Audio Engineering Society

Conference Paper

Presented at the Conference on
Semantic Audio
2017 June 22 – 24, Erlangen, Germany

Enhancing LSTM RNN-based Speech Overlap Detection by Artificially Mixed Data

Gerhard Hagerer[1,2], Vedhas Pandit[2], Florian Eyben[1], and Bjorn Schuller[1,2]

[1] *audEERING GmbH, Gilching, Germany*
[2] *Chair of Complex & Intelligent Systems, University of Passau, Germany*

Correspondence should be addressed to Gerhard Hagerer (gh@audeering.com)

ABSTRACT

This paper presents a new method for Long Short-Term Memory Recurrent Neural Network (LSTM) based speech overlap detection. To this end, speech overlap data is created artificially by mixing large amounts of speech utterances. Our elaborate training strategies and presented network structures demonstrate performance surpassing the considered state-of-the-art overlap detectors. Thereby we target the full ternary task of non-speech, speech, and overlap detection. Furthermore, speakers' gender is recognised, as the first successful combination of this kind within one model.

Introduction

Today speaker diarisation systems are at a stage where it is the speech overlap that contributes to the majority of the errors [1, 2]. The same can be said about the speech recognition systems. Due to its potential in improving speaker diarization and speech recognition performance [1, 3], there is now a growing interest in the overlap detection problem, and also in 'overlap-robust' systems such as [4]. Unsurprisingly, overlaps are at the active focus of NIST Rich Transcription (RT) evaluations since 2004[1].

In addition to being responsible for performance degradation, speech overlaps are also important from a pervasiveness point of view. Overlaps may represent as high as 40% of all between-speaker intervals in conversations [5]. Overlap additionally carries a lot of information regarding the conversational dynamics, e.g., interrelationships, the dominance/subordination of a speaker with respect to the others [6], the speaker's competitive versus non-competitive intentions [7, 8], agreement levels [9], extent of co-operation [10], gender roles [6, 11]. Modelling of such connotative aspects of turn-taking and overlaps is crucial in designing a convincingly realistic virtual agent or a dialogue system. Developments in this area therefore has a much wider impact with direct implications for the human computer interaction community as well [12].

The problem of overlap detection so far was mostly analysed by techniques like Gaussian Mixture Models or Hidden Markov Models [1, 3], which could be improved through utilization of prosodic [13] and spatial [14] features. Furthermore, impressive results could be achieved by decomposing a signal into its underlying contributory parts via convolutive non-negative sparse coding (CNSC) [15, 16]. In recent research, neural

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

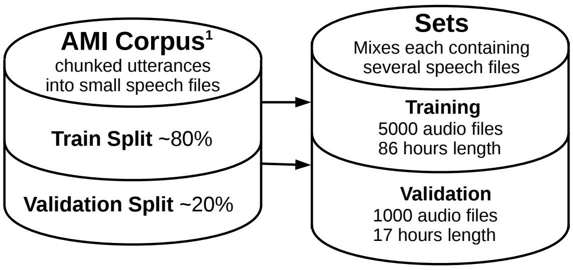

Fig. 1: Preparation of training and validation sets: AMI corpus recordings except test set plus recordings from some emotion corpora are cut into small utterance files. Finally, artificial mixes are created each containing several non-overlapping and overlapping segments.

network based systems arose which learn contextual information occurring around speech overlap [17]. This paper builds upon that work on overlap detection based on LSTM RNNs, but in contrast thereto, we argue here that artificially mixed instead of naturally occurring speech is sufficient to train LSTM RNN overlap detectors. Overlaps are known to be correlated with syllable boundaries [18], beginning of utterances [19], linguistic cues [2], the distribution of speech pauses [20], or speaker change statistics [21]. However, when the training data is artificially mixed, this context information is lost completely and the neural net has to focus merely on the overlap signal itself. The present works argues that this approach, i.e. speech overlap detection by ignoring sociolinguistic context, brings better measurements. Furthermore, mixing offers the possibility to generate as much unique data as necessary due to the randomisation involved in the mixing process. For machine learning approaches in general, and for neural networks in particular, this advantage is highly advantageous.

The rest of this paper is organised as follows. Section 2 contains the methodology, i. e. generation of the artificially mixed datasets, extracted features, and LSTM configurations. In Section 3, the experiments are described including LSTM training and performance improvement strategy thereof, and the evaluation of its segmentation performance. Finally, we conclude with our findings in Section 4.

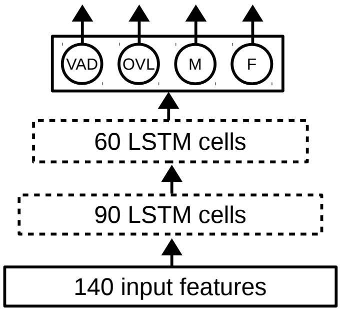

Fig. 2: Illustration of the LSTM structure: The inputs are audio feature vectors. Hidden layers are dashed. All layers are fully connected and LSTM layers with themselves recursively. The topmost feedforward output layer predicts voice activity, overlap, male, female.

Methodology

Data Generation

We use the complete AMI meeting corpus [22] without the test set as defined by Geiger [17], which is kept aside for evaluation tests later on. Furthermore, speech utterances from emotional corpora are added to cover tonal and loudness-related variations in speech [1]. The used recordings are cut into small utterance files based on speech pauses greater than 1.5 seconds. These cuts are performed on the recordings of the head-mounted microphones, where only one speaker is audible, instead of the complete meeting recordings that include speech activity from all the participants. In this way, naturally occurring overlaps are avoided completely. This is important, because the aim is to generate overlap completely artificially, by laying several utterances on top of each other through audio mixing. But before doing so, a train (80%) and validation (20%) split of the utterance files is created such that each speaker occurs only in one of the two sets. This is shown on the left of Figure 1. Mixing is then performed on the files of each split, by randomly choosing several utterances as input files (drawing with replacement) and randomly laying them either on top of or sequentially behind each other within a new mono audio signal. During this process, pitches, speeds and volumes of the input speech

[1] Emotion corpora being added to AMI are EmoDB, AVIC, FAU AIBO, and Semaine. The amount of emotional data is about one third compared to the AMI dataset.

files are varied and finally also the noise is added. In that context, normalization is applied to avoid clipping. Furthermore, speech pauses between utterances are randomised in length and gender occurrences are balanced (50% / 50%). We use the released annotations i. e., the pause and voice activity information, speaker identity with gender metadata, for this purpose. As a result, a training and validation set is created as shown in Figure 1.

Audio Features

From the mixed audio files the following acoustical features are extracted. First, 50 magnitudes are calculated based on 50 equidistant, triangular band pass filters that are applied on the power spectrum of a Mel scale. Secondly, MFCCs 1–20 are calculated based on these magnitudes. Thirdly, for each of these 70 features, the delta is calculated using the Delta Regression coefficient with a time radius of 2, i.e. 5 frames are considered [23]. In total, this results in 140 features per frame, where each frame has a length of 55 ms and the step-size between two consecutive frames is 20 ms.

Neural Network Architecture

LSTM RNNs

The problem of overlap detection requires audio classification over time, which in turn is in the domain of sequence labelling. Recurrent Neural Networks (RNNs) are suitable for this task, since their outcome of the current timestep depends not only on the current, but also on the previous inputs, i.e. they are context-sensitive. However, since standard RNNs are suffering from the vanishing gradient problem, their context sensitivity degrades rapidly with time.

As a consequence, Long Short Term Memory RNNs (LSTMs) were established by Hochreiter and Schmidhuber in 1997 [24], which became popular due to impressive recognition successes since at least 2007 [25]. LSTMs are Recurrent Neural Nets with fully connected recursive layers. Instead of sigmoidal neurons, the so-called LSTM blocks are used – consisting of input, forget and output gates. Each of these gates are sigmoidal units, and together they control the access to the cell's state memory, i. e., how much of the new input (input gate), and of the old state (forget gate) gets written to the current state, and how much of it gets emitted (output gate) by learning the weights associated. Their activations are computed similarly to what is already known of RNNs, except that the last internal state (in certain ways therefore, the memory) is fed into each gate, too. As a result, these cells can hold contextual information, like characteristics of a speech in a scene, for many time steps. This makes them particularly suitable for detecting overlap.

Outputs

For this paper, bidirectional LSTMs (BLSTMs) are incorporated, which tend to give better results when using the same number of cells. These work just as LSTMs, but half of the network is used to analyse the sequence backwards, while the other half is used to analyse the sequence in forward direction. The two results are then merged. This usually gives better results, since not only the past, but also the future context is considered by the RNN [26]. However, BLSTMs can only be used for offline batch analysis, since the sequence to be analysed must be complete, so that the backward run can be performed. This is the case for our experiments.

Regarding the structure of the utilised BLSTM RNN, explanations are given along the lines of Geiger [17] as follows. A sequence of feature vectors representing the audio signal is mapped to four regression outputs corresponding to the degree of voice activity, overlap, male, and female, by the network.

The input sequence of total length T is defined as

$$\mathrm{X_T} = [\mathrm{x_1}, ..., \mathrm{x_T}] \tag{1}$$

The output of RNNs depends both on the current input $\mathrm{x_t}$ at timestep t, as well as on all previous inputs $\mathrm{x_k}$ at timesteps $k < t$. The output is, therefore, a function of the sequence $\mathrm{X_t} = [\mathrm{x_1}, ..., \mathrm{x_t}]$ as follows:

$$o(t) = f(\mathrm{X}_t) \tag{2}$$

The targets for $o_i(t)$ are learned from the training set as

$$\hat{o}_i(t) = \begin{cases} +1 & \text{if output } i \text{ is active} \\ -1 & \text{otherwise} \end{cases} \tag{3}$$

where i corresponds to the respective output: general speech activity, speech overlap, male speech activity, and female speech activity.

The predictive regression output of $o_i(t)$ is classified by applying a threshold θ to it:

$$c_i(t) = \begin{cases} +1 & \text{if } o_i(t) \geq \theta \\ -1 & \text{otherwise} \end{cases} \tag{4}$$

Structure & Parameters

The network has two hidden layers, of which the first contains 90 and the second contains 60 fully connected hidden LSTM cells. The output is a feedforward layer and has four neurons corresponding to the voice activity detection (VAD), the overlap detection (OVL), the male speaker detection (M), and the female speaker detection (F) – see Figure 2. The hyperbolic tangent (also referred to as *tanh*) is used as an activation function . Backpropagation is implemented using the backpropagation through time algorithm (BPTT). Stochastic gradient descent is utilised, i. e., weight updates are applied by steepest gradient descent after each mini-batch of the training set. The learning rate is set to 10^{-5} and the momentum to 0.9. A weight noise, with standard deviation 0.01 and mean 0, is added before every weight update during backpropagation to avoid getting stuck in local minima. Since the used inputs contain noise already, no input noise is added to the input vectors during training. Unless mentioned otherwise, the sum squared error loss function (SSE) is used.

Experiments

Model Training

To train the BLSTM model described in Section 2.3, the training and validation sets are created by mixing as described in Section 2.1. The GPU-accelerated, LSTM capable, CURRENNT toolkit [27] is used to implement several different network architectures, including the one presented in this paper. Training is stopped after 20 epochs without accuracy improvement on the validation set to avoid overfitting. The resulting model is referred to as the standard BLSTM.

Two problems are not handled by the standard BLSTM. Firstly, the amount of overlap in natural speech tends to be relatively low as compared to non-overlapped speech, which is why the mixed datasets were created correspondingly. Secondly, the active speech, male, and female targets are over-represented compared to overlap. The corresponding errors therefore contribute much more to the weight updates during backpropagation than the errors due to missed overlaps. This is likely to create a high number of false positives, and thus low recall measures for overlap frames.

To tackle this, two modified training methods are proposed: post-training and weight penalisation. For post-training, a new training set called overlap set is generated based on the same principles as the actual training set with the only difference that it contains much more speech overlap as you would expect from normal conversational scenarios (80%). The previously trained standard BLSTM network is then taken as a basis for a new training which is done on the overlap dataset. By this training procedure the network is given a stronger focus on recognising speech overlap characteristics in the post-training, while it is at the same time robustly modelling non-overlapped speech characteristics from pre-training. In the best case it might be able to react better to overlap. In the worst case it just gets a stronger bias towards overlap recognition independently of if speech overlap is there in the input data or not. The resulting network is referred to as post-trained network.

For error penalisation, the overlap frames are weighted by penalising the corresponding target errors more than the other ones, so that they contribute more to the weight updates. This is done with a weighted SSE loss function where x is the input vector, i is the index for the output, M is the total number of outputs, o is the predictive output vector of the net, $\hat{o}$ the target vector, and p_i is the weight applied on the error of output o_i:

$$O_{\text{SSE}}(o_i, \hat{o}_i) = \frac{1}{2} \sum_{i=0}^{M-1} p_i (\hat{o}_i - o_i)^2 \tag{5}$$

The question here is what the best possible overlap penalisation weights are. An answer can be found by trying out several weights. The BLSTM is trained regularly on the training set in the same manner as the standard BLSTM, but each time with a different weighted SSE loss function. Afterwards, it is possible to choose the best among all resulting networks. As a reference measure, the highest F_1 score is taken into account, based on the frame-wise network predictions for all uncut AMI meeting conversations – except the ones from the test set.

The the overlap penalties are chosen differently depending on three cases: no speech, speech and no overlap,

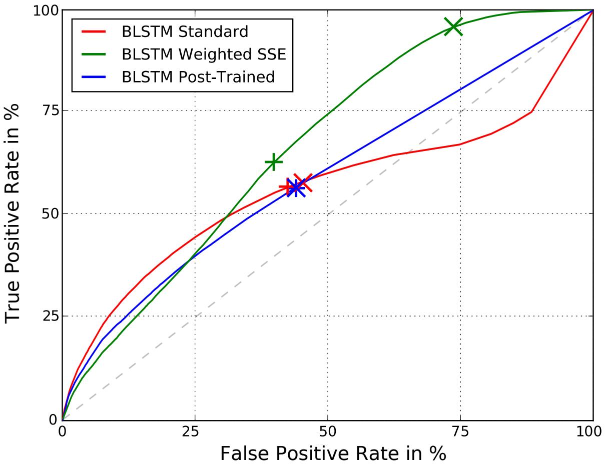

Fig. 3: Receiver Operator Characteristics of all BLSTM models for speech overlap: The frame-wise false positives are plotted against the true positives in %. The 'X' signs mark the optimal overlap detection error (ODE), '+' the equal error rate (EER) for each model.

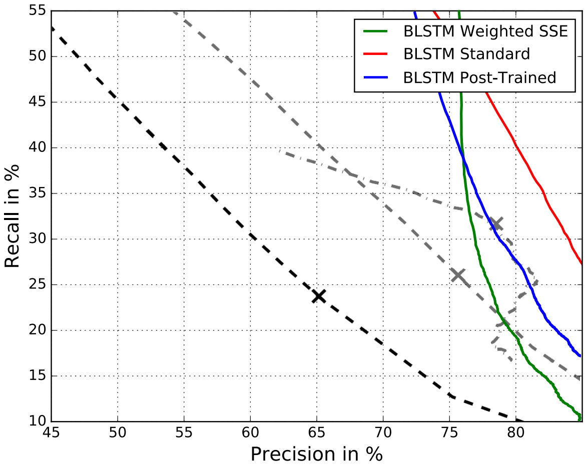

Fig. 4: The frame-wise overlap performances precision and recall of the proposed BLSTM models: standard and weighted SSE. The greyish dashed lines represent selected models of [17] on the same test set. 'X' symbols mark the point with the minimal overlap detection error (ODE).

speech and overlap. The first case needs the least penalisation and stays as it is, the middle one takes a bit more to avoid false positives of the trained model, the last gets the highest penalisation weighting. For all other outputs, the penalties are left at their default value 1.

By this method the focus of the speech overlap output of the neural network is precisely controlled so that it tries to optimise at most on the overlapping speech regions.

Results

We apply the three previously explained models frame-wise to the test set are defined in [17]. The precision and recall measures for each model are plotted in Figure 4, whereby the curves are created by shifting the regression threshold θ as defined in Section 2.3.2 from -1 to $+1$ in 0.01 steps for the predictive overlap output. In the figure, the three models proposed in this work are marked by colours, whereas the black and grey lines depict performance of comparable overlap detection models presented in Geiger [17].

It can clearly be seen in Figure 4 that overlap detection could be improved compared to the approach from Geiger [17]. The colorised line plots representing our BLSTM approaches mostly lie above all the others implying that the precision and recall value pairs are both higher than the ones from the models from related work. However, it turns out that the standard BLSTM approach is not better than weighted SSE, even though it appears so from Figure 4. This impression comes from the axes' limitation to have a comparable view to Geiger. A global picture would show the superiority of weighted SSE against all others, which is also underpinned by the following findings.

In Table 3 the minimal overlap detection error (ODE), i. e. the sum of false positives and negatives divided by the total number of frames, for each model is shown. Here, it can be highlighted that the 27.32% ODE of the weighted SSE approach is the lowest both among the models from this as well as from Geiger's work. The other approaches perform a bit worse, the standard BLSTM approach with 42.88% ODE and the post-trained with 43.39%.

Concerning receiver operator characteristics, Figure 3 shows how frame-wise false positives and true positives are correlated. The 'X' signs show the minimal ODEs, which maximise the relation of true to false positives.

	Standard BLSTM		Weighted SSE BLSTM		Post-Trained BLSTM	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
Gender	90.99	89.63	90.63	89.34	90.07	88.78
VAD	88.31	92.95	87.94	92.83	89.05	93.55

Table 1: Frame-wise accuracies and F1 scores from gender and VAD outputs of both proposed models applied on the test set in %. The measures are taken according to the threshold from minimal ODE, see Figure 4. VAD performances are more precise than apparent here, since the models recognise small speech pauses < 1 second between words which is not the case for the AMI speech segment annotations. Measurements on more accurately hand-labelled datasets yield accuracies up to 96% and F1 scores up to 98%.

Model	AUC	EER
BLSTM weighted SSE	66.69	38.48
BLSTM standard	56.36	42.89
BLSTM post-trained	59.33	43.72

Table 2: The values of area under the curve (AUC) and equal error rate (EER) for speech overlap in % for both models as they are plotted in Figure 3. While the AUC for weighted SSE approach is greater, the EER is nearly the same.

Model	Prec.	Recall	ODE
BLSTM weighted SSE	72.29	95.92	27.32
BLSTM standard	71.10	37.83	42.88
BLSTM post-trained	72.1	57.12	43.39

Table 3: Frame-wise speech overlap precision, recall, and minimal ODE (overlap detection error) of all models as they were applied on the test set in %. Measurement numbers correspond to the 'X' from Figure 4.

In that regard the weighted SSE approach appears to have a significant better performance as the standard BLSTM as well as the post-trained approach. This is also reflected by a greater area under the curve (AUC) – see Table 2 – and by a smaller equal error rate of the weighted SSE approach to the other two models.

The findings of the experiment results can be summarised as follows. The weighted SSE approach gives the best performance on nearly all measures compared both to the standard and post-trained BLSTM approaches. At the same time it is the only model which is consistently better than previously reported LSTM and combined HMM+LSTM models [17], which is especially apparent from Figure 4. However, all presented approaches show on average improvements to the referenced prior work. It appears that the bigger amount of artificially mixed data as well as the bigger and different BLSTM network structure can compensate the use of advanced features from [17].

Conclusion

Significant improvements to state of the art overlap detection was achieved in the present work by utilising BLSTM RNNs and modified weighted SSE loss functions. Thereby two novelties were introduced: Firstly, training a speech overlap detector was done on random speech mixes only so that the models learn overlap independently from sociolinguistic context like turn-taking patterns, speech pause distribution and more. This gave robust results on basic features outperforming different techniques trained on the same, even though unprocessed, speech corpus with highly sophisticated features [17]. Secondly, all trained networks additionally learn and predict voice activity and gender of the speaker with reasonable performance measures. This shows that all these outputs can be recognised together at the same time by one single model – to the author's best knowledge the first successfull combination of that kind.

Future work can benefit from these basic insights insofar, as artificially mixed speech overlap can be utilised to create much more accurate overlap detectors only by using more mixed data and more complex networks while using ordinary features. Furthermore, it can now be considered building speech analysis LSTMs that provide an additional overlap output.

Acknowledgements

This work is supported by the EU's Horizon 2020 Programme through the Innovative Action No. 688835 (DE-ENIGMA) and No. 645094 (SEWA). We further thank the NVIDIA Cooperation for their support of this research by Tesla K40-type GPU donation.

References

[1] Boakye, K., Trueba-Hornero, B., Vinyals, O., and Friedland, G., "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4353–4356, IEEE, 2008.

[2] Geiger, J. T., Eyben, F., Evans, N., Schuller, B., and Rigoll, G., "Using Linguistic Information to Detect Overlapping Speech," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, pp. 690–694, ISCA, ISCA, Lyon, France, 2013, (acceptance rate: 52 %, IF* 1.05 (2010)).

[3] Huijbregts, M., van Leeuwen, D. A., and Jong, F., "Speech overlap detection in a two-pass speaker diarization system," 2009.

[4] Suzuki, M., Kurata, G., Nagano, T., and Tachibana, R., "Speech recognition robust against speech overlapping in monaural recordings of telephone conversations," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5685–5689, IEEE, 2016.

[5] Heldner, M. and Edlund, J., "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, 38(4), pp. 555–568, 2010.

[6] West, C., "Against our will: Male interruptions of females in cross-sex conversation," *Annals of the New York Academy of Sciences*, 327(1), pp. 81–96, 1979.

[7] Chowdhury, A., Danieli, M., and Riccardi, G., "The role of speakers and context in classifying competition in overlapping speech," *INTERSPEECH, Dresden*, 2015.

[8] Chowdhury, S. A., Danieli, M., and Riccardi, G., "Annotating and categorizing competition in overlap speech," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5316–5320, IEEE, 2015.

[9] Galley, M., McKeown, K., Hirschberg, J., and Shriberg, E., "Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies," in *42nd Annual Meeting on Association for Computational Linguistics*, p. 669, Association for Computational Linguistics, 2004.

[10] Goldberg, J. A., "Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power-and rapport-oriented acts," *Journal of Pragmatics*, 14(6), pp. 883–903, 1990.

[11] Anderson, K. J. and Leaper, C., "Meta-analyses of gender effects on conversational interruption: Who, what, when, where, and how," *Sex Roles*, 39(3-4), pp. 225–252, 1998.

[12] Shriberg, E., Stolcke, A., and Baron, D., "Observations on overlap: findings and implications for automatic processing of multi-party conversation." in *INTERSPEECH*, pp. 1359–1362, 2001.

[13] Zelenák, M. and Hernando, J., "The Detection of Overlapping Speech with Prosodic Features for Speaker Diarization." in *INTERSPEECH*, pp. 1041–1044, 2011.

[14] Zelenak, M., Segura, C., Luque, J., and Hernando, J., "Simultaneous speech detection with spatial features for speaker diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2), pp. 436–446, 2012.

[15] Vipperla, R., Geiger, J., Bozonnet, S., Wang, D., Evans, N., Schuller, B., and Rigoll, G., "Speech Overlap Detection and Attribution Using Convolutive Non-Negative Sparse Coding," in *37th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2012*, pp. 4181–4184, IEEE, IEEE, Kyoto, Japan, 2012.

[16] Geiger, J. T., Vipperla, R., Bozonnet, S., Evans, N., Schuller, B., and Rigoll, G., "Convolutive Non-Negative Sparse Coding and New Features

for Speech Overlap Handling in Speaker Diarization," in *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, ISCA, ISCA, Portland, OR, 2012.

[17] Geiger, J. T., Eyben, F., Schuller, B., and Rigoll, G., "Detecting Overlapping Speech with Long Short-Term Memory Recurrent Neural Networks." in *INTERSPEECH*, pp. 1668–1672, 2013.

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

[19] Laskowski, K., Heldner, M., and Edlund, J., "On the dynamics of overlap in multi-party conversation," in *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, pp. 846–849, Curran Associates, Inc, 2012.

[20] Yella, S. H. and Valente, F., "Speaker diarization of overlapping speech based on silence distribution in meeting recordings," in *INTERSPEECH*, EPFL-CONF-192713, 2012.

[21] Yella, S. H. and Bourlard, H., "Improved overlap speech diarization of meeting recordings using long-term conversational features," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7746–7750, IEEE, 2013.

[22] McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., et al., "The AMI Meeting Corpus," in *5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, 2005.

[23] Eyben, F., *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*, Springer, 2016.

[24] Hochreiter, S. and Schmidhuber, J., "Long Short-Term Memory," *Neural computation*, 9(8), pp. 1735–1780, 1997.

[25] Graves, A. and Schmidhuber, J., "Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks," in *Advances in neural information processing systems*, pp. 545–552, 2009.

[26] Schuster, M. and Paliwal, K. K., "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, 45(11), pp. 2673–2681, 1997.

[27] Weninger, F., Bergmann, J., and Schuller, B., "Introducing CURRENNT: The Munich Open-Source CUDA Recurrent Neural Network Toolkit," *The Journal of Machine Learning Research*, 16(1), pp. 547–551, 2015.