INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Multi-PLDA Diarization on Children's Speech

Jiamin Xie[1], *Leibny Paola Garcia-Perera*[1], *Daniel Povey*[1,2], *Sanjeev Khudanpur*[1,2]

[1]Center for Language and Speech Processing &
[2]Human Language Technology Center of Excellence,
The Johns Hopkins University, Baltimore, MD, 21218, USA

jxie27@jhu.edu, lgarci27@jhu.edu, dpovey@gmail.com, khudanpur@jhu.edu

Abstract

Children's speech and other vocalizations pose challenges for speaker diarization. The spontaneity of kids causes rapid or delayed phonetic variations in an utterance, which makes speaker's information difficult to extract. Fast speaker turns and long overlap in conversations between children and their guardians makes correct segmentation even harder compared to, say a business meeting. In this work, we explore diarization of child-guardian interactions. We investigate the effectiveness of adding children's speech to adult data in Probabilistic Linear Discriminant Analysis (PLDA) training. We also train each of two PLDAs with separate objective to a coarse or fine classification of speakers. A fusion of the two PLDAs is examined. By performing this fusion, we expect to improve on children's speech while preserving adult segmentations. Our experimental results show that including children's speech helps reduce DER by 2.7%, achieving a best overall DER of 33.1% with the x-vector system. A fusion system yields a reasonable 33.3% DER that validates our concept.

Index Terms: speaker diarization, children's speech, probabilistic linear discriminant analysis

1. Introduction

Speaker diarization is a widely researched topic in the speech community, often referred to as the task of "who speaks when". It is a crucial first step that enables a single-speaker assumption necessary for downstream applications, including speaker verification and speech recognition, among others. Most diarization methods take short (1-2 sec) segments of a recording via a sliding window, identify the speaker(s) in each segment, and aggregate segments based on whether the same or different speaker is present. The process results in hypothesized segments of varying lengths belonging to each speaker.

Over the years, research in diarization has progressed from controlled microphone conditions like broadcast news to natural conditions such as telephone conversations or business meetings. The results of the work are also promising. The NIST rich transcription (RT) evaluation is an annual series of campaigns that systematically benchmarks diarization systems (e.g. [1, 2, 3]). In RT-04 for broadcast news, the diarization error rate (DER) ranges from 2% to 10% among the best systems [4]. Subsequently, in RT-09 for meeting recordings, the DER varies from 5.3% to 22.2% [5]. Though the DER in each condition may considered adequate, the increased magnitude and range of DER due to a shift in environment indicates that the diarization system is sensitive to noisy and vibrant settings.

Children's speech has been investigated in automatic speech recognition in studies that go from using GMM-HMM up to DNN based systems [6, 7, 8, 9]. However, diarization for children's speech has only recently started receiving attention. Recent work in [10] experimented with diarization on child-centered daylong recordings. The study reveals that children's recording is often accompanied by factors that are hard to isolate for diarization to work. For example, children lose their focus quickly and tend not to stay in one place for long. Thus, the recording involves various background noises and interactions between the child and people or objects around. The omnidirectional microphone is used to incorporate children's move-around, but can eventually pick up many unwanted acoustic events. In terms of diarization, recording of children's speech is covered with overlaps and non-speech events. The domain of children's speech even introduces a mismatch to the traditional diarization system which trains on adult speech from telephony or meetings. Thus, the intrinsic factors about children's recordings make diarization very difficult. The i-vector system built in [10] for child-centered recordings shows on average a 48.9% DER across different training datasets and methods. Moreover, the inclusion of kids data in training shows no significant improvements. In this research, we aim to extend this exploration to a more data-driven framework.

As mentioned, adult speech has been the focus for long. The DiHARD community challenge [11] opened a new dimension to the problem of diarization by including children's speech. From the competition, the worst result was surprisingly obtained on a sub-dataset (SEEDLingS [12]) in which children and adults interact. The SEEDLingS data are recorded by a proprietary device of the LENA^{TM} Foundation called DLP [13]. In a technical report from LENA [14], a segmentation accuracy between 71% and 82% is claimed between four general classes: "Adult", "Child", "TV", and "Other". However, the performance is restricted to using only the DLP device. The promising numbers are not guaranteed to be reproduced on general datasets. We study diarzation of SEEDLingS in this work.

In this spirit, we are tackling the problem by refining the models and diarization pipeline that are publicly available. One straightforward way is to incorporate children's speech data in the training of the i-vector or x-vector based system. The other way is to obtain better modeling for the Probabilistic Linear Discriminative Analysis (PLDA). We train two separate PLDAs for one coarser classification of the speaker categories and the other a finer classification of each speaker. Afterwards, the two PLDA scores are fused and an average is computed. Finally, the system performs the usual clustering with the averaged score and outputs the speaker labels for each segment.

The organization of the paper is as follows. Section 2 describes the main parts of our system. Section 3 outlines the steps of each experiment conducted. Section 4 describes the results. Section 5 opens the discussion of results and future directions for this research. Finally, Section 6 contains some conclusions from this work.

http://dx.doi.org/10.21437/Interspeech.2019-2961

2. System Description

In this section, we illustrate our i-vector and x-vector diarization system for children's speech. The baseline is an i-vector system to be compared with x-vector. The pipeline follows mainly the JHU system submitted for DiHARD 2018 [15] that is an open recipe in the KALDI toolkit [16]. The following paragraphs explain our systems in detail.

2.1. Pre-Processing

The audio data inflow to our system is all 16k-Hz sampled. Mel-frequency cepstral coefficients (MFCC) are used as features. Twenty-four cepstral coefficients are taken from 30 mel-frequency bins. The features are extracted over a 25ms window with a frame rate of 10ms. Both Δ and Δ-Δ features are appended. Cepstral mean normalization is applied.

The energy-based voice activity detection (VAD) is used to differentiate voiced frames from the silence. We apply VAD to both train and test data. However, the VAD keeps the non-speech events in the test data and gives no meaningful results, as explained later in section 4.4. Oracle speech activity detection (SAD) is used instead. Segmentation of the recording is performed as regards to the oracle time stamps. We refer to a similar take on the issue in [17].

2.2. Speaker Representation Extractor

After the pre-processing, each segment in the *test* and the *PLDA training* data is subsegmented by a 1.5s sliding window with a 0.75s overlap. The segments of extractor training data use a 3s sliding window and 10s hop. The speaker features are then extracted from the subsegments and length normalized [18]. The i-vector system uses a 2048-component UBM and a 400-dimensional representation. The x-vector system uses a 512-dimensional embedding.

2.2.1. i-vector

The i-vector [19] is a compressed representation of the speaker information. It has been the state-of-the-art for speaker and language recognition and diarization. This training approach is based on the idea of a new low-dimensional speaker- and channel-dependent space representation. The new space is known as the "total variability space" and encompasses the variabilities of the speaker and the channel. The formulation is defined as

$$M = s + T \cdot \omega, \tag{1}$$

where M and s is the supervector mean of the GMM and UBM, respectively. T is a low-rank rectangular matrix and ω is the identity vector (*i-vector*) with a normal distribution $N(0, I)$. The model projects an utterance onto the low-dimensional total variability space. For diarization, the model is learned by training on utterances from various speakers. The i-vector can then be extracted from each subsegment of the recording, compressing the speaker information within the subsegment.

2.2.2. x-vector

The x-vector [20, 21] is a low-dimensional speaker feature similar to the i-vector. But different from the formulated framework of i-vector, the x-vector is a DNN embedding. It is expected to encapsulate the most relevant and discriminative information of each speaker. For speaker recognition and diarization, a time-delayed neural network (TDNN) [22] is trained to discriminate between various speakers. The input to the TDNN are examples of frame-level features (MFCCs) with context. The objective function of the neural network uses a cross-entropy loss.

This DNN has two parts:

- A first part dedicated to collect frame level information
- A second part that models segment level information

The DNN is partitioned by a statistics pooling layer that computes the mean and standard deviation. Once trained, the softmax layer is removed and the x-vector embedding can be extracted from the layers after the pooling layer. The specific architecture is explained in greater detail here [20, 21].

2.3. PLDA

The PLDA is a generative model to handle *i-vectors* and *x-vectors*. It is first introduced for image recognition tasks [23] and extended to speaker recognition and diarization [24].

In its formulation, the supervector $r_{s,h}$ representing the i-vector of a single recording over a *channel* h by a *speaker* s is assumed to be composed of a collection of factors,

$$r_{s,h} = m + V \cdot y_s + \varepsilon_{s,h}, \tag{2}$$

where m is the class-independent mean of all the vectors, V is a low rank matrix defining the inter-class variability space, and $\varepsilon_{s,h}$ is the residual intra-class variability. The factor y_s follows a standard normal distribution and $\varepsilon_{s,h}$ follows a Gaussian distribution with zero mean and diagonal covariance.

For the PLDA scoring, we rely on a likelihood ratio of the form,

$$\theta(X) = \frac{p(r_s, r_{new}|H_1)}{p(r_s|H_0)\, p(r_{new}|H_0)}, \tag{3}$$

where r_{new} stands for the i-vector or x-vector from an unseen phrase, H_1 is the hypothesis of both vectors belonging to the same model (speaker), and H_0 represents the hypothesis of both vectors coming from different models.

2.4. Multi-PLDA Approach

The multi-PLDA approach is to train two PLDAs with separate objectives and fuse them. Here we refer each PLDA to a finer and coarser classification. The coarser classification is an easier task of the two. The goal is to compensate the results of the finer classifier for some high-level information of the speaker.

2.4.1. Finer classification

The finer classification (*finer-PLDA*) treats utterances of each distinct speaker as one class. This is a normal practice of using the PLDA in diarization. The utterances are labeled by the name of presented speaker and used to train the PLDA. In practice, the list of speaker to utterances mapping is provided to the training. As such, the PLDA can learn a projection space that maximizes the separation between speakers and minimizes variation within each speaker. The variation within each speaker can be caused by different channel effects.

2.4.2. Coarser classification

The coarser classification (*coarser-PLDA*) finds a general differentiation among speakers. Each utterance is labeled by the high-level categories that speakers belong to rather than the identities of speakers. For example, a natural utterance is commonly spoken by a female, male, or child. We map utterances to these three categories. The coarser classification becomes a simple task since more samples can be given to each class than in the finer classification. Thus, the *coarser-PLDA* can compensate some mistake made in the *finer-PLDA* through a fusion.

2.4.3. Fusion

The fusion process combines likelihood ratio scores obtained from the multiple PLDAs. The same strategy can also be found using in speaker verification or image recognition [25, 26].

From one PLDA, the pairwise scoring between features in each subsegments of a recording results a score matrix. Suppose the PLDA models P_f and P_c are trained by the finer and coarser classification accordingly. For a given utterance u, the score matrices $\mathbf{S}_u$ and $\mathbf{C}_u$ can be computed from P_f and P_c.

A fusion matrix $\mathbf{Q}_u$ can be written as,

$$\mathbf{Q}_{u|P_f,P_c} = a \cdot \mathbf{S}_{u|P_f} + (1-a) \cdot \mathbf{C}_{u|P_c} \tag{4}$$

where a is a scalar between 0 and 1. Through the parameter a, different weightings can be devised to join the discriminative power of multiple PLDAs in the fusion.

2.5. Agglomerative Hierarchical Clustering (AHC)

The last part of the system is a bottom-up agglomerative hierarchical clustering [27]. This is a greedy algorithm that starts with a number of clusters equal to the number of speech subsegments and merges the closest clusters iteratively until a stopping criterion is met. In the end, the system will assign a unique speaker label to all segments belonging to each cluster.

3. Experimental Setup

In this section, we describe the experimental setup of our study, the evaluation metric of the system is explained, and the datasets used are listed. The characteristics, partitions, and purposes of the datasets are illustrated as follows.

3.1. VoxCeleb

The *VoxCeleb1* and *VoxCeleb2* [28, 29] are two datasets that contain an on-line speech collection from interviews of celebrity. The speakers in the dataset are expected to be adults. A total of 7325 speakers are present. The *VoxCeleb1* training partition [28] and the entire *VoxCeleb2* data are used here.

3.2. CMU Kids Corpus

The *CMU Kids* corpus [30] is a dataset collected from children reading aloud. The age of the child varies from 5 to 10 years. Recordings from 75 speakers are used in our experiment.

3.3. CSLU Kids' Speech Corpus

The *CSLU Kids' speech* corpus [31] is recorded from children spontaneously saying simple words or sentences. The age among the speakers is between kindergarten and grade ten, which is similar to that of the *CMU Kids* corpus. Speech from 1116 speakers are used in our experiment.

3.4. Train Data Partitions

3.4.1. Training the extractor

VoxCeleb1&2 are combined and used for system training. In the i-vector system, 100k recordings are randomly selected from the combined set, pre-processed and used to train the UBM and T matrix. In the x-vector system, the combined set is augmented by reverberation and musical noise [21]. A set of 1 million recordings is sampled after data augmentation, and joined with the clean set to train the neural extractor.

3.4.2. Training the coarser- and finer-PLDA

The adult data is reduced to only *VoxCeleb1* here. The *finer-PLDA* is simply trained on the adult and all the kids data available, which is categorized by the unique speakers.

The *coarser-PLDA* is trained on three classes: *"female"*, *"kids"*, and *"male"*. We do not make a gender split of the "*kids*" speech: differences in vocal tract length are small at a young age [30]. The combined adult and kids data is partitioned in two ways as follows to be compared:

1. A balanced set with 190 files chosen in each category
2. A slightly unbalanced set with more *"kids"* data

The unbalanced set is composed of utterances from 1191 child speakers, 561 female speakers, and 690 male speakers. Note that all kids data are included in the unbalanced case.

3.5. SEEDLingS Corpus

The SEEDLingS corpus contains day-long speech recordings of 6-to-18 month olds in their home environment [12]. We use part of the SEEDLingS corpus provided in the DiHARD development dataset as our test set. A total of 23 files are available, each with a duration of about 5 minutes.

3.6. Diarization Error Rate

The primary evaluation metric in our experiments is diarzation error rate (DER). In our evaluation, *speech overlaps are included* and *no non-score collar* is used. The DER measures the cumulative duration of the following three types of errors: 1) false alarms – classifying non-speech as speech, 2) miss – classifying speech as non-speech, and 3) speaker match error – actual speaker differs from the labeled speaker.

4. Results

We conduct two main experiments. The first one studies the effectiveness of adding kids speech data to PLDA training. The second one tests the fusion concept for the multi-PLDAs. We also compare the results on training the *coarser-PLDA* by the balanced or unbalanced set. The system performance is evaluated using DER after a supervised calibration of the decision threshold. The main results from the two experiments are presented in Table 1 and Table 2, respectively. Details of the experiments are illustrated in subsections 4.1-4.3. We also point out our use of oracle speech segmentation, and discuss its impact in subsection 4.4.

4.1. Kids Speech in PLDA Training

Including the kids speech in the PLDA training data is shown to be effective for the x-vector system.

Table 1: *DER comparisons between using the finer-PLDA trained on only adults data and on both adults and kids data*

Feature Extractor	**VoxCeleb1**	**VoxCeleb1 & Kids**
i-vector	34.87%	36.71%
x-vector	35.89%	**33.15%**

As listed in Table 1, DER is reduced a notable 2.74% by training with additional kids data compared to training with only the adult data. The x-vector system with the *finer-PLDA* achieves the best overall performance with 33.15% DER[1]. The i-vector baseline shows no improvement from the inclusion.

4.2. Fusion PLDA

The *fusion-PLDA* is tested for only the x-vector system. The fusion is conducted separately for two setups which keep the *finer-PLDA* fixed and train the *coarser-PLDA* by either the balanced or unbalanced set. The computation follows the fusion equation (4). An average scheme ($a = 0.5$) is used.

Table 2: *DER comparisons between PLDAs before and after fusion for unbalanced (Unb.) and balanced (Bal.) training data*

	Finer-PLDA	**33.15%**
Train Data	**Coarser-PLDA**	**Fusion-PLDA**
Unb. Vox1 & Kids	37.05%	34.11%
Bal. Vox1 & Kids	35.76%	**33.29%**

In Table 2, we present DER results of each PLDA before and after the fusion. The fusion concept is shown to be valid. In the *coarser-PLDA* column, results show that training with balanced data has a lower DER than the unbalanced case. The same relationship is observed from results in the *fusion-PLDA* column. Thus, the performance difference between each coarser classifier is preserved before and after the fusion, which is expected.

However, no better overall result was seen from fusion. The reason is unclear, since the score is clustered by the AHC algorithm and gives the DER. The result of the *fusion-PLDA* in balanced training (33.29%) is close to the best overall (33.15%), so a better coarse-level classification would improve. In addition, the operating point of parameter a in Eqn. 4 is crucial as it determines the weighting of scores from the *finer-PLDA* and *coarser-PLDA* in the final score. The best value of parameter a can vary and shall be iteratively tuned for different setups with classifier choices, training data, and specific goal of the system.

4.3. Whitening Transformation and Balanced Training

The whitening transformation of the speaker features is a standard practice to improve diarization performance [18]. Different training data used in the *coarser-PLDA* and *finer-PLDA* can result two whitening transformation matrices to use.

Table 3: *DER comparisons among different choices of whitening transformation matrix in the fusion case*

	Fusion-PLDA	
Whitening Matrix	Unbalanced	Balanced
Coarser- & Finer- mixed	36.23%	34.23%
Coarser-PLDA	35.57%	35.20%
Finer-PLDA	34.11%	**33.29%**

In Table 3, results from each *mix & match* combination of the transformation matrix are compared. The best result (33.29%) is obtained when the training data of each PLDA in the fusion is whitened by the training set of the *finer-PLDA*. In addition, the results of balanced training the *coarser-PLDA* leads the unbalanced case by an average 1.06% DER of different setups. The balance in the training data among categories of the *coarser-PLDA* is shown to be important.

4.4. Oracle Speech Segmentation

A variety of non-speech events were observed in SEEDLingS since they are recorded in the daily home environment. This includes kids shouting, crying, laughing, among other vocalizations. The background noises are vibrant such as microphone scratches, toys drop, cartoons playing on tablets or TVs. The traditional energy-based VAD acted poorly on removing the non-speech events of high acoustic energy. The oracle speech segmentation, on the other hand, provides us a meaningful way to focus on the speaker diarization of children's speech. The worst case scenario on DER using the oracle speech segmentation is found to be 37.05%, where all segments are assigned to one speaker. Though a DNN-based speech activity detection (SAD) model is available in [16] developed for the ASpIRE project, it removes the kids speech that appears short or incomplete. Therefore, we have used oracle SAD in this work, and look forward to substantial work on the SAD that can incorporate kids speech.

5. Discussion

We have seen a DER reduction achieved by including children's speech (5-11 years old) to the adult training data of PLDA in the *x-vector* approach. There still remains an age difference between the train and our test dataset, SEEDLingS (6-18 months old). We need to study the effect of children's age and explore if addressing the difference can improve our results. We have observed a close-to-the-best result obtained by the *fusion-PLDA*. However, there is room for overall improvement. The following enhancements on the multi-PLDA system need to be investigated , 1) a better coarse-level classifier, e.g. DNN, 2) more diverse training data for the coarse-level classifier, and 3) an iterative tuning to find the best parameter a of a particular setup. In addition, the speaker diarization on children's speech shows to be a complex scenario, where multiple events occur simultaneously. To use this event information could help a better diarization. Much work needs to be done on the SAD, overlap, noisy condition, etc. The manual annotation does not always match the real segmentation. A smarter way to clean these annotations may help us gain a clearer analysis of our model.

6. Conclusions

In this paper, we presented work on diarzation of children's speech. Training the PLDA with additional children's speech and the normal adult data has shown to be effective in the *x-vector* diarization approach. The inclusion of kids data appears to give an adequate gain in performance. We developed a multi-PLDA diarization system. The concept is to train a finer and a coarser PLDA (classification) and fuse them. The coarse classification seems complementary to the fine one when fused, and a close-to-the-best result was observed by the fusion PLDA.

[1] We also tested the system with the DiHARD evaluation set. The mixed training data is expected to cause degradation on the adult speech tests. The DER result (29.85%) of mixed training is shown to be close to the result (29.34%) of using only the adult data. A better result (29.21%) is obtained by including the kids data in the DNN extractor.

7. References

[1] J. G. Fiscus, N. Radde, J. S. Garofolo, A. Le, J. Ajot, and C. Laprun, "The rich transcription 2005 spring meeting recognition evaluation," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 369–389.

[2] J. G. Fiscus, J. Ajot, and J. S. Garofolo, "The rich transcription 2007 meeting recognition evaluation," in *Multimodal Technologies for Perception of Humans*. Springer, 2007, pp. 373–389.

[3] N. Fall, "rich transcription (rt-04f) evaluation plan, august 2004," 2004.

[4] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, Sep. 2006.

[5] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, Feb 2012.

[6] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[7] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[8] S. Ghai and R. Sinha, "A study on the effect of pitch on lpcc and plpc features for children's asr in comparison to mfcc," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[9] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[10] A. Cristià, S. Ganesh, M. Casillas, and S. Ganapathy, "Talker diarization in the wild: the case of child-centered daylong audio-recordings," in *Interspeech*, 2018.

[11] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First dihard challenge evaluation plan," 2018.

[12] E. Bergelson, "Bergelson seedlings homebank corpus," *doi*, vol. 10, p. T5PK6D, 2016.

[13] D. Xu, U. Yapanel, and S. Gray, "Reliability of the lena (tm) language environment analysis system in young children's natural home environment (lena technical report ltr-05-2)," *LENA Foundation," Boulder, CO2009, Available: http://lena. org/wp-content/uploads/2016/07/LTR-05-2_Reliability. pdf*, 2009.

[14] M. Ford, C. T. Baer, D. Xu, U. Yapanel, and S. Gray, "The lenatm language environment analysis system: Audio specifications of the dlp-0121," *Boulder, CO: Lena Foundation*, 2008.

[15] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe *et al.*, "Diarization is hard: Some experiences and lessons learned for the jhu team in the inaugural dihard challenge," in *Proc. Interspeech*, 2018, pp. 2808–2812.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

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

[18] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Twelfth annual conference of the international speech communication association*, 2011.

[19] 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.

[20] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017, pp. 999–1003.

[21] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[22] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[23] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[24] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." in *Odyssey*, vol. 14, 2010.

[25] X. Pang and M.-W. Mak, "Fusion of snr-dependent plda models for noise robust speaker verification," in *The 9th International Symposium on Chinese Spoken Language Processing*. IEEE, 2014, pp. 619–623.

[26] V. Štruc, N. Pavešić, J. Žganec-Gros, and B. Vesnicer, "Patch-wise low-dimensional probabilistic linear discriminant analysis for face recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 2352–2356.

[27] D. A. V. Leeuwen, "Speaker linking in large datasets," in *Odyssey*, 2010, pp. 202–208.

[28] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[29] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[30] M. Eskenazi, J. Mostow, and D. Graff, "The cmu kids corpus," *Linguistic Data Consortium*, 1997.

[31] K. Shobaki, J.-P. Hosom, and R. Cole, "Cslu: Kids speech version 1.1," *Linguistic Data Consortium*, 2007.