Detecting Overlapping Speech with Long Short-Term Memory Recurrent Neural Networks

Jürgen T. Geiger[1], *Florian Eyben*[1], *Björn Schuller*[2,1] *and Gerhard Rigoll*[1]

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
[2]Institute for Sensor Systems, University of Passau, Germany
{geiger,eyben,rigoll}@tum.de, bjoern.schuller@uni-passau.de

Abstract

Detecting segments of overlapping speech (when two or more speakers are active at the same time) is a challenging problem. Previously, mostly HMM-based systems have been used for overlap detection, employing various different audio features. In this work, we propose a novel overlap detection system using Long Short-Term Memory (LSTM) recurrent neural networks. LSTMs are used to generate framewise overlap predictions which are applied for overlap detection. Furthermore, a tandem HMM-LSTM system is obtained by adding LSTM predictions to the HMM feature set. Experiments with the AMI corpus show that overlap detection performance of LSTMs is comparable to HMMs. The combination of HMMs and LSTMs improves overlap detection by achieving higher recall.
Index Terms: Speech Overlap Detection, Speaker Diarization, Neural Networks, Long Short-Term Memory

1. Introduction

In spontaneous, conversational speech, it occurs very often that two or more speakers are speaking simultaneously [1]. Overlap occurs at speaker turn points or as backchannel utterances or interruptions. Recent studies try to further analyse the nature of overlapping speech, revealing results about the duration and when it is likely to occur or about different types of overlapping speech [2, 3, 4]. Overlapping speech is still a major source of error for many speech processing applications [5]. For example, it is a problem for speech recognition systems for conversational speech [6]. In [7], overlap detection is applied to improve a speaker recognition system. Another important field of application for overlap detection systems is overlap handling in speaker diarization [8]. In speaker diarization, overlap leads to impure speaker models and furthermore directly provokes increases in the missed speaker rate. Overlap detection systems with good performance in precision and recall can address these two problems.

The first system for detection and handling of overlap in speaker diarization was presented in [9]. An HMM system with three classes (non-speech, speech, and overlapping speech) employing mainly spectral audio features (MFCCs, RMS energy, LPC residual energy, and diarization posterior entropy) was used to detect overlap. Detected overlap segments were then excluded prior to speaker clustering to obtain better speaker models. To reduce the missed speaker rate, a second speaker was introduced in all detected overlap segments. More recently, other features have been investigated for overlap detection. Prosodic [10] and spatial [11] features were able to improve the performance of an HMM-based overlap detection system.

In our previous work, we explored the use of convolutive non-negative sparse coding (CNSC) for overlap detection [12]. CNSC is a signal separation technique which is used to separate potentially overlapping speech signals into their contributing sources. The resulting speaker activations are used to detect overlap. In [13], we combined CNSC-based overlap detection with an HMM system by using features derived from CNSC activations within the HMM framework. In addition, more spectral, energy and voicing-related features were investigated for their suitability for overlap detection. An analysis of detected overlap segments showed that especially short segments of overlapping speech are hard to detect [14]. Such segments include backchannel utterances or interruptions, which are characterised by a low degree of actual acoustic overlap. Therefore, systems that go beyond pure acoustic features and try to analyse the context could help to improve overlap detection. An approach presented in [15] uses the output of a voice activity detection system and the silence distribution to detect overlap. This work was extended by exploiting long-term conversational features for overlap detection [16]. Neural networks could improve overlap detection by analysing the context. In the domain of speech recognition, tandem architectures which combine neural networks and HMMs have been applied succesfully [17]. However, the amount of context a conventional recurrent neural network (RNN) can exploit is limited. Long Short-Term Memory (LSTM) RNNs have been proposed to overcome this so-called vanishing gradient problem [18]. Recently, we used LSTMs for voice activity detection [19].

In this work, we apply LSTM-RNNs to the the task of speech overlap detection. Using conventional MFCC features and energy, spectral, vocing-related and CNSC-based features that were proposed in [13], LSTMs are used as a regressor to predict frame-wise overlap scores. These scores are employed to detect segments of overlapping speech. In addition, we use the predicted overlap scores as features within the HMM framework. Experiments are conducted with the AMI corpus of meeting recordings containing spontaneous speech. Results show the efficacy of LSTMs for overlap detection.

2. Overlap Detection System

The proposed overlap detection system is depicted in Figure 1. It consists of a conventional HMM system for overlap detection. In addition, extracted audio features can also be fed to the LSTM-RNN to generate overlap predictions. These overlap predictions are either directly used (by applying a threshold) to detect overlap or they are added to the other features and decoded with the HMM, resulting in a tandem HMM-LSTM system.

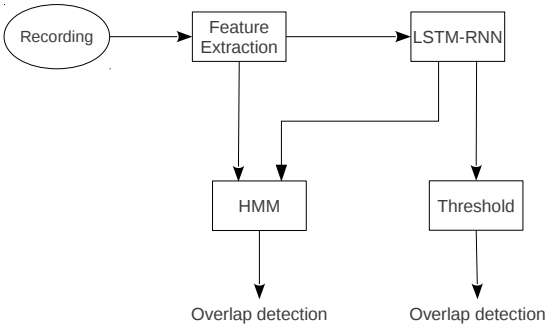

Figure 1: System overview for the overlap detection system

2.1. HMM System

As a baseline system, a standard HMM-based overlap detection system, as first presented in [9], is applied. Speech, non-speech and overlapping speech are each modelled by a three-state HMM. Observations are modelled 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. In the decoding grammar, self-transitions and transitions from non-speech to overlapping speech are forbidden. In order to trade off false positive detections versus false negatives, different system operating points are tested. The log-likelihood transition penalty from speech to overlapping speech is the tuning parameter to obtain these operating points. This parameter is also referred to as the overlap insertion penalty OIP. Higher OIP leads to fewer false positive detections and in turn higher precision and lower recall.

2.2. Audio Features

Two different sets of audio features are tested with the HMM and the LSTM. The first feature set (denoted as MFCC) consists of MFCCs 1-12. The second feature set (denoted as ESVC) contains, in addition to MFCCs, various energy, spectral, voicing-related and CNSC-based audio features. Table 1 lists all features in the ESVC feature set. This feature set has been determined in a previous work [13], where feature selection was employed, based on a larger feature set, to determine features that are best suited for overlap detection. In addition to conventional MFCCs, which have been used for overlap detection in prior work [9], energy features are expectably good indicators for overlapping speech. Jitter and shimmer are measures of fluctuations in fundamental frequency and amplitude respectively, and are thus also ideally suited. All energy, spectral and voicing-related features are computed with our freely available openSMILE feature extraction toolkit [20].

CNSC [21, 22] is an approach to represent non-negative, multi-variate data as a linear combination of lower rank bases. The application to overlap detection was first reported in [12]. In all work reported here, we used an approach proposed in [23, 24]. Bases W are learned for each speaker in an audio document using spectral magnitude features extracted from segments of preferably pure (non-overlapping) speech. These speaker-specific training data are obtained with a speaker diarization algorithm. For this work, the LIA-Eurecom speaker diarization system [25] was used. The base patterns of each

Energy & spectral features (18)
MFCC 1-12
loudness (auditory model based)
energy in band 250 - 650 Hz
energy in band 1 kHz - 4 kHz
spectral flux
spectral kurtosis
spectral harmonicity
Voicing-related features (3)
probability of voicing
jitter
shimmer (local)
CNSC-based features (2)
CNSC energy ratio
CNSC total energy

Table 1: Employed energy, spectral, voicing-related and CNSC-based (ESVC) features

speaker are concatenated to create a global basis. When the spectral magnitude features of a recording are decomposed or projected onto each speaker basis, the resulting activations H reflect each speaker's activity. Summing over all activations for a given speaker s leads to an estimate of the speaker energy $E_j(s)$ for frame j. Two features are computed from this speaker energy. The first CNSC-based feature is the CNSC energy ratio,

$$ER_j = \frac{E_j(\hat{s}_2)}{E_j(\hat{s}_1)} \tag{1}$$

which reflects the difference in activation energy between the two most active speakers. The second CNSC-based feature is the CNSC total energy,

$$ET_j = \sum_{s \in S} E_j(s) - \frac{f}{|J_{sp}|} \sum_{j \in J_{sp}} \sum_{s \in S} E_j(s) \tag{2}$$

which is the sum of all speaker energies, normalised by the mean over all the speech frames J_{sp}. Here, f is a regularisation factor tuned on held-out development data. Full details of the CNSC feature extraction are reported in [13].

Finally, the feature set is augmented with first order regression coefficients and normalised using the statistics of the training set to have zero mean and unity variance.

3. Long Short-Term Memory Recurrent Neural Networks for Overlap Detection

3.1. LSTM-RNNs

Recurrent neural networks (RNNs) are a widely used technique for context-sensitive sequence labeling. They exploit context in the form of inputs from past time steps by using cyclic connections. Due to the so-called vanishing gradient problem (the influence of a certain input on the hidden and output layer of the network decays exponentially over time), the context that is used by an RNN is limited. In order to overcome this problems, Long Short-Term Memory RNNs (LSTMs) were introduced in [18]. LSTMs use memory cells to store information over a longer period of time. An LSTM hidden layer is composed of so-called memory blocks. Each memory block con-

sists of multiple self-connected memory cells and three multiplicative gate units (input, output, and forget gates). These gates allow for write, read, and reset operations within a memory block. The amount of context information that the network uses is learned during training. Due to their ability to model long-range dependencies between the inputs, LSTMs seem to be a promising approach for overlap detection.

3.2. LSTM Regression to Generate Overlap Predictions

We apply LSTMs as linear regressors to predict framewise overlap scores. Therefore, the output layer of the network consists of a single linear unit with output $o(t)$ at time t. The input feature vectors to the network are defined as

$$\mathbf{X} = [\mathbf{x_1}, ..., \mathbf{x_T}] \tag{3}$$

where T is the number of frames in the target sequence. In our network, the output $o(t)$ at time t is dependent on the past input vectors $\mathbf{X_t} = [\mathbf{x_1}, ..., \mathbf{x_t}]$,

$$o(t) = f(\mathbf{X_t}) \tag{4}$$

due to the LSTM principle and recurrent nature of the network. During network training, targets are defined as

$$\hat{o}(t) = \begin{cases} 1 & \text{if } \mathbf{x_t} \in \text{overlap} \\ 0 & \text{if } \mathbf{x_t} \in \text{speech} \\ -1 & \text{if } \mathbf{x_t} \in \text{non-speech} \end{cases} \tag{5}$$

which results in a larger distance between nonspeech and overlap. This punishes confusions between non-speech and overlap more than between speech and overlap. The predictions $o(t)$ of the trained network are used for classification by applying a threshold θ,

$$c(t) = \begin{cases} 1 & \text{if } o(t) \geq \theta \\ -1 & \text{if } o(t) < \theta \end{cases} \tag{6}$$

where the predicted class $c(t)$ differentiates only between overlap and non-overlap. The threshold θ is varied to obtain different system operating points with a different trade-off between precision and recall.

The size of the input layer of the network is equivalent to the number of employed audio features. One recurrent hidden layer with four memory blocks is used and each memory block consists of 50 LSTM cells. This topology proved to be efficient for voice activity detection [19]. The LSTM-RNNs are trained and evaluated with the rnnlib by Alex Graves [26]. LSTM training is performed with the backpropagation through time (BPTT) algorithm; the weights are updated using the gradient descent algorithm with a learning rate of 10^{-5} and momentum 0.9. Weights are required to be initialised with non-zero values, thus we initialise the weights with uniform random values sampled from $]0; 0.1]$. To enhance generalisation, Gaussian noise with zero mean and standard deviation of 0.3 is added to all inputs. A maximum of 40 training epochs is run to avoid over-adaptation. We use an early stopping criterion by stopping training if there is no error improvement on the development set for 10 epochs. As an error measure during network training, we use the framewise root mean quadratic error between the targets $\hat{o}(t)$ and the network predictions $o(t)$.

Figure 2 shows LSTM predictions $o(t)$ for a 20-second excerpt from the test set. It can be seen that LSTM predictions are well correlated with the ground truth, yielding low values for non-speech regions and high values for overlap segments. By applying a threshold, overlap segments can be detected from these LSTM predictions.

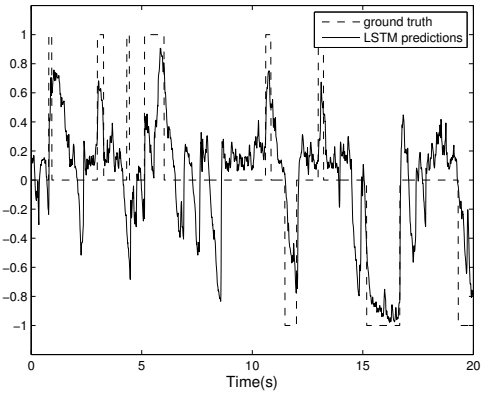

Figure 2: LSTM predictions for a 20-second excerpt from the test set. The dashed line marks the ground truth (-1: nonspeech, 0: speech, 1: overlap).

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

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

To combine LSTM and HMM overlap detection, the LSTM predictions $o(t)$ are used as an additional feature for the HMM. As for all other features, delta coefficients are computed for LSTM predictions. However, LSTM predictions are not normalised, based on results of preliminary experiments.

4. Experiments

4.1. Experimental Setup

Experiments are conducted using the AMI Corpus [27]. A subset of 40 meeting recordings is used for training, 6 meetings for tuning and 10 for testing. We use the same set of 10 recordings (see Table 2) for testing as was used in a previous study by other authors [28]. The length of the recordings in the test set varies between 17 and 57 minutes and in total, the length of the test set is more than 6 hours. All are single-channel, far-field microphone recordings – the most challenging scenario. On average the amount of overlapping speech is in the order of 20 % in the test set.

To obtain the MFCC and ESVC features, we apply the following system parameters, based on our previous experience: Energy, spectral and voicing-related features are computed every 20 ms. A window size of 60 ms is applied for MFCC and voicing-related features, whereas other energy and spectral features are determined using a window size of 25 ms. CNSC is applied using magnitude spectra computed from 40 ms windows with a window shift of 20 ms. We used $R = 35$ bases per speaker, a convolutional range of $P = 4$ and a sparseness parameter $\lambda = 0.05$. The regularisation factor in Eq. (2) is set to $f = 1.2$. Speaker bases are learned using speaker-specific training data obtained with the LIA-Eurecom speaker diarization system [25].

Experiments are performed with the HMM-based and the LSTM overlap detection systems. Both systems are tested with

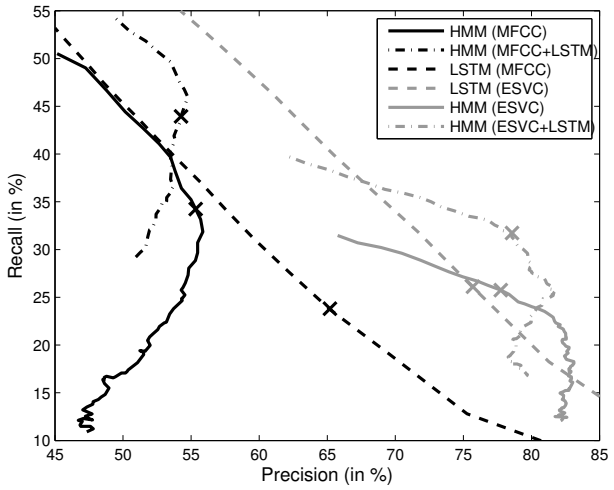

Figure 3: Precision and Recall for HMM (solid lines), LSTM (dashed lines) and their combination (dashed-dotted lines), each time using either MFCC (black) or ESVC (grey) audio features. An 'X' marks the operating point with minimal overlap detection error as determined with the development set.

MFCC features and with ESVC features. In addition, we evaluate the combination of HMM and LSTM, where LSTM predictions are added to the HMM feature set. This tandem system is also tested with MFCC features and with ESVC features. Thus, in sum, 6 different system configurations are tested.

System performance is measured in terms of frame-wise overlap precision, recall and detection error. The overlap detection error is equivalent to the total duration of missed and false positive overlap time divided by the reference overlap time. Note that, since overlap makes up only around 20 % of the recordings, false positive detections can result in an overlap detection error above 100 %. For a typical application such as overlap handling for speaker diarization, the detection error is the most meaningful metric.

4.2. Results

In Figure 3, recall is plotted vs. precision for all six tested systems. In addition, for each system, the operating point with the minimum overlap detection error on the development set is marked with an 'X' and displayed in Table 3. For the LSTM system, different operating points are obtained by varying the threshold for overlap detection, leading to a rather straight curve in the precision-recall plot. Different operating points for the HMM systems are obtained by increasing OIP. At first, increased OIP results in higher precision and lower recall, while for higher OIP, precision drops as well. This can be seen in the result curves for all HMM systems.

With MFCC features, LSTM-based overlap detection performance is comparable to HMM in the high-recall region. Beyond that, LSTMs with MFCC features are capable of achieving high-precision operating points. HMMs achieve a minimum error of $93.4\,\%$ while LSTMs have an error of $88.9\,\%$, which is the result of higher precision. Using ESVC features instead of only MFCCs in the HMM system results in higher precision and lower recall. Yet again, performance of LSTM is comparable to that of HMM with ESVC features, with minimum error rates of $81.6\,\%$ and $82.3\,\%$, respectively. For both feature sets, adding LSTM predictions to the HMM feature set to obtain the tandem system improves the recall performance while achieving similar precision values. In the case of using MFCC features, for the minimum-error operating point, recall increases from $34.2\,\%$ to $44.0\,\%$ while precision stays roughly the same. Due to the comparably low precision, the overlap detection error does not improve. With ESVC features, the system combination increases recall from $25.8\,\%$ to $31.7\,\%$, with precision staying constant. The minimum error decreases from 81.6 % to 76.9 %, which is the consequence of increased recall performance.

Features	System	Prec.	Rec.	Err.
MFCC	HMM	55.3	34.2	93.4
ESVC	HMM	77.8	25.8	81.6
MFCC	LSTM	65.2	23.8	88.9
ESVC	LSTM	75.7	26.1	82.3
MFCC + LSTM pred.	HMM	54.3	44.0	93.1
ESVC + LSTM pred.	HMM	78.6	31.7	76.9

Table 3: Precision (Prec.), recall (Rec.) and overlap detection error (Err.) on the test set for the six tested system and feature combinations. Operating points are tuned to achieve minimum overlap detection error on the development set.

The experimental results show that LSTMs alone perform comparable to HMMs in terms of overlap detection error. The combination of HMM and LSTM can substantially improve the overlap detection performance, due to higher recall. One reason for higher overlap detection recall with the tandem system could be that the ability of LSTMs to exploit long-range context helps to detect overlap segments which are hard to detect by the HMM with acoustic features alone, such as short backchannel utterances.

5. Conclusions

We presented a system for speech overlap detection based on LSTM networks. LSTMs are trained as a regressor to predict a frame-wise overlap score. This prediction score is either directly used to detect overlap by applying a threshold, or it is utilised in a tandem HMM-LSTM system. Experiments were conducted with the AMI corpus and two feature sets: standard MFCCs and a larger set of energy, spectral, voicing and CNSC-based features. LSTMs showed performance comparable to a standard HMM system for both feature sets. Realising the tandem system by adding LSTM predictions to the HMM feature set resulted in improved system performance. Overlap detection recall was improved (23 % relative improvement in the case of using ESVC features) while keeping precision constant. Thereby, the overlap detection error was substantially reduced.

While LSTMs are able to increase the overlap detection recall, there is still a lot of room for improvement. Exploiting information that goes beyond pure acoustical features, like linguistic information, might help to further improve overlap detection.

6. Acknowledgements

This research was supported by the ALIAS project (AAL-2009-2-049) co-funded by the EC, the French ANR and the German BMBF. We would like to thank Nicholas Evans, Ravichander Vipperla and Dong Wang for their contributions to the CNSC features.

7. References

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

[2] K. Laskowski, M. Heldner, and J. Edlund, "On the Dynamics of Overlap in Multi-Party Conversation," in *Proc. Interspeech*, Portland, OR, USA, 2012.

[3] M. Wlodarczak, J. Simko, and P. Wagner, "Temporal entrainment in overlapped speech: Cross-linguistic study," in *Proc. Interspeech*, Portland, OR, USA, 2012.

[4] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.

[5] E. Shriberg, A. Stolcke, and D. Baron, "Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 1359–1362.

[6] F. Brugnara, D. Falavigna, D. Giuliani, and R. Gretter, "Analysis of the Characteristics of Talk-show TV Programs," in *Proc. Interspeech*, Portland, OR, USA, 2012.

[7] H. Sun and B. Ma, "Study of Overlapped Speech Detection for NIST SRE Summed Channel Speaker Recognition," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 2345–2348.

[8] M. Huijbregts, D. van Leeuwen, and C. Wooters, "Speaker Diarization Error Analysis Using Oracle Components," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 393–403, 2012.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] J. Geiger, R. Vipperla, S. Bozonnet, N. Evans, B. Schuller, and G. Rigoll, "Convolutive Non-Negative Sparse Coding and New Features for Speech Overlap Handling in Speaker Diarization," in *Proc. Interspeech*, Portland, OR, USA, 2012.

[14] 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, pp. 340–344.

[15] S. H. Yella and F. Valente, "Speaker Diarization of Overlapping Speech based on Silence Distribution in Meeting Recordings," in *Proc. Interspeech*, Portland, OR, USA, 2012.

[16] S. H. Yella and H. Bourlard, "Improved Overlap Speech Diarization of Meeting Recordings using Long-Term Conversational Features," in *Proc. ICASSP*, Vancouver, Canada, 2013.

[17] Barry Chen, Qifeng Zhu, and Nelson Morgan, "Learning long-term temporal features in lvcsr using neural networks," in *Proc. Interspeech*, Jeju Island, Korea, 2004, pp. 612–615.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life Voice Activity Detection with LSTM Recurrent Neural Networks and an Application to Hollywood Movies," in *Proc. ICASSP*, Vancouver, Canada, 2013.

[20] 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.

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

[22] 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.

[23] D. Wang, R. Vipperla, and N. Evans, "Online pattern learning for non-negative convolutive sparse coding," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 65–68.

[24] D. Wang, R. Vipperla, N. Evans, and T. F. Zheng, "Online non-negative convolutive pattern learning for speech signals," *IEEE Transactions on Signal Processing*, vol. 61, no. 1, pp. 44–56, 2013.

[25] 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.

[26] A. Graves, S. Fernández, and J. Schmidhuber, "Multi-dimensional recurrent neural networks," in *Proc. of the 2007 International Conference on Artificial Neural Networks*, Porto, Portugal, 2007.

[27] 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.

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