EURECOM
Department of Multimedia Communications
2229, route des Crêtes
B.P. 193
06560 Sophia-Antipolis
FRANCE

Research Report RR-11-257

Speech overlap detection using convolutive non-negative sparse coding

May 4th, 2011
Last update May 4th, 2011

Ravichander Vipperla, Dong Wang, Simon Bozonnet and Nicholas Evans

Tel : (+33) 4 93 00 81 00
Fax : (+33) 4 93 00 82 00
Email : {vipperla, wangd, bozonnet, evans}@eurecom.fr

[1]EURECOM's research is partially supported by its industrial members: BMW Group, Cisco, Monaco Telecom, Orange, SAP, SFR, Sharp, STEricsson, Swisscom, Symantec, Thales.

Speech overlap detection using convolutive non-negative sparse coding

Ravichander Vipperla, Dong Wang, Simon Bozonnet and Nicholas Evans

Abstract

Overlapping speech is known to degrade speaker diarization performance with impacts on both speech activity detection, speaker clustering and segmentation (speaker error). While previous related work has made important advances the problem remains largely unsolved. This paper reports early work to investigate the application of non-negative matrix factorisation (NMF) to the overlap problem. NMF aims to decompose a composite signal into its underlying contributory parts and is thus naturally suited to tasks of detecting overlap and its attribution to contributing speakers. With additional sparse constraints the algorithm is shown to be effective in identifying overlapping speech and gives a relative improvement of 11% in terms of equal error rate over a baseline approach based on conventional Gaussian mixture models. Experiments with source attribution show a relative improvement in the order of 40%.

Index Terms

speech overlap detection, speaker diarization, convolutive nonnegative matrix factorization, sparse coding

Contents

List of Figures

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 occurance of overlap is typical in uncontrolled, spontaneous scenarios such as that of conference meetings which have been the focus of the NIST Rich Transcription evaluations since 2004[1].

The effects of overlapping speech in a speaker diarization context are well known and generally considered to be two-fold. Without some means of detection, segments of overlapping speech lead to impurities in speaker specific models and hence reduced segmentation performance. Further errors are incurred since it is then neither possible to attribute segments of overlapping speech to their contributing speakers; most systems assume that only a single speaker is active at any one time.

Only a small number of attempts to treat overlapping speech have been successful. Two problems need to be addressed. The first involves the detection of overlapping intervals so that they can be removed from speaker clustering and model training. The second problem involves the attribution of intervals 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 [3] and that a solution to speaker attribution is potentially more rewarding. Otterson [4] reaches similar conclusions.

The first work to detect overlap automatically appeared in 2008. Boakye et al. [5] investigated the use of multiple features for overlap detection and a post-processing step for attribution but results showed only modest improvements in diarization performance. This work was extended in [6] with new features and a new pre-processing step to remove intervals of overlap from initial clustering. Greater improvements in performance are reported and oracle experiments confirm the full potential. Huijbregts et al. [7] report a similar approach whereby a model of overlapping speech, trained on data localised around speaker turns, is used for overlap detection. A similar approach to that in [4] is applied to attribution and modest improvements in diarization performance are achieved. Finally we include reference to more recent work [8] which utilises spatial or localisation features in addition to conventional acoustic features. Our particular interest is in single-microphone data, however, where localisation features are not relevant.

Until recently our own efforts in automatic overlap detection have been mostly unsuccessful. Our initial efforts involved the analysis of speaker-specific Gaussian mixture model (GMM) likelihoods that emerge from speaker diarization but results were discouraging. More recent attempts with non-negative matrix factorisation (NMF) also gave poor results but led us to consider sparse coding constraints [9] which seems much more promising. NMF is a matrix decomposition technique that can be viewed as a parts of object based decomposition and has

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

found wide application in image processing [10]. Our motivation stems from the successful application of a convolutive variant of NMF (CNMF) which captures dynamics in the time series and has been used successfully in speech denoising applications [11]. A more recent development in this approach is the application of sparseness constraints during matrix decomposition [9, 12]. Sparseness constraint is more appealing from a speaker attribution perspective as will be discussed in the following sections.

Our approach relies upon the output of a standard diarization system in order to learn bases which span a speaker-specific acoustic space. The idea is to project segments of speech into the set of speaker spaces and hence to determine whether there is more than a single active in addition to *which* speakers are active. Thus the approach addresses both fundamental problems of overlap detection and source attribution. Initial results are encouraging and, while the work is at an early stage, we believe the approach warrants greater attention.

2 CNMF with sparseness constraints

Non-negative matrix factorisation [13] is an approach for the linear decomposition of a non-negative matrix $D \in \Re^{\geq 0}_{M \times N}$ with similar non-negative constraints on the decomposed matrices $W \in \Re^{\geq 0}_{M \times R}$ and $H \in \Re^{\geq 0}_{R \times N}$:

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

The columns of W can be seen as the basis vectors and the rows of H as the basis activations or weights to recompose an estimate of the original matrix. As described in [14], the decomposition is performed iteratively using elegant and computationally efficient multiplicative update rules to minimise the distance between the data 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 is typically comprised of windowed magnitude spectra which satisfy the non-negative constraint. The decomposition of this matrix results in basis 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) [11] addresses this shortcoming. The decomposition in CNMF 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 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} columns of W_p can be treated as one basis dimension that captures one of the prominent spectro-temporal patterns in the given signal.

The further application of sparse constraints [9, 12] leads to a sparse activation matrix H, which is a useful feature in applications where there is a need to force the decomposition onto as few bases as possible. This leads to the following optimisation criterion:

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

where H_{ij} denotes the elements of H. In our implementation, we use the update rule proposed in [9] for computing W and H:

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

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

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

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 to D. After each update of W, the 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 enforces the resulting activations to be sparse.

3 CNSC for overlap detection

We here describe our approach to apply CNSC to the detection of overlapping speech. Attribution, where we aim to determine the 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

According to the outline presented above, 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 Ws for all the speakers to create a global set W^G that spans the spectral patterns for 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 corresponding to the basis for each speaker therefore serve as an indication of that particular speaker's activity. Since the basis W is normalised, the sum of the activations in a column of H is strongly correlated to the signal energy from that particular speaker in the corresponding time or analysis window. The speaker energy is determined according to:

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

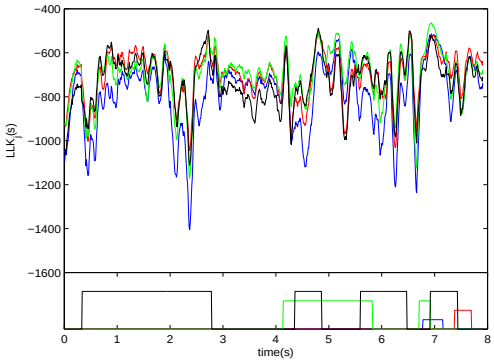
where, I_s is the set of rows in H corresponding to the basis of speaker s and j is the frame index.

We evaluated our approach to overlap detection using a set of 15 conference meeting files from the standard NIST Rich Transcription and AMI evaluation datasets. To compute the speaker basis for each evaluation file, pure speech was first obtained for each 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 basis vectors for each speaker with a convolutional range of 4.

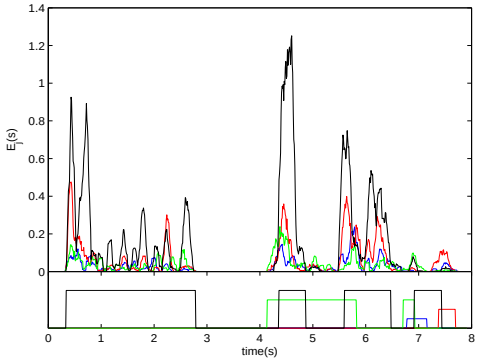
In order to compare performance with a traditional, baseline GMM-based approach, we undertook a similar experiment using speaker-specific GMMs which were trained on Mel-frequency cepstral coefficients, again using pure speech in a similar oracle-style setup. Each speaker model is comprised of 16 components. The log likelihood (LLK) for each of the speaker models ($LLK_j(s)$) is computed for each frame j and is used as a indication of each speaker's activity in the same way as the frame energy is used in the case of CNSC.

Results are illustrated in Figure 1 which shows the LLK and energy for the GMM (a) and CNSC (b) approaches respectively. Ground-truth reference speaker activities are plotted below using the same profile for corresponding speakers. The latter are plotted on different scales solely for clarity and show that, for the most part, there are only two active speakers. Between 6.5s and 8s, however, there are four active speakers. For the GMM approach, there is little correlation between the LLK and ground-truth speaker activity whereas for the CNSC approach the energy profiles appear to correlate well. We thus surmise that CNSC offers relatively better potential as an indicator of speaker activity.

In order to implement a classifier capable of detecting overlapping speech it is necessary to threshold the LLK or energy profiles. In the case of CNSC the ratio of the second highest to the highest speaker energy is computed for each frame:



(a) GMM



(b) CNSC

Figure 1: An illustration of the correlation between ground-truth speaker activity and (a) LLK scores obtained from the GMM approach and (b) energy scores with CNSC approach.

$$scoreCNSC_j = \frac{E_j(\hat{s}_2)}{E_j(\hat{s}_1)} \tag{9}$$

where $\hat{s}_i$ denotes the speaker with the ith highest energy. For overlapping segments we expect the ratio to be nearer to unity while for non-overlapping segments the ratio should be closer to zero.

An identical strategy is adopted for the GMM approach. Here, though, given that LLK scores are in the log-domain, a similar ratio is calculated as follows:

$$scoreGMM_j = LLK_j(\hat{s}_2) - LLK_j(\hat{s}_1) \tag{10}$$

Performance for both GMM and CNSC approaches with varying thresholds is illustrated in the detection error trade-off (DET) curves of Figure 2. EERs of 50.1% and 44.4% for the GMM and CNSC approaches respectively correspond to a relative improvement of 11% and confirm the potential of the CNSC approach.

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 [15].

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 is also plotted in Figure 2 and shows that, reasuringly, there is only a negligable difference in performance. This further supports our view that the use of overlapping segments for clustering is not overly problematic and that greater attention should be placed on attribution.

3.3 Discussion

We acknowledge that the EERs reported here are high but note that the correspondence of the EER to the diarization error rate (DER) is unknown and certainly complex; in our experience intermediate assessments with apparently poor results do not always correlate with the resulting DER. We also highlight that the EER is just one operating point and that, by choosing a different threshold, one may trade false alarms for missed overlap. Further work is required to investigate the resulting effects on speech/non-speech activity detection and speaker error rates.

Finally, other experiments, not reported here, show that actual precision and recall rates (for a given, fixed threshold, i.e. a single point on the DET profile) can be significantly improved through the smoothing or filtering of profiles shown in Figure 1. Such practice is common with speaker diarization. Thus performance is in practice significantly better than the impression given in Figure 2.

4 CNSC for speaker attribution

We now turn to the attribution of overlapping speech to contributing speakers. Reliable attribution has the potential to improve the DER by reducing missed speech errors where an interval of speech containing more than a single speaker is attributed to only one.

One simple approach involves the thresholding of each speaker's energy profile in order to detect active speakers. Let $E_{<k>}(s)$ be the total energy attributed to speaker s in segment k

$$E_{<k>}(s) = \sum_{j \in k} E_j(s) \tag{11}$$

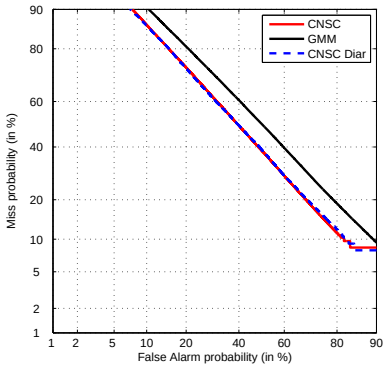


Figure 2: Detection error tradeoff curves using CNSC and GMM approaches with ground-truth references and CNSC using an automatic diarization output.

were j is the frame index. Speaker s is deemed to be active if $E_{<k>}(s) \geq \delta * E_{<k>}(\hat{s}_1)$, where δ is an empirically optimised threshold and where $\hat{s}_1$ is the speaker with the highest energy in the same segment. The latter acts to normalise speaker contributions and vocal effort and works well in practice.

Performance is again compared to that achieved using a similar criterion using the LLK and GMM-based approach. In this case a speaker is deemed to be active if $LLK_{<k>}(\hat{s}_1) - T(k) * G * \log(\delta)$ where $T(k)$ is the duration of the segment and G is the number of Gaussian components in the GMM. Since we are operating in the log domain, we need to scale the threshold as shown to obtain a fair comparison between the two approaches.

To assess the performance of each approach we use a metric which has been adapted from the standard formula for the DER and here concentrate on speaker error (SpkErr) only, i.e. we discount speech activity detection. Errors in speaker attribution are calculated over all segments containing overlap according to:

$$Error = \frac{\sum T(k)[max(N_{Ref}, N_{Hyp}) - N_{Corr}]}{\sum T(k) N_{Ref}} \tag{12}$$

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

The attribution error for both GMM and CNSC approaches is shown in Figure 3 against the threshold δ which varies between 0.1 and 0.9. Whereas the profile for the GMM approach is relatively flat it descends rapidly for the CNSC algorithm as the threshold increases. Optimum performance is achieved with a value of $\delta = 0.7$. CNSC thus clearly outperforms the GMM approach in the case of source attribution and delivers an improvement close to 40% relative.

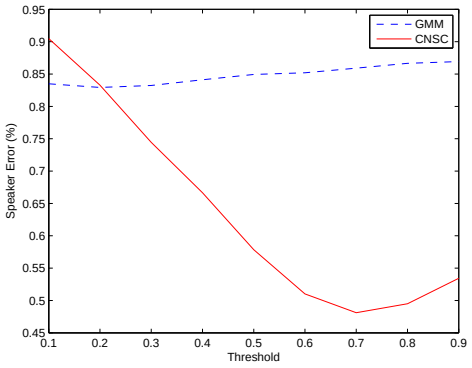


Figure 3: Speaker Error: CNSC and GMM

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 is seen to outperform a more conventional approach based on the likelihoods obtained from Gaussian mixture speaker models. A relative reduction of 11% in equal error rate is obtained in terms of detection but improvements in attribution are in the order of 40% relative. 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 are not orthogonal. The application of sparse contraints alleviates the problem to some extent by encouraging activations to be concentrated on a small number of bases but further work is required to optimise the number of basis dimensions, the convolution length and sparseness constraints to reduce cross projection and hence improve performance. Future work should include an analysis of different speaker bases to detect speakers with multiple models 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.

References

[1] M. Huijbregts and C. Wooters, "The blame game: performance analysis of speaker diarization system components," in *INTERSPEECH*, August 2007, pp. 1857–60.

[2] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, O. Vinyals, and G. Friedland, "Speaker diarization: A review of recent research," *submitted to IEEE*

Transactions on Audio, Speech, and Language Processing, Special Issue on New Frontiers in Rich Transcription, 2010.

[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*, September 2007.

[4] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *in Proc. ASRU 2007*, 2007, pp. 686–6.

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

[6] K. Boakye, O. Vinyals, and G. Friedland, "Two's a crowd: improving speaker diarization by automatically identifying and excluding overlapped speech," in *Proc. INTERSPEECH*, vol. 1, September 2008, pp. 32–5.

[7] M. Huijbregts, D. van Leeuwen, and F. de Jong, "Speech overlap detection in a two-pass speaker diarization system," in *Proc. Interspeech*, 2009, pp. 1063–1066.

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

[9] W. Wang, "Convolutive non-negative sparse coding," in *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, 2008, pp. 3681 –3684.

[10] D. Guillamet, B. Schiele, and J. Vitri, "Analyzing non-negative matrix factorization for image classification," in *In Proc. 16th Internat. Conf. Pattern Recognition (ICPR02), Vol. II, 116119. IEEE Computer Society*, 2002, pp. 116–119.

[11] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 1–12, 2007.

[12] P. O'Grady and B. Pearlmutter, "Convolutive non-negative matrix factorisation with a sparseness constraint," in *Machine Learning for Signal Processing, 2006. Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on*, 2006, pp. 427 –432.

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

[14] ——, "Algorithms for non-negative matrix factorization," in *NIPS*, 2000, pp. 556–562.

[15] 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 14-19 2010.