system achieves a precision/recall of 21.2% / 24.8%. The precision peaks around this value.

Table 1 illustrates a comparative summary of all experiments on overlap detection using bases obtained from oracle references or real speaker diarization system outputs. Also shown is the variation in performance when using an oracle VAD, a real VAD or no VAD. Performance is also compared to the baseline HMM system. Naturally, using oracle bases and oracle VAD gives the best results, and there is a decrease in performance when bases computed from the output of a real diarization system are used. This is due to the impurities in speaker training data. An obvious advantage for the HMM system is the inherit exponential duration decay which allows statistical-based smoothing, while in our CNSC based approach, the analysis was done at frame level. However, the fact that our system performs reasonably well in comparison to the HMM system is very encouraging and motivates future work in this direction.

4. SPEAKER ATTRIBUTION

We now turn to the attribution of overlapping speech to contributing speakers. Attribution has the potential to improve the (Diarization error rate (DER) by reducing missed speech errors in intervals of speech containing more than a single speaker. We assess attribution performance independently from overlap detection by using an oracle overlap detection component. For each segment of overlapping speech, the energy of each speaker is determined from the base activations in exactly the same way as described in Section 3.1 for all frames in the segment. We assume that, in each interval of overlapping speech and as is generally the case in practice, there are exactly two active speakers. We further assume that they correspond to the

System setup	Speaker error (in %)
Baseline diarization system	68.4
GMM posterior based system	44.7
CNSC, Oracle bases	31.5
CNSC, Real bases	40.4

Table 2. Results for speaker attribution. Shown is the speaker error on all true overlap segments.

two speakers with the highest energy as determined by CNSC activations. To assess attribution performance, we use a metric which is adapted from the standard DER formula to focus on speaker error (SpkErr) only. The speaker attribution error is calculated according to:

$$SpkErr = \frac{\sum T(k)[max(N_{Ref}, N_{Hyp}) - N_{Corr}]}{\sum T(k) N_{Ref}} \quad (10)$$

where, $T(k)$ is the duration of the k^{th} overlapping speech segment, N_{Ref}, N_{Hyp} and N_{Corr} are the number of speakers in the reference hypothesis, detection hypothesis and those correctly attributed to the segment respectively. Note that the metric is time-weighted in a similar manner as the standard DER.

We first compute the speaker error rate from the diarization system output focusing only on intervals of overlapping speech. In this case the speaker error rate is 68.4% and is naturally high since, with no provision for overlap attribution, the minimum error is 50%. Then as a baseline result, we use a state-of-the-art speaker attribution system [4], where frame-level GMM posteriors for each speaker obtained from the diarization system are summed over all the frames in an overlapping interval and the interval is attributed to the two speakers with highest scores. With this system we obtain a speaker error rate of 44.7%. When the CNSC activations (obtained with speaker bases computed with reference transcripts) are used for speaker attribution, the resulting speaker error rate falls to 31.5%. When using speaker bases created with the output of a real diarization system instead of the reference transcripts, the resulting speaker error rate is 40.4%. Thus, using CNSC, the speaker error rate in overlapping intervals is reduced by about 40% relative over the baseline diarization system, and compared to the GMM posterior score based system, the results are improved by about 9.6% relative. These results are summarized in Table 2. However, we must note that when performing speaker attribution experiments with real overlap detection output with low precision, some errors will be introduced in the falsely detected overlapping segments.

Finally, the contribution of overlap attribution on the overall DER is shown in Table 3. According to speaker attribution results, the overlapping intervals in the diarization output were relabelled with the detected speakers. With oracle overlap detection and real speaker attribution, the errors due to missed speakers can be reduced by 64% relative. There is a small increase in false alarms due to some incorrect attributions, but overall the diarization results improve by about 6.5% relative.

System setup	DER	Missed	False alarm	SpkErr
Baseline diarization	28.50	6.4	2.9	19.2
CNSC, Oracle bases	26.15	2.3	4.2	19.6
CNSC, Real bases	26.65	2.3	3.6	20.7

Table 3. Influence of overlap attribution, using oracle overlap detection, on overall DER (%).

5. CONCLUSIONS

This paper reports an investigation into the use of convolutive non-negative matrix factorisation with sparse constraints (CNSC) for the detection and attribution of overlapping speech in the context of speaker diarization. The CNSC approach gives overlap detection results which are comparable to a state-of-the art HMM overlap detection approach. It is also seen to perform well in the case of attributing an overlapping speech interval to contributing speakers. A limitation of the approach relates to the cross-projection of a speaker's energy onto the bases of other speakers. This is to be expected since the bases are purely spectral representations and are thus not entirely decorrelated across speakers. The application of sparse constraints alleviates the problem to some extent by encouraging activations to be concentrated on a small number of bases. Further work is nevertheless required to optimise the number of bases, the convolution length and sparseness constraints to reduce cross projection. Our current work aims to integrate CNSC activations into HMM overlap detection framework to exploit the benefit of duration modelling. Future work could include an analysis of different speaker bases to detect speakers with multiple models in a typical diarization system and the full integration of CNSC into a regular speaker diarization framework. This should include a thorough study of the impact of overlap on speaker diarization.

6. REFERENCES

[1] M. Huijbregts and C. Wooters, "The blame game: performance analysis of speaker diarization system components," in *Proc. INTERSPEECH'07*, Antwerp, Belgium, September 2007, pp. 1857–1860.

[2] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, O. Vinyals, and G. Friedland, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing, Special Issue on New Frontiers in Rich Transcription*, 2011.

[3] C. Fredouille and N. Evans, "The influence of speech activity detection and overlap on the speaker diarization for meeting room recordings," in *Proc. INTERSPEECH'07*, Antwerp, Belgium, September 2007, pp. 2953–2956.

[4] K. Boakye, O. Vinyals, and G. Friedland, "Improved overlapped speech handling for speaker diarization," in *Proc. INTERSPEECH'11*, Florence, Italy, August 2011, pp. 941–944.

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

[6] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved diarization in multi-party meetings," in *Proc. ICASSP'08*, Las Vegas, Nevada, USA, 2008, pp. 4353–4356.

[7] 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.

[8] W. Wang, "Convolutive non-negative sparse coding," in *IEEE International Joint Conference on Neural Networks.*, 2008, pp. 3681–3684.

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

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

[11] P. D. O'Grady and B. A. Pearlmutter, "Convolutive non-negative matrix factorisation with a sparseness constraint," in *Machine Learning for Signal Processing*, 2006, pp. 427–432.

[12] S. Bozonnet, N. W. D. Evans, and C. Fredouille, "The LIA-EURECOM RT'09 Speaker Diarization System: Enhancements in speaker modelling and cluster purification," in *Proc. ICASSP'10*, Dallas, Texas, USA, March 2010, pp. 4958–4961.