LEVERAGING LSTM MODELS FOR OVERLAP DETECTION IN MULTI-PARTY MEETINGS

Neeraj Sajjan, Shobhana Ganesh*, Neeraj Sharma*, Sriram Ganapathy*, Neville Ryant* +

*Learning and Extraction of Acoustic Patterns (LEAP) Lab, Indian Institute of Science, Bangalore-560012
+Linguistic Data Consortium, University of Pennsylvania, USA

ABSTRACT

The detection of overlapping speech segments is of key importance in speech applications involving analysis of multi-party conversations. The detection problem is challenging because overlapping speech segments are typically captured as short speech utterances far-field microphone recordings. In this paper, we propose detection of overlap segments using a neural network architecture consisting of long-short term memory (LSTM) models. The neural network architecture learns the presence of overlap in speech by identifying the spectrotemporal structure of overlapping speech segments. In order to evaluate the model performance, we perform experiments on simulated overlapped speech generated from the TIMIT database, and natural multi-talker conversational speech in the augmented Multi-party Interaction (AMI) meeting corpus. The proposed approach yields improvements over a Gaussian mixture model based overlap detection system. Furthermore, as an application of overlap detection, integration of overlap detection into speaker diarization task is shown to give improvement in diarization error rate.

Index Terms— Overlap Detection, LSTM modeling, Speaker Diarization, Conversational Speech Analysis.

1. INTRODUCTION

Overlap speech segments comprise of speech from more than one talker (as shown in Fig. 1). Such segments are found in almost every multi-talker conversational speech setting, such as dialogues, meetings, debates, and broadcast news. As described in [1], overlapped speech can be demarcated into four types, namely, (a) short feedback, no interruption of the speaker, (b) premature turn-taking at the end of the speakers turn, (c) simultaneous starting after longer silence, and (d) barge-in, aiming to take the turn over.

The presence of overlapped speech segments in recordings adversely impacts speech applications such as automatic speech recognition (ASR), speaker identification and speaker diarization. Presence of overlap segments in the captured recordings, for instance in speaker identification, leads to poor modeling of individual speakers (when used in training), and poor test accuracies (when used in testing) [2, 3]. Accurate detection of overlapping speech segments can significantly improve analysis of conversational speech recordings. For example, Shokouhi et al. [4] report improvement in word-count estimation (in conversational recordings) by providing data pruning using overlap detection, and Charlet et al. [5] report improvement in diarization error rates by excluding overlap speech segments. Over-

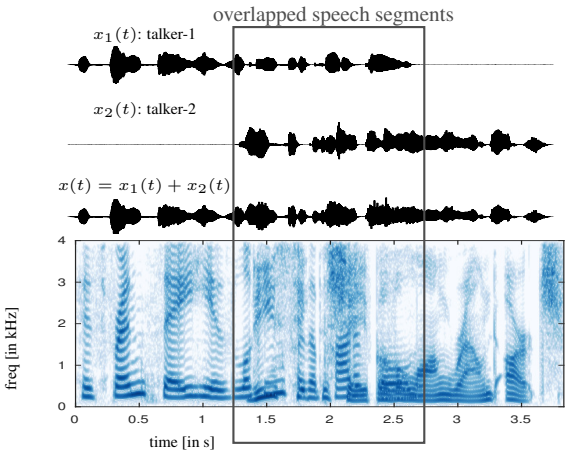

Fig. 1. Illustration of spectrotemporal structure in a synthetically generated overlapped speech. Key distinctions are seen in voiced region overlaps; two different harmonic complexes intermix in these regions.

lap detection can be approached as a feature design and classification problem. From the feature design perspective, assuming overlapping segments contain voicing, harmonicity [6, 7, 8], kurtosis [9, 10] of short-time segments, LPC residual energy [2], spectral-flatness measure (SFM) [6], and MFCCs [6, 2] have been explored as useful features for overlap detection. In spite of these efforts, the performance on overlap speech detection tasks is well below acceptable levels (for example, precision rates of 67% with a recall rate of 34% reported in [6] on TIMIT corpora [11]).

An attempt to model overlapped speech as a non-linear feature transformation in the cepstral domain is proposed in [12], and a convolution sparse coding approach is proposed in [13]. Apart from using short-time segments ($20-30$ ms), Yella et al. [14] has shown improved overlap detection by augmenting silence and speaker change statistics computed over $3-4$ s with short-time features. Interestingly, owing to the four kinds of overlaps in natural conversations [1], the overlapping segments can be associated with overlaps of voiced and unvoiced segments, and also speech and non-speech (such laughter) segments. Further, depending on reverberation, ambient noise, and context of the conversation, the energy of the talkers can be widely different, thereby making single talker to interfering talker ratio vary drastically in the recording. These aspects make the task of robust overlap detection in natural conversational recordings

The work reported here was started at the Jelinek Summer Workshop on Speech and Language Technology (JSALT) 2017 held at CMU, supported by JHU, and gifts from Amazon, Apple, Facebook, Google and Microsoft. The work done was also partly supported by funds from Defense Research Development Organization (DRDO) under the DST0689 project and from the Prathiksha Grant. The views expressed in this paper are from the authors and do not express the views of the funding agencies.

challenging.

We experiment with two time-frequency representations, namely, mel-spectrograms and gammatone spectrograms as features, in addition to the traditionally used features such as SFM, kurtosis and MFCCs. With these features we develop an overlap detection system using deep learning architectures. The methods developed provide a frame-level posterior probability indicating the presence of overlap speech in the given speech segment. We find that the proposed modeling architectures show significant improvements over the baseline performance obtained with Gaussian mixture models. Interestingly, the recurrent modeling methods, with the best performance obtained using long short term memory networks (LSTMs), are able to extract representations from the spectrograms that are useful in predicting overlaps. In a parallel development, on using neural networks for speech analysis, recently, performance benefits for tasks such as speech activity detection [15] and speaker change detection [16] has been obtained.

The main focus of the paper is to improve overlap detection on conversational speech. For evaluation we use the augmented multiparty interaction (AMI) meeting corpus [17]. We also experiment with an artificial overlap dataset constructed with the TIMIT corpus. [11]. The synthetic dataset is used for data augmentation to improve performance on the AMI corpus. As an application of overlap detection on conversation speech analysis, we use the detection system as a front-end pre-processing method in an i-vector based speaker diarization system. We show that integrating the obtained overlap region annotations provides reduction in diarization error rate (DER).

The organization of the paper is as follows. Sec.2 describes the dataset used in the experiments. Sec.3 details the approach used for overlap detection. The results of the various experiments are presented in Sec.4, followed by a summary in Sec. 5.

2. DATASET

2.1. TIMIT

The TIMIT corpus [11] is composed of 630 speakers, each reading 10 phonetically rich English sentences, and recorded in an anechoic environment at 16 kHz. We design an overlap speech dataset using the utterances from the corpus with the following procedure. A pair is created by choosing two recordings taken from two different speakers, and these recordings are superimposed together. The instant of superimposition is chosen randomly with the constraint that the instant is prior to the end of the last spoken word in the first recording. The phoneme level annotations of each recording in the superimposition are used to obtain the ground truth overlap labels. To simulate real life scenarios, the superimposed recordings are convolved with the room impulse responses (RIR) of meeting rooms (drawn from the publicly available Aachen RIR dataset [18] and corrupted at 3-10 dB SNRs levels with 4 kinds of additive noises (meeting, lecture, conference, and hallway; drawn from the DEMAND database [19]). Training is done on speech files corresponding to 400 speakers and testing on separate 168 speakers. The partitioning is as given in the dataset.

2.2. AMI

The AMI corpus [17] is a meeting dataset containing close to 100 hrs of multi-talker conversational meeting recordings. A meeting has at least three talkers, and there are a total of 171 talkers in the whole corpus (114 male and 57 female). Each meeting is recorded using a set of different devices, namely, microphone array composed of eight single distant microphones, headset and lapel microphones. We use the first microphone channel (denoted by `Array-1`) from the microphone array. In addition, the officially designated training dataset and validation dataset in the AMI corpus was used for training and testing, respectively, in the experiments. The training and testing dataset comprises of 102 and 12 meeting recordings, respectively. The microphone array channel are distant recordings, and hence contain reverberation and ambient noise, in addition to the naturally occurring overlapped speech regions. The AMI corpus is annotated by human annotators using the headset recordings (as these have higher single talker SNR). On analysis of the annotations, 7.9% frames (at 10 ms) of the total spoken speech frames are composed of overlapped speech. The rest of the recording is either single talker speech or unlabeled speech/silence. However, on cross-checking, it was found that some of the annotations have errors. For example, several regions containing non-speech sounds (like laughter) are labeled as speech, and several unlabeled regions include single, multi-talker speech and large amounts of silence. To rectify the annotation errors, we do force alignment using a pre-trained automatic speech recognition (ASR) model. The model was pre-trained on single distant microphone (SDM) files of AMI dataset using the Kaldi ASR toolkit [20]. The forced alignment resulted in changes in the annotations at certain segments of the recordings. This had a significant impact on the ground truth labeling of single and overlapped speech segments. For example, close to 40% of the frames were assigned different labels after force alignment.

3. DETECTING OVERLAP SEGMENTS

3.1. Baseline Approach

As a baseline we use Gaussian mixture models (GMMs). The features used in training include kurtosis, SFM, and MFCC+D, generating a 26 dimensional feature vector per frame. This is similar to the setup used in [6, 14] (the aperiodicity measure is omitted as its computation is not well defined for natural recordings containing ambient noise and reverberation. Further, this measure was not found to be of much significance in [14]). The features are computed for every 25 ms window, with a window hop size of 10 ms. The performance of baseline approach is reported in Table 1.

3.2. Neural Network Approach

The substandard performance of baseline, shown in Table 1, can be attributed to the choice of features and the use of GMMs. To validate this hypothesis, we harness the modeling capabilities of neural networks. We explore the performance with few variants of these networks. A brief description of the features and the neural network models used is presented below.

Features. Instead of extracting features from the short-time segments, we use the spectrogram itself as feature. We use the mel-spectrogram, a frequency warped time-frequency representation of speech. This is obtained by binning the short-time (25 ms, with hop size of 10 ms) segment magnitude Fourier transform of speech with non-linearly arranged triangular frequency filters. The logarithm of the weighted sum of spectral energy, in each bin, gives the mel-spectrogram representation of 40 dimensions. We denote this feature by fbank. A feature vector corresponds to context of 11 frames, 5 frames on either side of the current frame.

Deep neural network (DNN). A feed forward DNN with 3 hidden layers and each layer containing 256 neurons is used. Rectified Linear Units (ReLU) are used as activations. The input feature vec-

Table 1. Detection accuracy % with GMMs (Baseline approach) [4].

Data	Feature	Single	Overlap	Avg.
TIMIT	kurt.+SFM+MFCC+D	59.6	69.4	64.5
AMI	kurt.+SFM+MFCC+D	43.1	61.9	52.5

Table 2. Detection accuracy % on TIMIT Overlap dataset with NNs on fbank features (Proposed approach)

Model	Single	Overlap	Avg.
DNN[3 layers]	73.0	87.0	**79.9**
CNN2D [3 layers]	79.2	71.9	75.5
lstm[512 cells]	73.7	83.1	78.4
blstm[256 cells]	78.7	79.5	78.9
blstm[512 cells]	72.5	87.0	79.7
clstm[1 Conv-lstm(512)]	89.8	52.0	71.8
clstm[3 Conv-lstm(512)]	87.0	63.0	74.9

tor is a vectorized version of each context patch. As a comparison with

Convolutional network models (CNN) The CNN contains convolutional layers which are weight shared feature extraction layers. CNN model used in experiments has 3 convolutional layers, with $\{64,\ 128,\ 256\}$ filters and kernel shapes of $(3, 7)$, $(3, 5)$ and $(3, 3)$. Finally a pooling layer was added, the output of which was fed into three dense layers with 1024, 512, and 256 neurons, respectively.

Long short term memory (LSTM) network models The LSTMs explicitly make use of evolving temporal structure in the features, and hence have been found suitable for speech recognition applications [21]. We use a single layer of 512 LSTM cells which are followed by three dense layers with 1024, 512 and 256 neurons. The bi-directional LSTMs which process both past and future information by looking at the input from both time directions are also explored. We used 512 blstm units with 3 dense layers similar to network with lstms. Finally, we also experiment with two variants of convolutional LSTM network, which perform front-end processing with convolutional model, followed by a recurrent architecture [22].

The first variant comprised of one convolutional layer, the output of which was fed into the LSTM. The second had 3 layers of convolution with same number of filters and neurons as mentioned before, and was fed after pooling into the LSTM layer comprising 512 units followed by 3 dense layers.

4. RESULTS

Training of the models was posed as a three class classification problem, with the classes being single, overlap, and filler, using the Keras toolkit [23]. The filler class models any extraneous sounds that are neither single speaker nor overlapped speech segments. The data for the filler class is derived from the non-speech regions of the dataset. Using the filler class obviates the requirement for an accurate voice activity detector (VAD, designing an accurate VAD for natural noisy speech recordings is challenging). The performance of the trained models are reported on test/evaluation data at frame level accuracies of single speaker class and overlap class. In the AMI dataset, even after forced alignment, a significant number of segments are not annotated. As it is not possible to identify false alarms in such cases, we do not opt for reporting Equal Error Rate (EER). We also report the performances on only the single speaker class and overlap class due to the above mentioned issues in the dataset.

The baseline performance using GMMs is shown in Table 1.

Table 3. Detection accuracy % on AMI Meeting Dataset with original annotations, data augmentation, on fbank features.

Model	Single	Overlap	Avg.
DNN[3 layers]	56.3	73.0	64.7
lstm[512 cells]	76.0	60.6	**68.4**
blstm[256 cells]	51.4	75.3	63.4
blstm[512 cells]	58.3	71.8	65.1
clstm[1 Conv-lstm(512)]	49.5	74.5	62.0
clstm[3 Conv-lstm(512)]	57.8	68.0	63.0

Table 4. Detection accuracy % on AMI Meeting Dataset with force aligned annotations using fbank features.

Model	Single	Overlap	Avg.
DNN[3 layers]	63.9	78.0	70.9
CNN2D[3 layers]	73.0	63.8	68.4
lstm[512 cells]	77.0	68.0	**72.5**
blstm[256 cells]	68.9	75.4	72.1
blstm[512 cells]	57.8	79.0	68.4
clstm[1 Conv-lstm(512)]	36.3	87.4	61.8
clstm[3 Conv-lstm(512)]	39.3	87.2	63.2
lstm[512 cells][without data aug]	66.37	69.23	67.8
lstm + Viterbi decode	**87.9**	**71.0**	**79.4**

Each of the three classes were modeled using 512 component GMMs. The train and test dataset contain TIMIT segments without any additional noise and reverberation effects. The filler class models silence. The GMM based approach performs reasonably well on this dataset. As expected, the performance degrades on the AMI database. For comparison, a 3 layer DNN with the same features as that of the baseline was trained. This gave an average accuracy of 68.3%, a 30% improvement over GMM approach. Basing on this improvement, the following experiments focused on using neural network architectures with contextual features.

- **TIMIT -** The results for overlap detection on the TIMIT dataset using neural network approach are reported in 2. As seen here, the CNN and LSTM models models' results are significantly better than that of the baseline GMM results. The best result is obtained with the 3 layer DNN model, although the bidirectional LSTM model is similar in performance.
- **AMI -** In the first set of experiments with AMI dataset we used original AMI annotations. Data augmentation is used to have a more enriched training for overlap detection on the AMI dataset. For this, we use the overlap dataset created using TIMIT corpus, together with samples containing additive noise and reverberation distortions (as discussed in Sec. 2). The results are shown in Table 3.

  Next, we considered forced aligned annotations as these rectify erroneous labeling of training and evaluation data. The model architectures and features used on the data without force alignment was used to evaluate the data with forced aligned labels to outline the improvement in performance (reported in Table 4). The GMM results for the AMI dataset reported in 1 also uses forced aligned annotations. As seen here, forced alignment provides less noise data labels and improves the system performance.

  Further improvements on AMI dataset is achieved using posterior Viterbi decoding [24] with a 3 state HMM (filler/single

Table 5. Detection accuracy % on AMI Meeting Dataset using different features with the LSTM model.

Feature	Single	Overlap	Avg.
gammatone	66.3	75.1	70.7
gammatone + kurt.+SFM	67.9	73.5	70.7
fbank + kurt.+SFM	79.1	62.3	70.7

speaker/overlapped speech). Parameters were estimated using the forced alignment labels on the training set and each recording was decoded in a single pass. As seen in Table 4, Viterbi smoothing improved the accuracy substantially, increasing frame-level accuracy to 87.9% for single and 70.9% overlapped speech, respectively.

In addition to fbank features, we also experimented with gammatone spectrogram [25]. In comparison to the mel-spectrogram, gammatone spectrogram provides a more accurate modeling of the peripheral auditory filter bank [25]. We used 64 gammatone cochlear filters in our implementation. The results are tabulated in Table 5. The overlap class detection accuracy is better than the single class. This is in contrast to the results obtained using fbank features. Interestingly, using kurtosis and SFM features with fbank or gammatone spectrogram does not provide any noticeable improvements. We hypothesize that improved modeling capabilities elicited by LSTM models may subsume any additional advantages using spectral features like SFM, and kurtosis.

4.1. Impact on Speaker Diarization

To evaluate the impact of overlap detection on speaker diarization, we implement a state-of-the-art i-vector based speaker diarization system [26] based on i-vector model [27] (trained on fixed 1.5 s speech segments, with shift of 0.75 s). The pairwise distance metric computed using probabilistic linear discriminant analysis (PLDA) is used. An agglomerative clustering procedure is applied on the PLDA scores to obtain clusters associated with each speaker. The performance of the speaker diarization system is measured using the diarization error rate (DER) (with collar of 250 ms and discarding multi-talker segments in scoring). The diarization system was implemented using the Kaldi toolkit [20].

The speaker diarization result on a development set composed of 12 meetings and an evaluation set of 16 meetings is reported in Table 6. Here, we experiment with three conditions: (i) use of the force-aligned AMI segments without discarding overlaps (serves as baseline), (ii) pruning the baseline labels by keeping single speakers intact (based on reference) and using the predicted LSTM output (on the overlap regions), and (iii) pruning the baseline labels by keeping single talker intact and using the ground truth labels for the overlap regions. This way of evaluation ensured that all the methods had no contribution of speech detection error in DER computation. The main contribution of the overlap detector is to provide regions of single speaker speech which can generate pure talker segments for agglomerative clustering. As seen in Table 6, the proposed overlap detection method improves the state-of-art diarization system considerably (relative improvements of 22.1% over the baseline system for development meetings and 21.0% for the evaluation meetings).

5. CONCLUSION

The findings from the experiments on both the TIMIT and AMI datasets suggest that the LSTM based models provide better separation between single speaker and overlap classes. In order to understand the working of the LSTM model on this task, we have provided a visualization of the input fbank representation as well as the LSTM outputs (the first two dimensions of t-stochastic neighborhood embedding (t-SNE) [28]) in Fig. 2. This is plotted for a meeting file from the AMI dataset not used in training. The LSTM representations show significant separation in the single/overlap classes compared to the input fbank representation. Hence, the recurrent network is able to extract representations useful for overlap detection from the fbank input.

In summary, we have developed an overlap detection based on neural network models. The previous work on overlap detection used GMMs with a variety of features. In this work, we have found that: (i) recurrent network models like LSTM are quite effective for overlap detection, and (ii) instead of designing specialized features, spectrograms can be used as features for this task. With experiments on the artificial overlap data on TIMIT, we have illustrated that the proposed approach performs significantly better than the baseline methods. In addition, using the AMI meeting dataset, we show that the usefulness of the proposed system extends to natural conversational settings, as well, and improvements in a speaker diarization application[1]. The AMI meeting dataset has multichannel recordings. In future, we plan to build a multi-channel framework to exploit these recordings, and analyze the implications on overlap detection performance and robust conversational speech analysis.

Table 6. Diarization error rate (DER) % for development and evaluation meetings in the AMI corpus obtained without any pruning of overlap segments (baseline), pruning using proposed LSTM output (predicted), and pruning using ground truth labels.

Model	Dev.	Eval
Baseline	25.3	19.4
Predicted	19.7	15.2
Ground truth	19.3	16.7

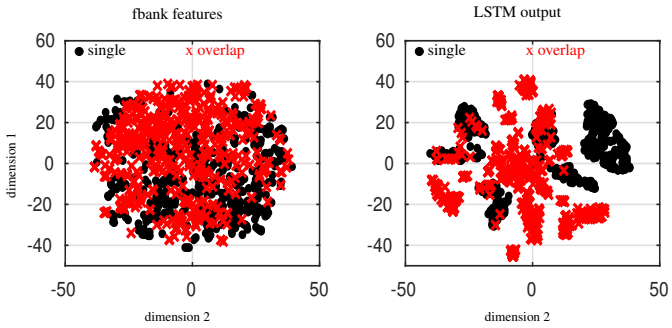

Fig. 2. t-SNE [28] scatter plots of input fbank features with context (11×64) and the LSTM 1^{st} layer activation, for single speaker and overlap frames.

6. ACKNOWLEDGEMENTS

The authors would like to thank the participants of the conversational analysis team in this workshop, and especially, Matthew Maciejewski for his help in setting up the diarization system.

[1]The forced aligned AMI dataset labels and the codes used in the paper can be accessed at `https://github.com/BornInWater/Overlap-Detection`

7. REFERENCES

[1] Ingo Siegert, Ronald Bock, Andreas Wendemuth, Bogdan Vlasenko, and Kerstin Ohnemus, "Overlapping speech, utterance duration and affective content in HHI and HCI- an comparison," in *IEEE Intl. Conf. on Cognitive Infocommunications (CogInfoCom), 2015*. IEEE, 2015, pp. 83–88.

[2] Kofi Boakye, Beatriz Trueba-Hornero, Oriol Vinyals, and Gerald Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.* IEEE, 2008, pp. 4353–4356.

[3] Martin Zelenák and Javier Hernando, "The detection of overlapping speech with prosodic features for speaker diarization," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[4] N. Shokouhi, A. Ziaei, A. Sangwan, and J. H. L. Hansen, "Robust overlapped speech detection and its application in word-count estimation for prof-life-log data," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, April 2015, pp. 4724–4728.

[5] D. Charlet, C. Barras, and J. S. Liénard, "Impact of overlapping speech detection on speaker diarization for broadcast news and debates," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, May 2013, pp. 7707–7711.

[6] Navid Shokouhi, Amardeep Sathyanarayana, Seyed Omid Sadjadi, and John HL Hansen, "Overlapped-speech detection with applications to driver assessment for in-vehicle active safety systems," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.* IEEE, 2013, pp. 2834–2838.

[7] Yang Shao and DeLiang Wang, "Co-channel speaker identification using usable speech extraction based on multi-pitch tracking," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2003, vol. 2, pp. 205–8.

[8] N. Shokouhi and J. H. L. Hansen, "Teager-kaiser energy operators for overlapped speech detection," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1035–1047, May 2017.

[9] J. P. LeBlanc and P. L. De Leon, "Speech separation by kurtosis maximization," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 1998, vol. 2, pp. 1029–1032.

[10] K. R. Krishnamachari, R. E. Yantorno, J. M. Lovekin, D. S. Benincasa, and S. J. Wenndt, "Use of local kurtosis measure for spotting usable speech segments in co-channel speech," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2001, vol. 1, pp. 649–652.

[11] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[12] Pranay Dighe, Marc Ferras, and Hervé Bourlard, "Detecting and labeling speakers on overlapping speech using vector taylor series," in *Proc. INTERSPEECH, ISCA*, 2014.

[13] J. T. Geiger, R. Vipperla, N. Evans, B. Schuller, and G. Rigoll, "Speech overlap detection using convolutive non-negative sparse coding: New improvements and insights," in *2012 Proc. European Signal Processing Conference (EUSIPCO)*, Aug 2012, pp. 340–344.

[14] Sree Harsha Yella and Hervé Bourlard, "Overlapping speech detection using long-term conversational features for speaker diarization in meeting room conversations," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1688–1700, 2014.

[15] G. Gelly and J. L. Gauvain, "Optimization of RNN-based speech activity detection," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 646–656, March 2018.

[16] Ruiqing Yin, Hervé Bredin, and Claude Barras, "Speaker change detection in broadcast tv using bidirectional long short-term memory networks," in *Proc. INTERSPEECH, ISCA*, 2017, pp. 3827–3831.

[17] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al., "The ami meeting corpus: A pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.

[18] Marco Jeub, Magnus Schafer, and Peter Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Intl. Conf. on Digital Signal Processing, 2009*. IEEE, 2009, pp. 1–5.

[19] Cassia Valentini-Botinhao et al., "Noisy speech database for training speech enhancement algorithms and tts models," 2017.

[20] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE ASRU*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[21] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.* IEEE, 2013, pp. 6645–6649.

[22] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[23] François Chollet et al., "Keras: Deep learning library for theano and tensorflow," *URL: https://keras. io/k*, 2015.

[24] Piero Fariselli, Pier Luigi Martelli, and Rita Casadio, "A new decoding algorithm for hidden markov models improves the prediction of the topology of all-beta membrane proteins," *BMC Bioinformatics*, vol. 6, no. 4, pp. S12, 2005.

[25] D. P. W. Ellis, "Gammatone-like spectrograms," Last accessed: Oct. 2017.

[26] Gregory Sell and Daniel Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *IEEE Spoken Language Technology Workshop (SLT), 2014*. IEEE, 2014, pp. 413–417.

[27] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[28] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.