Mixture of Speaker-type PLDAs for Children’s Speech Diarization

Jiamin Xie\textsuperscript{1}, Suzanna Sia\textsuperscript{1}, Paola García\textsuperscript{1,2}, Daniel Povey\textsuperscript{3}, Sanjeev Khudanpur\textsuperscript{1,2}

\textsuperscript{1}Center for Language and Speech Processing & \textsuperscript{2}Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA
\textsuperscript{3}Xiaomi Corp., Beijing, China
{jxie27, ssia1, lgarci27}@jhu.edu, dpovey@gmail.com, khudanpur@jhu.edu

Abstract

In diarization, the PLDA is typically used to model an inference structure which assumes the variation in speech segments be induced by various speakers. The speaker variation is then learned from the training data. However, human perception can differentiate speakers by age, gender, among other characteristics. In this paper, we investigate a speaker-type informed model that explicitly captures the known variation of speakers. We explore a mixture of three PLDA models, where each model represents an adult female, male, or child category. The weighting of each model is decided by the prior probability of its respective class, which we study. The evaluation is performed on a subset of the BabyTrain corpus. We examine the expected performance gain using the oracle speaker type labels, which yields an 11.7% DER reduction. We introduce a novel baby vocalization augmentation technique and then compare the mixture model to the single model. Our experimental result shows an effective 0.9% DER reduction obtained by adding vocalizations. We discover empirically that a balanced dataset is important to train the mixture PLDA model, which outperforms the single PLDA by 1.3% using the same training data and achieving a 35.8% DER. The same setup improves over a standard baseline by 2.8% DER.

Index Terms: speaker diarization, children’s speech, transformer encoder, mixture of PLDAs

1. Introduction

Speaker diarization aims to answer the question of “who speaks when?” in a recording. It is the crucial first step that ensures the single-speaker assumption necessary for downstream tasks, including speaker verification and speech recognition, among others. Most diarization methods take short (1-2 sec) overlapping segments of a recording, estimate similarities between each pair of segments, and cluster/separate segments to same/different speaker(s). The process results in hypothesized speech segments of various lengths that belong to each speaker.

Research in diarization has mainly focused on adult speech. The benchmark diarization error rates (DER) in controlled speech environment, such as telephone conversations or business meetings, typically range from 2% to 10% among the best systems \[11,12]\). With more natural conditions, such as telephone conversational speech, the diarization performance varies between 5% and 30% DER \[3,4]\). However, these results are made possible under a rather easy setup of few number of speakers or with similar speaking style shared by participants. Recent studies have found diarization is hard under realistic conditions that involve overlapping speech, noisy background, and diverse modalities of speech \[5,6]\).

Children’s speech is one of the realistic domains that poses challenges for speaker diarization \[7\]. The acoustic and linguistic properties of children differ from that of adults, such as higher pitch and formant frequencies and longer phoneme duration \[8\]. In addition, children utter spontaneous vocalizations during their speech, which increases the need for intra-speaker variations to be appropriately modeled. These spontaneous vocalizations can also occur when others are speaking, which calls for overlap diarization. The above factors introduce a large performance discrepancy between the diarization system of children’s speech and adult speech. One of the studies which analyze the language exposure of children in a home environment revealed on average a 48.9% DER performance across different training datasets and the state-of-the-art diarization systems \[9\].

Our previous work \[10\] focused on the adaptation of a PLDA model through both children’s speech data augmentation and a discrimination of speaker representations by adult female, adult male, and child type of speakers. In this paper, we extend the later idea to a mixture PLDA model that explicitly captures the variation of speakers across speaker types. The organization of the paper is as follows. Section 2 addresses the related work and inspiration of mixture PLDA model. Section 3 describes the main methods. Section 4 outlines the experimental setup and data preparation. Section 5 presents the results. Finally, Section 6 concludes this work and mentions future work.

2. Related Work

Child speech has long been studied for automatic speech recognition (ASR) in \[11,12\]. Early development of diarization system on children’s speech has focused on child language acquisition analysis \[13\] or proprietary smart home devices \[14,15\]. The diarization of four classes among a primary child, secondary child, adult, and non-speech was first explored in \[13\]. The speaker-independent DNN-HMM system \[13\] achieved around 20% DER per child category. However, disentangling the non-speech events and children’s speech remains a challenge, as about 10% of the true child speech was misclassified to be non-speech. One recent work in \[16\] studied the speech enhancement of the noisy environment in daylong realistic recordings, including the SEEDLinkS \[17\], a child-centered dataset. The proposed LSTM-based enhancement preprocessor with a built-in diarization system achieved a 39.2% DER \[18\]. The mixture PLDA model has been mainly studied for speaker verification tasks \[19,20\]. The work of \[20\] showed a better performance using the mixture of two gender-dependent PLDAs than a single gender-independent PLDA. Our work extend from \[20\] to three classes of speakers and further develops the mixture of PLDAs for diarization.

3. Methods

In this section, we illustrate the methods to incorporate speaker type information to diarization. Subsection 3.1. explains an
ideal segmentation step that takes account of oracle speaker type labels. Subsection 3.2, explains the concept of the mixture of PLDA models that encompasses speaker type priors. Lastly, subsection 3.3 describes an estimator of speaker type confidence scores.

3.1. Speaker Type Segmentation

The speaker type segmentation refers to the process which splits speech into three parts that each belongs to an adult female, adult male, or child class. Diarization is subsequently performed on each speech region of a class. As illustrated on the left of figure 1, each of the speech splits then goes through the uniform segmentation and scored by a PLDA model trained on the data from the corresponding speaker type. Since the speaker types are mutually exclusive, the diarization proposals of speakers in each of the speech splits will be differentiated, i.e., the speaker1 of female speech is not the speaker1 of male speech. This is of course an ideal setup if provided with the gold speaker type labels. We find that the empirical performance using predicted labels from a classifier is worse compared to a standard baseline because the confusion made early between speaker types can cause the wrong assignments of speakers later.

3.2. Speaker Type Informed Scoring

To prevent the hard assignment of speaker types, a probabilistic framework is considered. The similarity scoring in diarization is performed because the confusion made early between speaker types is figuratively shown on the right of figure 1.

Under a mixture PLDA model, the numerator of equation (1) can be written as a convex combination of speaker-type dependent PLDA models,

\[
P(z_1, z_2 | H_d) = \sum_{g_1 \in G} P(g_1, g_2 | H_d) \sum_{g_2 \in G} P(z_1, z_2 | g_1, g_2, H_d)
\]

(3)

where \(g_1\) and \(g_2\) are the speaker types in \(G = \{M', F', C'\}\) corresponding to \(z_1\) and \(z_2\), and \(M', F', C'\) are the adult male, adult female, and child speaker types, respectively. The denominator of equation (1) can be written as

\[
P(z_1, z_2 | H_d) = \sum_{g_1 \in G} P(g_1, g_2 | H_d) P(z_1, z_2 | g_1, g_2, H_d)
\]

(4)

where there are 9 terms in the denominator (all pairwise combinations of speaker types). Given the same speaker, both \(g_1\) and \(g_2\) must belong to the same speaker type,

\[
P(g_1, g_2 | H_d) = P(g_1) = P(g_2)
\]

\[
P(z_1, z_2 | g_1, g_2, H_d) = P(z_1 | \psi_{g_1}) \times P(z_2 | \psi_{g_2})
\]

Under different speakers,

\[
P(g_1, g_2 | H_d) = P(g_1) \times P(g_2)
\]

\[
P(z_1, z_2 | g_1, g_2, H_d) = P(z_1 | \psi_{g_1}) \times P(z_2 | \psi_{g_2})
\]

where \(\psi_{g_1}\) and \(\psi_{g_2}\) are parameters of the \(g_1\)-type PLDA and \(g_2\)-type PLDA, respectively. Here, we assumed the distribution of a speaker type is independent under the different-speaker condition \(H_d\). The single likelihood \(P(z_1 | \psi_{g_1})\) or the joint likelihood \(P(z_1, z_2 | \psi_{g_1})\) can be obtained from the single PLDA model as described in [22]. But the prior distribution \(P(g)\) is a design choice to make.

3.3. Speaker Type Confidence Estimator

As explained in the previous section, the prior distribution of a speaker type is key to the mixture PLDA formulation. One simple way to adjust is to assume a constant prior for all recordings encountered in the evaluation. For instance, the prior of child speakers should be above the uniform threshold of 0.33 for diarization on children’s speech. We illustrate briefly the other approach to estimate the speaker type confidence from frame-level features.
3.3.1. Problem Formulation

Our goal is to obtain an informed prior probability $P(g)$ for the mixture PLDA. We can consider to use the posterior $P(g|X)$ given an input feature sequence $X$. The confidence estimates for each speaker type can be adopted by taking the softmax output of a neural network $g$. The prior distribution $P(g)$ that a speech segment $z_i$ belongs to a speaker type $g$ for $\forall i \in \{1, 2\}$ can be approximated by,

$$P(g_i) \approx P_{nn}(g_i|z_i) = \frac{1}{T_i} \sum_{t=0}^{T_i} P_{nn}(g_i|X)$$

where $z_i$ has $T_i$ frames and $P_{nn}(g_i|X)$ is the frame-wise posterior output of the network given the whole input sequence. We experimented with this using various sequence-to-sequence and Transformer architectures $[23, 24]$, but found that although such a trained system can predict the correct speaker type label with around 75% accuracy, the performance gains are not transferred when used as mixture PLDA weights, motivating future work on calibration of neural network output probabilities.

4. Experimental Setup

The experimental setup is illustrated in this section.

4.1. System Description

Our diarization system mainly follows the x-vector-based system from the DiHARD 2018 $[25]$ recipe in Kaldi $[26]$. We focus on extending the PLDA model within this pipeline. The audio data input is sampled at 16k-Hz. Mel-frequency cepstral coefficients (MFCC) are used as features and 30 cepstral coefficients are taken from 30 mel-frequency bins. The features are extracted over a 25ms window with a frame rate of 10ms. Both the Delta and the Delta-Delta features are appended. Cepstral mean normalization is applied over a sliding window up to 3 seconds. After the pre-processing, each segment in the evaluation and the PLDA training data is subsegmented by a 1.5s sliding window with a 0.75s overlap. The speaker features are then extracted from the subsegments and length normalized $[27]$. The x-vector embedding has 512 dimensions.

4.2. Datasets

4.2.1. Adult speech

We use the VoxCeleb $[28, 29]$ datasets for the adult speech. The VoxCeleb1 and VoxCeleb2 $[28, 29]$ are two versions of a large scale dataset that contains interview videos of celebrities uploaded to YouTube. The speakers in the dataset are expected to be mainly adults. There are a total of 7325 speakers with 61% being male and 39% being female. We filter by gender in each dataset to train individual PLDA models of adult speaker types.

4.2.2. Child speech

The CMU Kids $[30]$ and the CSLU Kids corpus $[31]$ are used for child speech training. Both datasets were collected for speech recognition tasks, so the audio quality is considered clean. The age of children from both datasets cover a range from 5 to 10 years old. The combined set contains 1191 speakers and about 42.6 hours of speech. The average duration of an utterance is about 4 seconds long.

4.2.3. Baby vocalization

We use a subset of the data provided in the Interspeech ComParE challenge $[32]$ as the augmentation dataset, which collects mostly baby crying sounds. The dataset contains 5.6k recordings with about 2.8 hours of baby vocalizations. We highlight that the average duration of a recording is only 1.8 seconds long, which makes this small dataset hardly sufficient to train models on baby speakers alone.

4.2.4. BabyTrain test

The babyTrain is a newly aggregated dataset of 9 child-centered corpus $[17]$ with daylong recordings in the home environment. The train, dev, and test split of the dataset is prepared by the JSALT 2019 workshop $[6]$ and covers a total of 270 hour recordings. The age of children in the dataset varies between 5 to 60 months old. We adopt the provided dataset splits of 57.5% train, 27% development, and 15.5% test set of the total audio length. The oracle mapping of speakers to categories of key child, child, adult female, adult male, and others is provided. The distribution of speaker type in train is similar to the test, which about 46% is child, 50% is female, and 4% is male. We exclude recordings where the distant speakers are annotated but with an undefined speaker type label. This leaves us with 329 out of 413 files.

4.3. Baseline Setup

Our baseline system uses the single PLDA model trained on the VoxCeleb1, and CMU Kids and CSLU Kids corpus.

4.4. Speaker Type Experiment

To obtain an estimate of the upper bound on the performance of the speaker type informed diarization system, we evaluate the system based on the oracle speaker type labels, that is $P(F) = 1$ when the speaker for the speech segment $z$ is Female. This effectively shifts the responsibility of likelihood ratio scoring to one PLDA model trained on the true speaker type.

4.5. Mixture of Speaker-type PLDAs

The mixture PLDA is compared to the single PLDA model in the evaluation. Both models are trained on the same dataset, where we further split the data by speaker types to train the mixture PLDA. To study the influence of a data imbalance, we either use the whole dataset or randomly select 1000 speakers from each speaker type to compose the mixture PLDA. We further compare between a nonuniform and uniform prior of speaker types, where the nonuniform distribution on female, child, and male is 40%, 40%, and 20%, respectively.

4.6. Evaluation Metric

The primary evaluation metric of our experiments is the diarization error rate (DER). We score speech overlaps and do not use non-score collar. The DER measures a cumulative duration of the following three types of errors over a total duration of valid scoring regions,

1. False alarm (FA) classifying non-speech as speech
2. Miss (MS) classifying speech as non-speech
3. Speaker mismatch (SM) actual speaker differs from the claimed speaker
5. Results

We conduct three main experiments. The first one examines the upper bound of performance gain using the oracle speaker type segmentation. The second one studies the effectiveness of baby-vocalization augmentation. The last one evaluates the single (UniPLDA) and mixture (MixPLDA) model as well as the influence of training data balance. The system performance is evaluated using DER under the gold number of speakers and oracle speech activity detection. The main results from the three experiments are presented in Table 1, Table 2, and Table 3. Details of the experiments are illustrated in section 4.

5.1. Performance Upper Bound on using Speaker Types

To study the benefit of speaker type information, the evaluation recording is split into three speech regions of each speaker type, using the oracle label. We instead had to use a score threshold to stop the clustering since the number of speakers in each speaker-type segmented audio is unknown. Shown in Table 1

| UniPLDA Baseline | Oracle Speaker Type | Same Speaker |
|------------------|---------------------|--------------|
| 39.90 (±0.2)     | 28.20 (±0.1)        | 40.26 (+)    |

Table 1: DER(%) (threshold) comparison between baseline and oracle speaker type

using the oracle speaker type reduces the UniPLDA baseline by a significant 11.7% DER. This verifies our claim that extra speaker type information is beneficial for diarization. The last entry shown in Table 1 illustrates the worst scenario when the system outputs only one speaker. This result also implies the dominant speaker accounts for about 60% of the speech (~40% DER), and of the remaining 40% belongs to other speakers.

5.2. Baby Vocalization Augmentation

The baby vocalization augmentation is found to be an effective domain adaptation of the child-type PLDA. We apply different augmentation techniques to the clean children’s speech and compared the results in Table 2

| System | Clean | mn | v | vn | vmus |
|--------|-------|----|---|----|------|
| CHI-PLDA | 39.61 | 39.28 | 38.74 | **37.70** | 39.51 |

Table 2: DER(%) comparison between clean and different augmentations of the child-type PLDA. The music, noise, vocalization, and MUSAN augmentation are labeled accordingly as mn, v, vn, and mus.

As shown from above, using the vocalization (column 4) alone is found to reduce 0.9% DER from the clean baseline. Compared to the gains from adding double-sized or triple-sized samples generated by music and noise augmentation (column 3) or the MUSAN augmentation (column 6), the vocalization seems to provide the most matched information to the target domain. Lastly, we find the vocalization is complementary with the noise augmentation, which further reduces the clean baseline by an absolute 1.9% DER.

5.3. Mixture PLDA and Data Balance

We observe combining the kids data and VoxCeleb to train the UniPLDA baseline achieves a 38.62% DER. Our proposed mixture of PLDA models with uniform weights (33%) on each speaker type, ‘Male’, ‘Female’, and ‘Child’ has a very comparable performance with the UniPLDA. However, with a matched estimation on the speaker type prior to the evaluation, the MixPLDA outperforms the UniPLDA model (row 3 and 5), showing the potential of the speaker type informed model. The cost of using a mixture model close to the single model since training three models in parallel is possible. The value of the constant prior may be elicited from human expert knowledge, but making either a manual inspection from sampling or an intuitive estimate should be sufficient.

We find the importance of keeping the data balanced while training the MixPLDA. Comparing row 2 to 3 or row 4 to 5 in Table 3 we find on average 1.9% DER improvement in the nonuniform MixPLDA, even though data size is reduced from the balancing operation. On the contrary, the performance of the UniPLDA depends heavily on the size of the data (the largest is the vmus, the second goes the vn, and baseline is the least). The formulation of [20] shows the likelihood ratio obtained under the MixPLDA is equivalent to a weighted sum of each likelihood ratio scored obtained from one of the mixed PLDAs. We suspect this may explain why data balancing is helpful since similarly constrained data can limit the modeling space that each PLDA covers.

6. Conclusion and Future Work

In this paper, we presented a diarization framework using the mixture PLDA model that is targeted at children’s speech domain. We discovered speaker type information is beneficial and verified a large upper bound of improvement. Empirically, the mixture of speaker-type PLDA models outperforms the single PLDA model when a balanced training data is used. Though the best result is obtained by the single PLDA, the best mixture PLDA system comes close with an absolute 0.2% difference in DER and a half of the training size required. Using baby vocalizations as additive background noises has shown matching to the age and acoustic condition of children’s speech. Our future work is to develop a confidence score estimator of speaker types using neural networks, as illustrated in the paper.

7. Acknowledgements

The authors would like to thank Jess Villalba for the constructive discussion on the mixture PLDA formulation.
8. References

[1] 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.

[2] 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.

[3] A. McCoy, G. Sell, and D. Garcia-Romero, “Speaker diarization using leave-one-out gaussian plda clustering of dnn embeddings,” Proc. Interspeech 2019, pp. 381–385, 2019.

[4] M. Senoussaud, P. Kenny, T. Stufylakis, and P. Dumouchel, “A study of the cosine distance-based mean shift for telephone speech diarization,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 1, pp. 217–227, 2013.

[5] S. Watanabe, M. Mendel, J. Barker, and E. Vincent, “Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings,” arXiv preprint arXiv:2004.09249, 2020.

[6] P. García, J. Villalba, H. Bredin, J. Du, D. Castan, A. Cristia, L. Bullock, L. Guo, K. Okabe, P. S. Nidadavolu et al., “Speaker detection in the wild: Lessons learned from jslt 2019,” arXiv preprint arXiv:1912.09038, 2019.

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

[8] M. Gerosa, S. Lee, D. Giuliani, and S. Narayanan, “Analyzing children’s speech: An acoustic study of consonants and consonant-vowel transition,” in 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol. 1. IEEE, 2006, pp. 1–4.

[9] A. Cristia, S. Ganesh, M. Casillas, and S. Ganapathy, “Talker diarization in the wild: the case of child-centered daylong audio recordings,” in Interspeech, 2018.

[10] J. Xie, L. P. Garcia-Perera, D. Povey, and S. Khudanpur, “Multiple plda diarization on childrens speech,” 2019.

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

[12] S. Ghai and R. Sinha, “A study on the effect of pitch on lpc and plpc features for children’s asr in comparison to mfcc,” in Twelfth Annual Conference of the International Speech Communication Association, 2011.

[13] M. Najafian and J. H. Hansen, “Speaker independent diarization for child language environment analysis using deep neural networks,” in 2016 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2016, pp. 114–120.

[14] D. Xu, U. Yapanel, and S. Gray, “Reliability of the lena language environment analysis system in young childrens natural home environment,” Boulder, CO: LENA Foundation, pp. 1–16, 2009.

[15] 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.

[16] L. Sun, J. Du, T. Gao, Y.-D. Lu, Y. Tsao, C.-H. Lee, and N. Ryant, “A novel lstm-based speech preprocessor for speaker diarization in realistic mismatch conditions,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5234–5238.

[17] M. VanDam, A. S. Warlaumont, E. Bergelson, A. Cristia, M. Soderstrom, P. De Palma, and B. MacWhinney, “Homebank: An online repository of daylong child-centered audio recordings,” in Seminars in speech and language, vol. 37, no. 02. Thieme Medical Publishers, 2016, pp. 128–142.

[18] D. Vijayasenan and F. Valente, “Diar: An open source toolkit for research in multistream speaker diarization and its application to meetings recordings,” in Thirteenth Annual Conference of the International Speech Communication Association, 2012.

[19] M.-W. Mak, X. Pang, and J.-T. Chien, “Mixture of plda for noise robust i-vector speaker verification,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 1, pp. 130–142, 2015.

[20] M. Senoussaud, P. Kenny, N. Brümmer, E. d. Villiers, and P. Dumouchel, “Mixture of plda models in i-vector space for gender-independent speaker recognition,” in Eleventh Annual Conference of the International Speech Communication Association, 2011.

[21] 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.

[22] S. Ioffe, “Probabilistic linear discriminant analysis,” in European Conference on Computer Vision. Springer, 2006, pp. 531–542.

[23] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, “Transformer-xl: Attentive language models beyond a fixed-length context,” arXiv preprint arXiv:1901.02860, 2019.

[24] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in Advances in neural information processing systems, 2014, pp. 3104–3112.

[25] G. Sell, D. Snyder, A. McCoy, 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 Interspeech, vol. 2018, 2018, pp. 2808–2812.

[26] 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.

[27] D. García-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.

[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.

[32] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. B. Pokorny, E.-M. Rother, K. D. Bartl, A. Cristia, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, D. Garcia-Romero, and J. Villalba, “The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats,” in Interspeech, 2018, pp. 122–126.

[33] D. Snyder, G. Chen, and D. Povey, “Musan: A music, speech, and noise corpus,” arXiv preprint arXiv:1510.08484, 2015.