Detection of Overlapping Speech for the Purposes of Speaker Diarization

Marie Kunešová[1,2][0000−0002−7187−8481], Marek Hrúz[1][0000−0002−7851−9879],
Zbyněk Zajíc[1][0000−0002−4153−6560], and Vlasta Radová[1,2][0000−0002−3258−8430]

University of West Bohemia, Faculty of Applied Sciences
[1]NTIS - New Technologies for the Information Society and [2]Dept. of Cybernetics
Univerzitní 8, 306 14 Plzeň, Czech Republic
{mkunes, mhruz, zzajic}@ntis.zcu.cz, radova@kky.zcu.cz

Abstract. The presence of overlapping speech has a significant negative impact on the performance of speaker diarization systems. In this paper, we employ a convolutional neural network for the detection of such speech intervals and evaluate it in terms of the potential improvements to speaker diarization. We train the network on specifically-created synthetic data, while the evaluation is performed on the AMI Corpus and the SSPNet Conflict Corpus.

Keywords: Overlapping Speech · Speaker Diarization · Convolutional Neural Network

1 Introduction

In natural human conversations, there are often instances where multiple individuals speak at the same time – this includes interruptions, backchannel responses (e.g. “yeah”, “uh-huh”), or simply brief natural overlaps during rapid turn-taking. Such overlapping speech can prove problematic for automatic speech processing, particularly for speech recognition and for speaker diarization.

Specifically, in our recent paper [17], we found that accurate detection of overlapping speech would have improved the results of our diarization system by a significant margin: on the development set of the DIHARD II corpus, the use of ground-truth overlap labeling decreased the Diarization Error Rate (DER) from 20.78 to 16.16% (22% relative improvement). Similar observations have previously been made by other authors, e.g. in [8]. This potential for improvement is what motivated our work on overlap detection.

The research of this topic has evolved over the last decade with only mild success: The more traditional approaches rely on a careful selection of handcrafted features, to be fed into a HMM decoder [2,16] or a neural network [1,3]. A more recent alternative is to let a neural network extract the relevant information form “raw” input, such as a spectrogram of acoustic signal [10,14]. Our work is also based on this latter approach.

1.1 Problems with Data

During our work on the overlap detector, we have encountered some difficulties, particularly with the lack of suitable data.

Training and evaluating an overlap detector generally requires a large amount of well-annotated data with frequent overlaps. Unfortunately, there do not appear to be any publicly available datasets made specifically for this purpose, and other corpora often lack sufficiently precise labels.

Like some other authors (e.g. [1,10,14]), we resorted to creating our own synthetic training data – we describe this in section 3.1. However, the same problem with inadequate labels also applies to subsequent evaluation of the overlap detector on real corpora, and its use in a speaker diarization system.

It is difficult, as well as very time-consuming, to precisely annotate overlapping speech. For this reason, reference annotations often tend to exclude very short occurrences ($< 0.5\,\mathrm{s}$), especially those at the boundaries between speakers. This can be a problem if the overlap detector is more sensitive, as such detected overlaps will be incorrectly evaluated as false alarms.

A similar issue is also with the classification of overlaps with voiced non-speech sounds such as laughter or humming. On the one hand, these sounds can often be identified as a specific speaker and can negatively affect speaker diarization. On the other hand, these events are often not included in speech transcripts, especially when they happen in the background of another speakers' speech, so such regions may be (in this case incorrectly) marked as non-overlap in the reference. This may again lead to a seemingly high false alarm rate of an overlap detector evaluated on such data.

When evaluating overlap detection, various authors deal with these issues in different ways, such as by ignoring very short intervals, applying generous tolerance windows, or, if they can be identified by other means, excluding intervals with non-speech from evaluation.

2 Overlap Detector

We have previously [6,7] used a Convolutional Neural Network (CNN) for the detection of speaker changes in an audio stream. In this paper, we employ the same general approach for the detection of overlapping speech.

A summary of the network architecture can be found in Table 1. The input of the network is a spectrogram of a short window of the acoustic signal. The output of the last layer is a value between 0 and 1, indicating the probability of overlapping speech in the middle of the window. Training references use a fuzzy labeling function, with a linear slope (width 0.4 s) at the boundaries between overlap and non-overlap (see the lower two plots of Figure 3 for an example). The sliding window has a length of 1 s and is shifted with a step of 0.05 s.

We use a median filter with a window length of 5 samples to smooth the raw network output, then apply a threshold to obtain overlap / non-overlap classification. Additionally, we fill in any gaps (non-overlaps within a longer overlap) which are shorter than 0.1 s, and then discard overlaps under 0.5 s, as these are unlikely to be included in the reference labeling (as discussed in section 1.1).

Table 1. Summary of the network architecture.

Layer	Kernels	Size	Shift
Convolution	128	8 x 16	2 x 2
Max Pooling		2 x 2	2 x 2
Batch Normalisation			
Convolution	256	4 x 4	1 x 1
Max Pooling		2 x 2	2 x 2
Batch Normalisation			
Convolution	512	3 x 3	1 x 1
Max Pooling		2 x 2	2 x 2
Batch Normalisation			
Fully Connected	1024		
Fully Connected	256		
Fully Connected	1		

3 Data

3.1 Synthetic Training Data

Given the lack of sufficient real data (as mentioned in section 1.1), we resorted to artificially creating training data from two corpora of read English speech, LibriSpeech [13] and TIMIT [5], using an automated and randomized process. In the creation of this synthetic dataset, we used some of the ideas previously described in [4,14].

TIMIT - The TIMIT corpus consists of the recordings of single English sentences, approx. 2–5 s long. We used the data from 320 speakers for training.

To obtain overlapped data, we first concatenated all utterances from a single speaker into one file of approx. 30 s, with random-length pauses (up to 2 s) in-between. In order to avoid noticeable seams, the silence at the beginning and end of each utterance is linearly tapered. Then, files from two random speakers are combined at different volumes and augmented with added background noise (office, hallway, meeting) from the DEMAND database [15] and, for 50% of the files, reverberation (via convolution with room impulse response from the AIR database [9]). The result is illustrated in Figure 1.

Reference labels were created with the use of the original phone-level transcripts - so that only the intervals where both speakers are truly active are labeled as overlap.

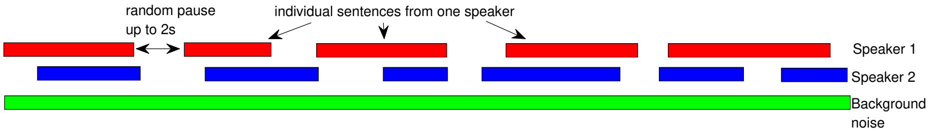

Fig. 1. Creation of artificial overlap data from the TIMIT corpus.

LibriSpeech - We also used data from the “train-other-500” set of the LibriSpeech corpus - this consists of approx. 500 hours of speech from over 1000 speakers, in the form of 10–15 s long recordings derived from audiobooks.

Given the very large amount of available LibriSpeech data, we were able to create several different types of overlaps, to better represent the possibilities which may occur in real data (see Fig. 2):

a) Two full length (approx. 10–15 s) utterances, with an overlap of 1/2 length
b) Two utterances with a short overlap (up to 2 s) or pause (up to 1 s) in-between
c) A single utterance with an inserted word or phrase from another speaker: Utterance 1 is split on pauses and a randomly selected speech interval (0.25–2 s) is placed over utterance 2, either: fully overlapping speech, fully inside a pause, or randomly placed.

In the case of b) and c), the resulting file is shortened to 5 s of non-overlap data on each side of the overlap or pause, as seen in Fig. 2. The is done to keep a better ratio between non-overlaps and overlaps.

As with TIMIT, we added noise and reverberation. Speech/non-speech labelling was obtained using a voice activity detector (VAD) on the original single speaker data without added noise.

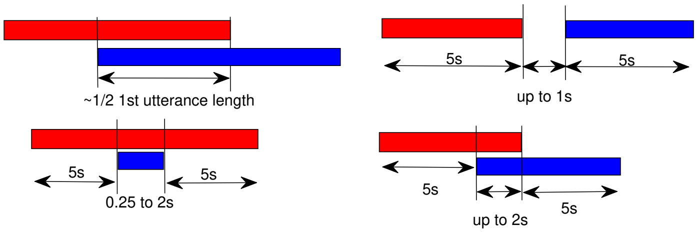

Fig. 2. Different types of synthetic overlap created from the LibriSpeech corpus. (Additive noise not shown.)

3.2 Test Data

We evaluated our overlap detector on three different sets of data: one artificially created dataset and two corpora of real conversations.

LibriSpeech Test Data - We created synthetic test data from the “test-other” subset of the LibriSpeech corpus, in a very similar way to the TIMIT training data - but with 5-10 s pauses between a single speaker’s utterances, and without the added noise or reverberation.

SSPNet Conflict Corpus [1] [11] - This is a dataset of French-language political debates, consisting of 1430 clips of exactly 30 s each, cut from 45 separate debates. Each clip usually involves between 2–5 people and, as these are spontaneous discussions, there are frequent instances of overlapping speech. The same corpus was also used for overlap detection in [10].

We selected 5 debates (06-05-31, 06-09-20, 06-10-11, 07-05-16, and 08-01-15; 161 files total = 80.5 minutes of audio data) as development data for tuning the decision threshold, the remainder (1269 files = 10.6 hours) was used for evaluation.

As the corpus hadn’t been created with overlaps in mind, the original reference labels are relatively rough in this regard - they do not include very short overlaps at speaker changes or during isolated backchannel responses (e.g. “Oui, ... oui.”), nor shorter non-overlap intervals within a longer overlap region (e.g. pauses in the speech of one speaker). However, our network proved capable of detecting all of the above. For this reason, we also selected a small number of audio clips (30 files = 15 minutes) and manually corrected the labels[2] to better correspond to the audio data (example shown in Fig. 3). These 30 files were then evaluated separately, using both the original and corrected labels, to illustrate how labelling quality affects the reported results (see Table 2).

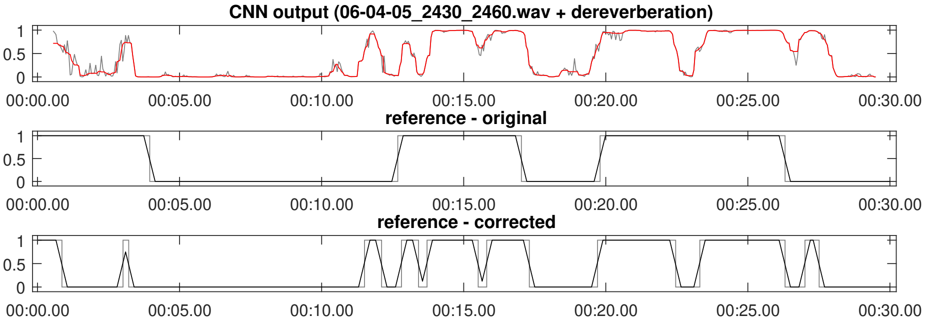

Fig. 3. Example output (raw + median filter) for dereverberated SSPNet data and the corresponding reference labels - original (middle) and manually corrected (bottom).

[1] http://www.dcs.gla.ac.uk/vincia/?p=270

[2] The corrected labels and the code used in this paper can be found at https://github.com/mkunes/CNN-overlap-detection

AMI Meeting Corpus [3] - A set of recordings from meetings between 3–6 people. We tested the overlap detector on the “headset mix” data, using the same train/validation/test split as Sajjan et al. [14]. We used the original transcripts as ground truth, rather than Sajjan et al.’s force-aligned labels[4], as we found the latter to be less accurate in some regards, but both versions have errors – in particular, there are many instances where overlaps with non-speech such as laughter are not labeled.

The corpus consists of several subsets of meetings which were recorded at different sites and vary in audio and transcription quality. We particularly found the Idiap scenario meetings (IS) to have very different optimal settings from the rest of the test set, so we also evaluate them separately.

4 Evaluation

Previous works on overlap detection use a variety of different evaluation metrics, including frame-level precision and recall or F-score [1], or per-overlap miss and false alarm rate [10] (see Table 3). However, as our main motivation is the improvement of speaker diarization, we decided to primarily evaluate the overlap detector in terms of the potential gains in diarization performance.

There are two main ways in which overlap information can be used in a diarization system: First, by excluding such intervals from any clustering process, we can avoid “polluting” the clusters and negatively influencing the clustering decisions. Secondly, in the final output, we assign multiple labels to each overlap region. The exact benefits of the first point depend on the diarization system in question. Thus, in this paper, we concentrate on the latter point, which is easier to quantify.

Diarization systems are usually evaluated in terms of Diarization Error Rate (DER), which consists of three types of error: *missed speech* (including missing speakers in overlaps), *false alarm* (silence mislabelled as speech or non-overlap as overlap), and *speaker error* (wrong speaker). In an ideal diarization system with no overlap handling, false alarm and speaker error will be zero, while missed speech will correspond to the amount of overlapping speech in the data.

In our evaluation of the potential benefits of overlap detection, we assume that the diarization system assigns two speaker labels to every detected region of overlapping speech (regardless of the true number speakers), and that (for correctly detected overlap) it does so perfectly – the *speaker error* is still zero.

In such a scenario, correctly detected overlaps will directly decrease the amount of *missed speech* compared to the baseline system, while false overlaps will increase the *false alarm*. Thus, we can obtain the potential improvement as the difference between the two values.

Note: By the correct definition, DER is calculated as a ratio of total speech (excluding silence), with overlaps being counted multiple times – once for each

[3] http://groups.inf.ed.ac.uk/ami/corpus/

[4] https://github.com/BornInWater/Overlap-Detection

speaker. However, for simplicity, we calculate the potential improvements here as relative to the total length of the audio data.

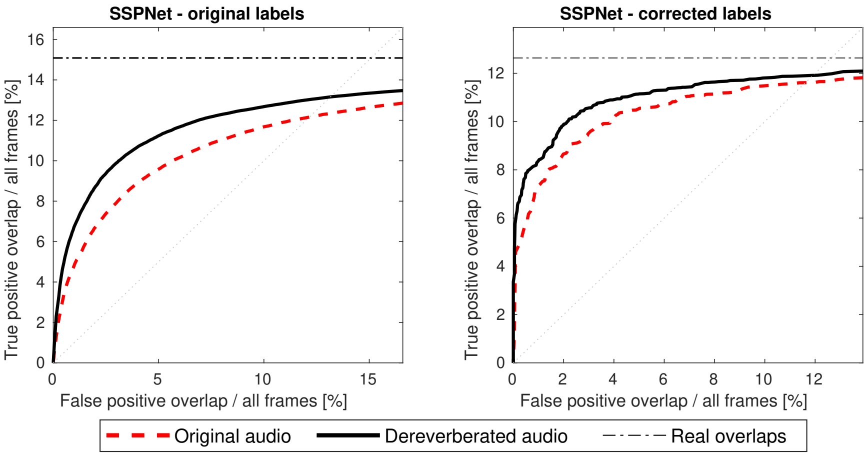

Fig. 4. False Positive vs True Positive for SSPNet data (frame-level percentage of all audio). Original labels (all 1269 test files) on the left, corrected labels (30 files, 15 minutes total) on the right. “Real overlaps” denotes the overlap percentage in the ground truth.

4.1 Results

The results we achieved on the different corpora are shown in Table 2 and in Figures 4 and 5.

The overlap detector appears to work very well on clean audio, such as the synthetic LibriSpeech data and the SSPNet Conflict Corpus. The network also seems to be very sensitive and capable of detecting even very short overlaps and non-overlaps, down to the level of individual words – a much greater precision than typically found in the reference annotations (as illustrated by the example output in Figure 3).

On the other hand, the detector had issues with the AMI corpus. This may be in part due to errors in the reference labels – we have found instances of missing speech, or long unlabeled intervals where multiple people are laughing, which our network also considers to be overlaps. However, the lower performance is likely also caused by the higher level of noise in the these recordings, as well as the sometimes very large differences in the voice volumes of individual speakers. This is evidenced by the fact that we were able to improve the results to some extent by including the training set of the AMI corpus in the training data – this suggests that we may need to improve the synthetic dataset.

Table 2. Results of overlap detection on evaluation data. Overlap percentages are relative to total audio length, precision and recall are calculated per frame. (Ref. = Real overlap ratio according to the reference, TP = True Positive, FP = False Positive, Δ = TP - FP $\simeq$ potential DER improvement).

	Overlaps [% of all frames]						
Dataset	Ref.	TP	FP	Δ	Prec.	Rec.	Thresh.
LibriSpeech test mix	16.32	11.99	2.82	**9.18**	0.81	0.73	0.25
SSPNet - original (10.6 h)	14.77	7.86	2.94	4.92	0.73	0.52	0.80
+ dereverberation		9.58	2.68	**6.90**	0.78	0.63	0.70
SSPNet - precise (15 min)	12.62	8.05	1.42	6.63	0.85	0.65	0.80
+ dereverberation		8.90	1.41	**7.49**	0.86	0.71	0.70
SSPNet - original (15 min)	12.86	7.47	2.00	5.47	0.79	0.59	0.80
+ dereverberation		8.60	1.71	**6.89**	0.83	0.68	0.70
AMI test (all subsets)	12.21	2.25	0.96	1.30	0.70	0.19	0.50
+ dereverberation		2.38	1.03	**1.34**	0.70	0.20	0.25
AMI test (only "IS")	7.82	2.75	1.34	1.41	0.67	0.36	0.80
+ dereverberation		3.71	1.76	**1.95**	0.68	0.48	0.60
Retrained network - with added AMI training data:							
AMI test (all subsets)	12.21	5.73	2.35	**3.38**	0.71	0.48	0.50
+ dereverberation		4.92	1.61	3.31	0.75	0.41	0.25
AMI test (only "IS")	7.82	3.28	1.24	2.04	0.73	0.43	0.90
+ dereverberation		3.73	1.61	**2.12**	0.70	0.48	0.80

Initial experiments also suggested that the network had problems with reverberant speech, which was often incorrectly labeled as overlap. We have partly mitigated this effect by adding reverberation to the training data (as described in section 3.1). However, we have also experimented with dereverberation of the test data - to evaluate the potential benefits, we used the WPE Dereverberation package[5] created by Nakatani et al. [12]. Even with the default settings without any adjustments, this has proven to be clearly beneficial for SSPNet data, but for AMI the difference is negligible (with the exception of the Idiap scenario (IS) meetings).

Finally, in Table 3 we present a comparison of our overlap detector with some other works on the topic. This comparison is somewhat complicated by the fact that other authors have used many different combinations of datasets (or their parts) and metrics to evaluate their systems. For instance, while 3 other systems in the table used the AMI corpus, each of them selected different files. Similarly,

[5] http://www.kecl.ntt.co.jp/icl/signal/wpe/index.html

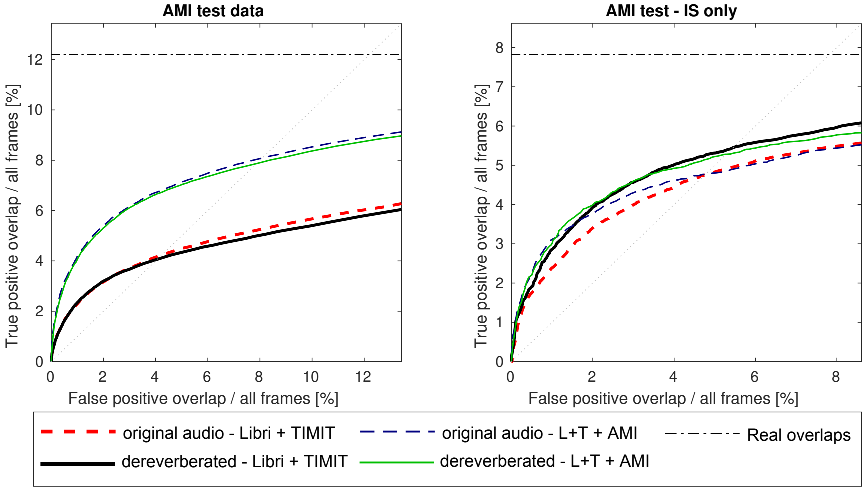

Fig. 5. False Positive vs True Positive for AMI data (frame-level percentage of all audio), with overlap detector trained only on synthetic LibriSpeech + TIMIT data or with the addition of AMI training data. Results are for all test files (left) and only for the Idiap scenario meetings (right).

the results of [10] on the SSPNet Conflict Corpus are not directly comparable with ours, as their system was evaluated only on voiced frames.

5 Conclusion

In a previous paper [17], we measured the improvement achievable with a ground-truth overlap labelling in a real diarization system. Here, we looked at the problem from the opposite angle - evaluating an overlap detector under the assumption that the diarization system otherwise functions without error.

The results we achieved here appear to be very promising, particularly those on relatively clean and noise-free data, although some more work will be required in order to improve the performance on data with higher levels of noise. The next step in our research will be to connect the two systems and to evaluate the full effects of the overlap detector on the entire diarization pipeline.

Acknowledgements

This research was supported by the Ministry of Culture Czech Republic, project No.DG16P02B048. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme “Projects of Large Research, Development, and Innovations Infrastructures” (CESNET LM2015042), is greatly appreciated.

Table 3. Comparison of the proposed system (selected results from Table 2, without added AMI training data) with prior works. With the exception of our "all subsets" and [14]'s "original labels" AMI results, no two systems used identical test data and ground-truth labelling.

System	Dataset	Prec.	Rec.	F-score	Accuracy
proposed	LibriSpeech test mix	0.81	0.73	0.77	0.93
	SSPNet (original labels)	0.73	0.52	0.61	0.90
	+ dereverberation	0.78	0.63	0.70	0.92
	AMI (16 files - all subsets)	0.70	0.19	0.30	0.89
	+ dereverberation	0.70	0.20	0.31	0.89
	AMI (4 files - only "IS" subset)	0.67	0.36	0.47	0.94
	+ dereverberation	0.68	0.48	0.56	0.94
[1]	Custom dataset	0.81	0.78	0.8	0.802
[10]	SSPNet (voiced frames only)	0.71	0.78	0.75	0.92
[2]	AMI (12 "IS" files, force aligned)	0.67	0.26	0.38	–
[14]	AMI (16 files, original labels)	–	–	–	76.0 / 60.6*
–	AMI (16 files, force aligned labels)	–	–	–	87.9 / 71.0*
[16]	AMI (25 files)	–	–	0.51	–

(* overlap-detection accuracy / single-speaker detection accuracy)

References

1. Andrei, V., Cucu, H., Burileanu, C.: Detecting overlapped speech on short time-frames using deep learning. In: Proc. Interspeech. pp. 1198–1202 (2017)
2. Boakye, K., Vinyals, O., Friedland, G.: Two's a crowd: Improving speaker diarization by automatically identifying and excluding overlapped speech. In: Proc. Interspeech. pp. 32–35 (2008)
3. Diez, M., Landini, F., Burget, L., Rohdin, J., Silnova, A., Žmolíková, K., Novotný, O., Veselý, K., Glembek, O., Plchot, O., Mošner, L., Matějka, P.: BUT system for DIHARD speech diarization challenge 2018. In: Proc. Interspeech. pp. 2798–2802 (2018)
4. Edwards, E., Brenndoerfer, M., Robinson, A., Sadoughi, N., Finley, G.P., Korenevsky, M., Axtmann, N., Miller, M., Suendermann-Oeft, D.: A free synthetic corpus for speaker diarization research. In: Proc. SPECOM. pp. 113–122 (2018)
5. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., Zue, V.: TIMIT acoustic-phonetic continuous speech corpus LDC93S1 (1993)
6. Hrúz, M., Kunešová, M.: Convolutional neural network in the task of speaker change detection. In: Proc. SPECOM. pp. 191–198 (2016)
7. Hrúz, M., Zajíc, Z.: Convolutional neural network for speaker change detection in telephone speaker diarization system. In: Proc. ICASSP. pp. 4945–4949 (2017)
8. Huijbregts, M., Wooters, C.: The blame game: Performance analysis of speaker diarization system components. In: Eighth Annual Conference of the International Speech Communication Association (2007)
9. Jeub, M., Schafer, M., Vary, P.: A binaural room impulse response database for the evaluation of dereverberation algorithms. In: 16th International Conference on Digital Signal Processing. pp. 1–5 (2009)

10. Kazimirova, E., Belyaev, A.: Automatic detection of multi-speaker fragments with high time resolution. In: Proc. Interspeech. pp. 1388–1392 (2018)
11. Kim, S., Valente, F., Filippone, M., Vinciarelli, A.: Predicting continuous conflict perception with bayesian gaussian processes. IEEE Transactions on Affective Computing **5**(2), 187–200 (2014)
12. Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., Juang, B.: Speech dereverberation based on variance-normalized delayed linear prediction. IEEE Transactions on Audio, Speech, and Language Processing **18**(7), 1717–1731 (2010)
13. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: LibriSpeech: An ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5206–5210 (2015)
14. Sajjan, N., Ganesh, S., Sharma, N., Ganapathy, S., Ryant, N.: Leveraging LSTM models for overlap detection in multi-party meetings. In: Proc. ICASSP. pp. 5249–5253 (2018)
15. Thiemann, J., Ito, N., Vincent, E.: The Diverse Environments Multi-channel Acoustic Noise Database: A database of multichannel environmental noise recordings. The Journal of the Acoustical Society of America **133**(5), 3591–3591 (2013)
16. Yella, S.H., Bourlard, H.: 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) **22**(12), 1688–1700 (2014)
17. Zajíc, Z., Kunešová, M., Hrúz, M., Vaněk, J.: UWB-NTIS speaker diarization system for the DIHARD II 2019 challenge. In: Submitted to Interspeech (2019), https://arxiv.org/abs/1905.11276