Multi-PLDA Diarization on Children's Speech

Jiamin Xie, Leibny Paola Garcia-Perera, Daniel Povey, Sanjeev Khudanpur

Center for Language and Speech Processing, Johns Hopkins University

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

Motivation

The research in diarization of children's speech has been only explored recently. The work in [3] tackled the problem through an *i-vector* approach, but discovered poor performance that is inconsistent with the adult scenario. The recording of children's speech is filled with vocalizations, media playing in background, and speech overlaps. These factors, perhaps intrinsic to children, hinder a good diarization performance. In this research, we aim to modify components of the developed diarization pipelines to improve diarization of children's speech.

Main Objectives

1. Explore data-driven *x-vector* framework [9] in comparison to an *i-vector* [4] baseline
2. Incorporate children's speech data into training process to reduce domain mismatch
3. Train two PLDAs on the *coarse-level* classification of speaker types and *fine-level* classification of speaker identities, and perform a score fusion

System Description

The diarization systems is a modified version of the JHU system in Kaldi, which participated in the Dihard 2018 [7]. The components of a general pipeline are illustrated in Figure 1. The speaker embedding extractor can either be the GMM/T matrix of *i-vector* or the TDNN of *x-vector*.

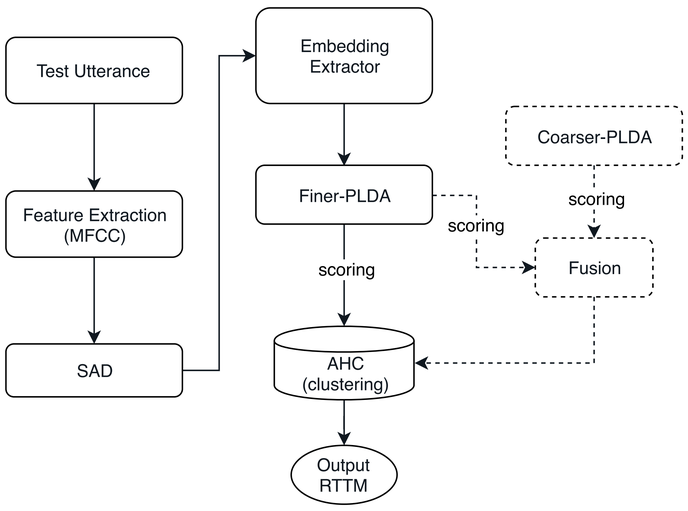

Figure 1: A general diarization pipeline (- - - is the alternative path for scoring)

Multi-PLDA and Fusion

The multi-PLDA approach trains two PLDAs with different objectives. We refer each PLDA as a coarser and finer classification. The former is an easier task of the two. The goal is to compensate the results of the finer classifier with some high-level information of the speaker.

- **Coarser classification**

  The *coarser-PLDA* space maximizes separation between speaker types. Assume voiced segments contain one speaker, then the voice can only be from a **male, female, or child**. We map speakers into these three categories and train the coarser-PLDA with labels.

- **Finer classification**

  The *finer-PLDA* space maximizes variations among speaker's identity. This is the conventional practice of PLDA in diarization.

- **Fusion**

  The fusion of two PLDA models, $\mathcal{M}_{fine}$ and $\mathcal{M}_{coarse}$ is captured in the equation below. The scoring is a likelihood ratio test between a segment coming from same and different speaker model. For an utterance u, the score matrices $\mathbf{S}_u$ and $\mathbf{C}_u$ can be computed from $\mathcal{M}_f$ and $\mathcal{M}_c$. A fusion score matrix $\mathbf{Q}_u$ can be written as,

$$\mathbf{Q}_{u|\mathcal{M}_f,\mathcal{M}_c} = a \cdot \mathbf{S}_{u|\mathcal{M}_f} + (1-a) \cdot \mathbf{C}_{u|\mathcal{M}_c} \tag{1}$$

  where a is a scalar weighting parameter between 0 and 1. We averaged scores with $a = 0.5$.

Data and Experimental Setup

The *VoxCeleb1* and *VoxCeleb2* are speech extracted from YouTube videos of interviews [6, 2]. There are 7325 adult speakers. The *CMU Kids Corpus* is collected from children reading aloud [5]. There are 75 speakers present. The *CLSU Kids' Speech Corpus* [8] is recorded from children spontaneously saying simple words or sentences. There are 1116 speakers. The age of kids is similar for both datasets (5 ˜10 years old). We combine CMU and CSLU corpus together as the kids speech data.

	VoxCeleb1	*VoxCeleb2*	*CMU Kids Corpus*	*CSLU Kids' Speech*	*Seedlings*
Extractor	✓	✓	✗	✗	✗
PLDA (each)	✓	✗	✓	✓	✗
Test	✗	✗	✗	✗	✓

Table 1: Data used in different system components

We conduct two main experiments and evaluate the performance on the Seedlings dataset [1]:

- ⋆ Compare the *x-vector* and *i-vector* systems with/without adding the kids training data
- ⋆ Compare using *coarser-PLDA*, *finer-PLDA* and a fusion of the two (*fusion-PLDA*)

	Embedding dimension	Num of GMM components	Window size (s)	Shift size (s)
i-vector	400	2048	1.5	0.75
x-vector	512	N/A	1.5	0.75

Table 2: Parameters of i-vector and x-vector system

The MFCC features with 24 cepstrals are extracted from the 16-kHz sampled audio input. The Δ and Δ-Δ features are appended. The cepstral mean normalization and oracle speech activity detection (SAD) are applied. The voiced segments are then sub-segmented to extract speaker embeddings. Table 2 lists extraction setup of the speaker embeddings.

Results

Results are evaluated by the diarization error rate (DER), which counts the total of missed, false alarm speech and speaker match error. In our evaluation, *speech overlaps are included* and *no non-score collar* is used.

- ∗ Inclusion of the kids' speech data in the PLDA training is shown to be effective for the *x-vector* system, reducing DER of the best *i-vector* baseline by 1.72%.

Speaker Embedding	**VoxCeleb1**	**VoxCeleb1 & Kids**
i-vector	34.87%	36.71%
x-vector	35.89%	**33.15%**

Table 3: DER change before and after including kids' data in training

- ∗ The *fusion-PLDA* taking an average between the finer and coarser scores is shown to be valid. The balanced data is shown effective for the *coarser-PLDA*.

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

Table 4: DER change from fusion under a balanced or unbalanced training

Discussion & Future Work

- ◇ Tuning of the parameter a can help achieve better result for the *fusion-PLDA*.
- ◇ Diverse training data and neural network may boost the *coarser* classification.
- ◇ Oracle SAD is used to give meaningful results. A conventional SAD will not be able to handle the vocalizations and etc. More work is needed for better SAD.
- ◇ An age difference still exists between train (5-10 years) and test (6-18 months). To analyze the effects will be helpful for further research.

Conclusions

- A data-driven framework is presented for diarization of children's speech
- Inclusion of kids' data in training shows improvement over using adult data
- Fusion of a coarser and finer PLDA is experimented and results are studied

References

[1] E Bergelson. "Bergelson seedlings homebank corpus". In: *doi* 10 (2016), T5PK6D.

[2] J. S. Chung, A. Nagrani, and A. Zisserman. "VoxCeleb2: Deep Speaker Recognition". In: *INTERSPEECH*. 2018.

[3] Alejandrina Cristià et al. "Talker Diarization in the Wild: the Case of Child-centered Daylong Audio-recordings". In: *Interspeech*. 2018.

[4] N. Dehak et al. "Front-End Factor Analysis for Speaker Verification". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (May 2011), pp. 788–798. ISSN: 1558-7916. DOI: 10.1109/TASL.2010.2064307.

[5] Maxine Eskenazi, Jack Mostow, and David Graff. "The CMU kids corpus". In: *Linguistic Data Consortium* (1997).

[6] A. Nagrani, J. S. Chung, and A. Zisserman. "VoxCeleb: a large-scale speaker identification dataset". In: *INTERSPEECH*. 2017.

[7] Gregory Sell 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.

[8] Khaldoun Shobaki, John-Paul Hosom, and Ronald Cole. "CSLU: Kids' speech version 1.1". In: *Linguistic Data Consortium* (2007).

[9] David Snyder et al. "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.