Interspeech 2018
2-6 September 2018, Hyderabad

BUT system for DIHARD Speech Diarization Challenge 2018

Mireia Diez, Federico Landini, Lukáš Burget, Johan Rohdin, Anna Silnova, Kateřina Žmolíková, Ondřej Novotný, Karel Veselý, Ondřej Glembek, Oldřich Plchot, Ladislav Mošner, Pavel Matějka

Brno University of Technology, Speech@FIT, Czechia
{mireia,landini}@fit.vutbr.cz

Abstract

This paper presents the approach developed by the BUT team for the first DIHARD speech diarization challenge, which is based on our Bayesian Hidden Markov Model with eigenvoice priors system. Besides the description of the approach, we provide a brief analysis of different techniques and data processing methods tested on the development set. We also introduce a simple attempt for overlapped speech detection that we used for attaining cleaner speaker models and reassigning overlapped speech to multiple speakers. Finally, we present results obtained on the evaluation set and discuss findings we made during the development phase and with the help of the DIHARD leaderboard feedback.

Index Terms: Speaker Diarization, Variational Bayes, HMM, i-vector, x-vector, Overlapped speech, DIHARD

1. Introduction

The efforts on speaker diarization (SD) have lately focused mainly on meetings and conversational telephone speech. Although remarkable improvements have been obtained under these conditions, the state-of-the-art systems are still not that successful in other domains. The aim of the 2018 DIHARD challenge [1] was to bring attention to other corpora such as interviews collected in both interior and exterior environments, conversations in restaurants, child language acquisition recordings in family environments or web videos. These considerably different domains imply that different approaches need to be considered when dealing with clean and noisy speech, silence and overlapped speech to enhance the capabilities of the SD systems. For further information on the different domains, we kindly refer the reader to [1].

In this paper, we mainly describe the BUT system that produced the best results on the evaluation set. However, many other ideas were considered and analyzed during the challenge. Some of them were the application of denoising and dereverberation algorithms in order to enhance the audio signals, the analysis of the level of reverberation in a per utterance manner, different types of normalization of the signal and the use of different corpora with different characteristics to train the models. We considered also the addition of a domain identification system in order to take advantage of the best per domain settings. For the lack of space, we present results only on a selection of these experiments. All these experiments are presented for the development set as by the submission deadline of this paper (the same as for the DIHARD challenge system submission) the evaluation set labels were still not available. The numbers presented for the eval set are those obtained by the challenge leaderboard feedback.

When devising our system, we considered different aspects as described in the following sections. They include different corpora in addition to the DIHARD development dataset; mechanisms to enhance the audio signal and to obtain acoustic features; strategies for initializing the model; a domain identification system; an overlapped speech detector; and unsupervised use of the evaluation data.

2. Variational Bayes diarization System

Our main approach to the diarization problem is based on a Bayesian Hidden Markov Model (HMM) with eigenvoice priors [2]. This model assumes that the sequence of speech features representing a conversation is generated from a HMM, where each state represents one speaker and the transitions between the states correspond to speaker turns. The ergodic HMM is used, where transitions from any state to any state are possible. However, the transition probabilities are set in a way that discourages too frequent transitions between states in order to reflect speaker turns durations of a natural conversation. The HMM state (or speaker) specific distributions are modeled by Gaussian Mixture Models (GMMs) with informative eigenvoice prior imposed on the GMM parameters. Such prior, which is essentially the same as in i-vectors [3] or Joint Factor Analysis (JFA) [4] models, allows us to robustly estimate speaker distributions, which facilitates discrimination between the speaker voices in the input recording. The proposed Bayesian model offers an elegant approach to SD as a straightforward and efficient Variational Bayes (VB) inference in a single probabilistic model addresses the complete SD problem: For each input conversation, we construct a HMM with preferably more states than what is the assumed number of speakers in the conversation and we start with some initial (possibly random) assignment of frames to HMM states. Then, each VB training iteration refines the HMM state specific distributions and recalculates the soft (probabilistic) assignment of frames to the HMM states. During the VB training, the complexity control inherent in the Bayesian learning automatically drops the redundant HMM states (i.e. learns zero transition probabilities into such states) and decides on the number of speakers in the conversation. The final assignment of frames to the "surviving" HMM states gives the solution to the diarization problem. An open source code for the algorithm is provided in [5] and more details on the algorithm can be found in [2]. In the rest of this paper, we will refer to this approach as VB diarization.

The work was supported by Czech Ministry of Interior project No. VI20152020025 "DRAPAK", European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 748097, the Marie Sklodowska-Curie cofinanced by the South Moravian Region under grant agreement No. 665860, Google Faculty Research Award program, Czech Science Foundation under project No. GJ17-23870Y, Technology Agency of the Czech Republic project No. TJ01000208 "NOSICI", and by Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602".

10.21437/Interspeech.2018-1749

2.1. Baseline initialization of the algorithm

The VB diarization system can be initialized setting an upper bound on the number of speakers for the input utterance and using a random assignment of frames to speakers. Also, it can be initialized using a labeling attained with an external diarization algorithm.

For all the experiments presented in section 3 we initialize the VB diarization using the output from the diarization system described in [6], which works as follows: each utterance is segmented into 2 second speech segments, overlapped by 0.5 seconds. 64 dimensional i-vectors [3] are extracted from each segment and projected by means of Principal Component Analysis to 3 dimensions [7]. These segments are then clustered using Agglomerative Hierarchical Clustering (AHC) using calibrated Probabilistic Linear Discriminant Analysis (PLDA) similarity scores [8, 9].

3. Experimental Set Up

3.1. Data resources

To evaluate the performance of our approach on the development set, we prepared two independent training sets, and two corresponding evaluation strategies. The first set (*Tel*) consists of 8 kHz (mostly telephone) data from NIST SRE 2004 - 2008 datasets, which amounts to around 1400h of speech. When using *Tel* set to train our systems, we evaluated them on the whole development set. To build the second set (*devPart*), we partitioned the DIHARD development set [10, 11] into two speaker-disjoint subsets with balanced number of utterances across all domain. The domains that could not be partitioned with this criteria (SCOTUS and SLX) were assigned each to one of the partitions. To evaluate the systems trained using *devPart*, we trained on one partition and evaluated on the other, and vice-versa. We report the pooled results from both partitions.

Several attempts were made to extend both *Tel* and *devPart* training sets. The training sets were augmented with various combinations and proportions of data from librispeech [12], VoxCeleb [13], NIST SRE08 interviews, AMI [14] and VanDam [15], but the results obtained on the development set were always worse than when training the systems only with the training sets *Tel* or *devPart*.

For the alternative initializations mentioned in section 4, the x-vector extractor was trained on data from NIST SRE 2004-2008, Fisher English and Switchboard.

3.2. Voice Activity Detection (VAD)

For the Track 2 of the challenge, in which no golden segmentation labels were provided, a VAD system was used in order to discard silence and feed the rest of the system only with speech segments. Moreover, for the Track 1, we explored using the VAD output together with the golden segmentation labels in order to discard even more silence when training the systems on DIHARD development data. However, while the results improved for some domains, they deteriorated for others producing similar results overall.

Our VAD is based on a neural network (NN) trained for binary, speech/non-speech, classification of speech frames. The 288-dimensional NN input is derived from 31 frames of 15 log Mel filter-bank outputs and 3 pitch features. The NN with 2 hidden layers of 400 sigmoid neurons was trained on the Fisher English with labels provided from Automatic Speech Recognition alignment. Per-frame logit posterior probabilities of speech were smoothed by averaging over consecutive 31 frames and thresholded to at the value of 0 to give the final hard per frame speech/ non-speech decision. See [16] for more detailed description of the VAD system.

For the *Tel* system training, using a more aggressive VAD gave better results. The VAD is based on the BUT Czech phoneme recognizer [17], dropping all frames that are labeled as silence or noise. The recognizer was trained on the Czech CTS data, but we have added noise with varying Signal to Noise Ratio (SNR) to 30% of the database.

3.3. Features

Several efforts were made exploring the best set of features. Our baseline feature set are htk-based Mel Frequency Cepstral Coefficients (MFCC), with 19 coefficients plus Energy extracted from 8kHz audios. Cepstral Mean and Variance Normalization (CMVN) methods had proven to be harmful for diarization, as they remove channel information that could be useful to identify speakers in certain scenarios. Still, given the variety of domains for the DIHARD dataset, we decided to revisit the different feature extraction methods. Also, features extracted from 16kHz data were tested. We highlight the most significant findings.

System	Tel	devPart
No normalization	29.75	26.34
floating CMVN	25.28	28.25
Per-utterance CM	26.41	27.02
16KHz	–	27.91
16KHz+deltas	–	27.78

Table 1: *%DER on the DIHARD development set*

In Table 1 we present results for different feature extraction methods. We can observe that for the system trained on telephone data, floating window CMVN and per-utterance Cepstral Mean substraction enhance performance, whereas they hurt for the system trained on the dev set. Extracting features from 16kHz does not provide any gains on the development set.

3.4. Data Processing

Denoising We used a NN autoencoder [18] which consists of three hidden layers with 1500 neurons in each layer. The input of the autoencoder is a central frame of a log-magnitude spectrum with a context of +/- 15 frames (in total 3999-dimensional input). The output is a 129-dimensional enhanced central frame. We used Mean Square Error (MSE) as objective function during training. The Fisher English database parts 1 and 2, approximately 1800 hours of audio, were used to train the autoencoder. The datasets were artificially corrupted with additive and convolutive noise at SNR level 0-21dB from Freesound library [19] and room impulse responses were taken from AIR [20], C4DM [21, 22], MARDY [23], OPENAIR [24], RVB 2014 [25], and RWCP [26].

Dereverberation We used the Weighted Prediction Error (WPE) [27] method to remove late reverberation from the data. We estimated a dereverberation filter on Short Time Fourier Transform (STFT) spectrum for every 100 second block of an utterance. To compute the STFT, we used 32 ms windows with 8 ms shift. We set the filter length and prediction delay to 20 and 3 respectively for 16 kHz, and 10 and 2 respectively for 8 kHz data. The number of iterations was set to 3.

Results in Table 2 show the benefits of using denoising and dereverberation on the development set. As it can be seen, both

System	Tel		devPart	
Set	allDEV	ADOS	allDEV	ADOS
Baseline	29.75	33.55	26.34	28.90
+denoising	27.33	26.60	26.65	24.02
+dervberb.	27.89	32.52	26.82	23.62

Table 2: *%DER for different data processing approaches*

approaches seem to foster performance when training the systems with *Tel*, and degrade when using *devPart*. Also, the improvements are not consistent in all domains, to illustrate this, we show results computed only over utterances coming from the subset corresponding to ADOS, from the development set.

4. Alternative initializations

For initialization with the i-vector-PLDA AHC approach explained in section 2.1 we experimented with several different i-vector configurations as well as NN-based speaker embeddings.

To extract speaker embeddings, referred to as *x-vectors*, we employed the architecture described in [28] (embedding A). We trained the NN with the corresponding Kaldi recipe [29] except that we used the data described in Section 3.1 in order to comply with the rules of the DIHARD challenge and that we reduced the minimum number of utterances a speaker needs to have in order to be included in the training set from 8 to 6.

For both i-vectors and x-vectors we experimented with several different post-processing methods. The configuration that finally gave the best performance was x-vectors, projected to 150 dimensions by LDA with no length normalization, and with mean over the PLDA training set subtracted. It is worth noting that, although this system provided the best initialization for the VB diarization system in terms of final performance, it was not the best system in terms of performance on its own. The x-vector based initializations generally provided more speakers which the VB diarization seems to benefit from. Results for only two of the best performing systems using VB with i-vector based (Baseline) and x-vector based initializations are shown in Table 3 where we see the clear difference between the two approaches for the whole set. However, this pattern was different depending on the domain as we can see, for example, for RT04S and SEEDLINGS.

System	Tel		
Set	allDEV	RT04s	SEEDLINGS
Baseline (ivec)	29.75	48.13	47.38
xvec	24.65	41.05	48.59

Table 3: *%DER for the VB SD system using various initializations*

5. Domain Identification

DIHARD development and evaluation data originate from several different domains varying in channel conditions, number of speakers, etc. We believed that individual adjustment of model parameters to different domains could improve the overall performance of our system. For this reason, we built a subsystem that automatically classify evaluation recordings according to the domains given in the dev set. At the end, we found that the only strategy that generalized to the evaluation data was to detect LibriVox recordings, which always contain one speaker.

To classify domains, we trained a Gaussian Linear Classifier (i.e. Gaussian distributed classes with shared covariance matrix) on 64 dimensional i-vectors extracted from the whole recordings. The i-vector extractor was trained on the *Tel* dataset with addition of the LibriSpeech dataset [12]. The classifier was trained on the development data and 150 randomly chosen files from previously released Librivox data [19].

6. Overlapped speech detection

Since the current diarization system outputs one speaker label per frame, a post processing of the output was carried out. An overlapped speech detector was trained using three corpora in which overlaps are annotated: AMI[14] - 98h (1st microphone from 1st microphone array), Callhome - 17h (multi-lingual subset of train-sets), SRE08 test set - 186h (LDC2011S08). The training data were selected to contain a rich mixture of languages and domains. The model is a modified version of our VAD from section 3.2. The difference is that the NN has 3 outputs: 'speech', 'non-speech' and 'overlapped speech'. The per-frame score is the logit of posterior of 'overlapped speech' NN output. The rest is the same as for the VAD described in 3.2: fbank+pitch feature front-end, 2 hidden-layer NN topology and averaging of logit-scores over a window of 31 frames.

The detector was applied using two thresholds: one aggressive and one precise. The aggressive threshold was used to filter out *any overlapped speech* in order to feed the first pass of the VB algorithm only with reliable speech frames. Then, the precise threshold was used to detect speech segments that are overlapped speech with high probability. In a second pass of the VB algorithm the speaker models were kept fixed and those frames filtered out by the aggressive threshold but not by the precise one were assigned to speaker models. We saw that this approach helped the most for the noisiest domains on dev data. Then, only the frames spotted by the precise detector were given two speaker labels in order to reduce the false alarm rate. The frames were tagged according to the following rules:

- If the neighboring frames are assigned to different speakers, the overlap segment is assigned to those speakers.
- If only one of the neighboring segments is assigned to a speaker (the other to silence), or both were assigned to the same speaker, the overlap segment is assigned to that speaker and to the next most likely speaker according to the diarization model output.
- If both neighboring segments were silence segments, the overlap segment was assigned to the two most likely speakers according to the diarization model output.

System	devPart	
Set	allDEV	VAST
No Overlap handling	27.85	38.93
+ overlap cleaning	27.44	38.98
+ overlap reassignment	27.51	36.48

Table 4: *%DER for overlap speech handling techniques*

Table 4 shows the comparison of the two stages of the overlap handling. Applying the aggressive threshold allows to train cleaner models; however, this does not imply improvements for all domains, for example for VAST. When using the precise threshold together with the aggressive one, the results in general are subtlety worse because of false alarm speech that was close to 0% before applying the overlap handling. Nevertheless, the improvement on noisy conditions such as VAST showed us that this option would perform better in the evaluation set since

	DEV										EVAL
System	**ALL**	ADOS	DCIEM	LIBRIVOX	RT04s	SCOTUS	SEEDL.	SLX	YP	VAST	**ALL**
Baseline	29.75	33.55	13.90	22.93*	48.13	18.47	47.38	24.77	11.92	37.65	35.85
Sys1	19.96	16.56	6.80	7.95*	35.16	12.18	30.12	21.85	3.59	34.87	25.39
Sys2	–	–	–	–	–	–	–	–	–	–	25.07

Table 5: *%DER for VB SD systems for development, development domains and evaluation set (as reported by the leaderboard). *Note that Librivox domain identification is not used for reporting results on the dev set, as it was trained on the same data. Sys2 is Sys1 adapted in an unsupervised way on the eval data, therefore we only report results on the eval set.*

the domains added for such set that were not in the development set had similar characteristics.

Although processing overlapped speech provided some improvement on the final results, there is still a considerable margin for gains in comparison to the same strategy but using the oracle overlap labels so this part of the model needs further investigation, see section 8 for some insights on this matter.

7. Best performing systems

Our baseline system trained on telephone data *Tel*, which is the system described in our previous work [2], attains 29.75 %DER on dev set and 35.85% DER in eval as reported by the leaderboard feedback. Results for this system are provided in Table 5 also per domain on the development set.

One of our final best performing system on eval (Sys1) resulted as a combination of some of the techniques presented above, which were not always the best performing on the development set. We used 19 MFCC coefficients+energy+deltas extracted from dereverberated 16kHz signals. The VB algorithm was trained on the development set, excluding the VAST data, as we found the labeling too noisy to generate reliable speaker models. As unsupervised usage of the evaluation set was allowed, the eval data was used for the UBM training. The diarization system was initialized with an x-vector based PLDA AHC system. The overlapped speech treatment described in section 6 was applied. Files identified as LibriVox by our domain identification system were labeled as single speaker files. This combination resulted in a 25.39% DER on the evaluation set. In Table 5, we also provide results for the development set when applying the same configuration. Note, however, that the system applied to eval is trained using the whole dev set, whereas the one applied to dev uses the *devPart* training set and pooled evaluation.

The final best performing system on the first track of the challenge (Sys2) was identical to the one just described, except for one significant difference: we re-trained the eigenvoice subspace for the VB diarization on the pooled dev and eval data. However, this procedure required speaker labels for the evaluation data, which were not available. We obtained such labels in an unsupervised way as follows: the evaluation data was labeled using 5 diverse diarization systems developed for this challenge using different features, initializations, etc. For each evaluation recording, we clustered frames in such a way that all the systems agreed on having only one speaker in each cluster (i.e. cluster labels were given by concatenated speaker labels from all systems). We selected the largest cluster as training data representing one speaker, we discarded all clusters that were believed to be the same speaker by any of the system and we continued with the next largest cluster. This system attained 25.07% DER.

For the second track of the challenge, where part of the problem was also to provide VAD and the use of the golden segmentation of the eval set was not allowed, our submission was the Sys1 system but using the VAD described in section 3.2 instead of the golden segmentation, which attained the best results in the challenge for this track, 35.51% DER.

8. Discussion

The first DIHARD speech diarization challenge has brought for the first time in a long time a new framework for the evaluation of SD systems and new challenges including evaluation without collars, inclusion of overlapped speech in the evaluation and files from very distinct domains.

The DIHARD dataset with audios from several domains posed a problem for system optimization. Several of the improvements that we would observe in the development set when testing different training sets, feature extraction methods, data processings, etc. would prove to be beneficial for audios coming from a specific domain. In hopes of getting some benefit from these findings, we attempted using automatic domain identification on the eval set. Unfortunately our findings would often not generalize to the evaluation set, so we mainly focused on a single model for all domains. Once we get the evaluation data labels, research will be done to analyze this matter, which will reveal whether the domain identification has to be improved, whether the amount of files in development set per domain were not sufficient to make solid assumptions about the optimizations or bring some other insights.

The inclusion of overlapped speech on the evaluation of the systems had a high impact on the results: around a 9% DER on the development set which accounts for about 40% of the total error for our best systems. We attempted using automatic overlapped speech detection to deal with this. When analyzing the performance on the development set we observed the unavoidable inconsistency between the labeling for overlapped speech between the audios from different domains. Some labelings would ignore long segments of overlapped speech (and silences), whereas others would mark overlapped speech with quite a significant collar around the real audible overlapped speech. This makes it almost impossible to make proper use of an overlapped speech detector. Even though the use of no collar results in a much more objective evaluation of the systems, it remains to discuss whether the collar should not be used for the sake of homogenizing the datasets which we need to process by a single system.

9. Conclusions

The first DIHARD speech diarization challenge has proven to be a highly demanding and interesting contest, providing a dataset and framework with potential for new research lines focusing on different areas. We have presented and compared several approaches to deal with some of the posed challenges and discussed problems or difficulties found in the new framework. Research will be done on the evaluation set once the labels are released to gain insights on the real effects of the approaches presented in the paper: data augmentation, overlap speech detection, data processing and specially system fusion techniques will be revisited.

10. References

[1] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First DIHARD Challenge Evaluation Plan," https://zenodo.org/record/1199638, 2018.

[2] M. Diez, L. Burget, and P. Matejka, "Speaker diarization based on bayesian hmm with eigenvoice priors," in *Proceedings of Odyssey 2018, The speaker and Language Recognition Workshop*.

[3] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[4] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions Audio Speech Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[5] L. Burget, "VB Diarization with Eigenvoice and HMM Priors," http://speech.fit.vutbr.cz/software/vb-diarization-eigenvoice-and-hmm-priors, 2013, [Online; January-2017].

[6] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2014, pp. 413–417.

[7] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, Oct 2013.

[8] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," Jun. 2010.

[9] P. Matějka, O. Glembek, F. Castaldo, J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Černocký, "Full-covariance UBM and heavy-tailed PLDA in I-vector speaker verification," in *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011*. IEEE Signal Processing Society, 2011, pp. 4828–4831.

[10] N. Ryant and et al., "DIHARD Corpus. Linguistic Data Consortium." 2018.

[11] E. Bergelson, "Bergelson Seedlings HomeBank Corpus," 2016.

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5206–5210.

[13] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, 2017, pp. 2616–2620. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0950.html

[14] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus: A pre-announcement," in *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, ser. MLMI'05. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 28–39. [Online]. Available: http://dx.doi.org/10.1007/11677482_3

[15] M. VanDam, A. Warlaumont, E. BergelsonMartin, and M. A. Przybocki, "Homebank: An online repository of daylong child-centered audio recordings." in *Seminars in speech and language.*, vol. 37, no. 2, 2016, pp. 128–142.

[16] P. Matějka and et.al., "BUT-PT system description for nist lre," in *Proceedings of NIST Language Recognition Workshop 2017*. National Institute of Standards and Technology, 2017, pp. 1–6.

[17] P. Matějka, L. Burget, P. Schwarz, and J. Černocký, "Brno University of Technology System for NIST 2005 Language Recognition Evaluation," in *Proceedings of Odyssey 2006*, San Juan, PR, 2006.

[18] O. Plchot, L. Burget, H. Aronowitz, and P. Matějka, "Audio Enhancing With DNN Autoencoder For Speaker Recognition," in *Proceedings of the 41th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), 2016*. IEEE Signal Processing Society, 2016, pp. 5090–5094. [Online]. Available: http://www.fit.vutbr.cz/research/view_pub.php?id=11139

[19] "LIBRIVOX data," https://librivox.org/.

[20] "Aachen Impulse Response Database," http://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/aachen-impulse-response-database/.

[21] "C4DM (Center for Digital Music) RIR database," http://isophonics.net/content/room-impulse-response-data-set.

[22] R. Stewart and M. Sandler, "Database of omnidirectional and B-format room impulse responses," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 165–168.

[23] "Multichannel Acoustic Reverberation Database at York," http://www.commsp.ee.ic.ac.uk/ sap/resources/mardy-multichannel-acoustic-reverberation-database-at-york-database/.

[24] "OpenAir Impulse Response Database," http://www.openairlib.net/auralizationdb.

[25] "Reverb Challenge ," http://reverb2014.dereverberation.com/index.html.

[26] "RWCP Sound Scene Database," http://www.openslr.org/13/.

[27] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, Sept 2010.

[28] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Interspeech 2017*, Aug 2017.

[29] Kaldi, "SRE16 v2," https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2, [Downloaded: 2017-12].