A COMPARISON STUDY ON INFANT-PARENT VOICE DIARIZATION

Junzhe Zhu, Mark Hasegawa-Johnson and Nancy L. McElwain

University of Illinois at Urbana-Champaign

ABSTRACT

We design a framework for studying prelinguistic child voice from 3 to 24 months based on state-of-the-art algorithms in diarization. Our system consists of a time-invariant feature extractor, a context-dependent embedding generator, and a classifier. We study the effect of swapping out different components of the system, as well as changing loss function, to find the best performance. We also present a multiple-instance learning technique that allows us to pre-train our parameters on larger datasets with coarser segment boundary labels. We found that our best system achieved 43.8% DER on test dataset, compared to 55.4% DER achieved by LENA software. We also found that using convolutional feature extractor instead of logmel features significantly increases the performance of neural diarization.

Index Terms— Child Speech, Language Development, Speaker Diarization, Voice Activity Detection, Multiple Instance Learning, Transfer Learning

1. INTRODUCTION

Mental health disorders and behavioral problems first emerge in early childhood [1]. Early diagnosis and intervention may help to ameliorate or prevent some types of behavioral disorders, but findings for the effectiveness of interventions are often mixed [2]. Intensive assessments of parent-infant interactions in naturalistic home environments, as well as normative data, are needed. Yet, such assessments conducted manually pose logistical challenges, including time and labor required by researchers, as well as privacy concerns of participating families. Automatic or semi-automatic diarization has the potential to address these limitations and provide insight into parent-infant vocal interactions – including the relative timing, duration, volume, and tone of voice – and would permit the establishment of normative data while minimizing privacy concerns. Greater volume of normative data, in turn, would facilitate the creation of effective evidence- based interventions in support of child mental health [3].

In contrast to broadcast news recordings [4], automatic diarization of naturalistic parent-child recordings is difficult because: (1) most speakers are usually recorded at a distance, (2) most utterances are informal and brief, (3) infants voices and types of utterances differ significantly from those of adults. Diarization of distant recordings [5] and informal speech has been well-studied, and end-to-end neural solutions [6] exist. Automatic diarization of children's speech, however, has only been extensively studied in the past three or four years, and remains challenging. Perhaps the most influential child speech diarization system is LENA [7, 8], a propriety system that segments audio recorded from LENA wearable devices, and that has been commonly used as a baseline [8, 9]. Additional recent work on child speech diarization has been inspired by the Second DiHARD Challenge [10], which includes childrens' speech (the Seedlings Corpus [11]) as one of its test corpora.

Because of the difficulty of the task, most systems submitting results to DiHARD use oracle voice activity detection (VAD) [12, 13, 14]; other solutions in the literature include automatic VAD [9, 15] or explicit models of one or more non-speech classes [8, 16, 17]. When oracle VAD is not provided, it is not always clear how a system should respond. Parent-infant interaction involves overlap of speech events, interspersed by long periods of silence. Instead of assigning each frame to a single type, it is more natural to formulate the problem as a multi-label sequence-to-sequence problem [6, 15, 17]. Furthermore, the permutation-invariant [18] labelling rules typical of other diarization tasks are less appropriate for infant databases, in which there is typically one infant, one female adult, one male adult, and sometimes one other child of a different age. Infant diarization accuracy may improve by pre-training and/or clustering models of 3 or 4 classes, e.g., key child, other child, female adult, male adult [8, 9, 9, 15, 16, 17].

This paper demonstrates end-to-end neural diarization of infant vocalizations. We also present a way to pre-train the neural network on large datasets with inaccurate boundary labels. In addition, we study the effect of different loss functions and input segment lengths of the network. Sec. 2 is the system description, Sec. 3 describes our approach to pre-training, Sec. 4 describes experimental methods and results, Sec. 5 concludes.

Thanks to Jiahao Xu from University of Sydney for help with server. This work was supported by funding from the National Institute on Drug Abuse (R34DA050256-01), the National Institute of Mental Health (R21MH112578-01) and the National Institute of Food and Agriculture, U.S. Department of Agriculture (ILLU-793-339).

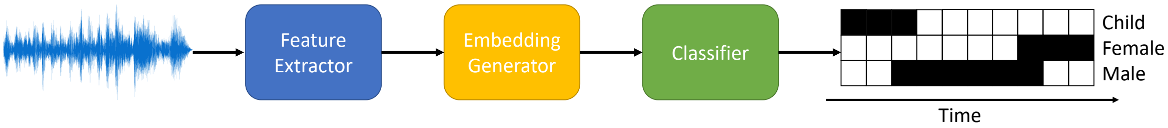


Fig. 1. An overview of system components. Feature extractor, embedding generator, and classifier are successively applied to input waveform, yielding a prediction for each speaker class at each frame.

2. SYSTEM DESCRIPTION

Suppose we have a dataset $\mathcal{D} = \{(\mathbf{x}, \mathcal{Y})\}$ where $\mathbf{x} \in \mathbb{R}^T$ is a waveform of T samples, and $\mathcal{Y} \in \mathbb{R}^{C \times L}$ is the label. C is the number of speaker classes that we define, and L is the number of frames, where $\mathbf{y}_{c,l}$ is the binary label indicating whether any speaker from class c is speaking at any sample in the range of $\left[\frac{l}{L}T, \frac{l+1}{L}T\right)$. We wish to design a system F_θ parameterized by θ that approximates the true mapping

$$\mathbf{x} \xrightarrow{F_{\text{true}}} \mathcal{Y} \tag{1}$$

from the audio to binary speaker class labels for each frame.

We decompose our system as follows:

$$F_\theta(\mathbf{x}) = (\text{Sigmoid} \circ F_{\text{cls}} \circ F_{\text{embed}} \circ F_{\text{feat}})(\mathbf{x}) \tag{2}$$

where

$$F_{\text{feat}} : \mathbb{R}^T \to \mathbb{R}^{H \times L} \tag{3}$$

is a time-invariant mapping from signal samples to feature space,

$$F_{\text{embed}} : \mathbb{R}^{H \times L} \to \mathbb{R}^{E \times L} \tag{4}$$

maps features to speaker embeddings with a one-to-one correspondence, and is conditioned on the rest of the frames. Finally,

$$F_{\text{cls}} : \mathbb{R}^{E \times L} \to \mathbb{R}^{C \times L} \equiv \mathbb{R}^E \to \mathbb{R}^C \tag{5}$$

maps each frame in the embedding sequence to a set of logits, and does not depend on other frames.

Features F_{feat}: Two types of features are tested.

The first tested feature vector is based on a 23-dimensional log-Mel-filterbanks with a window size of 25ms and hop size of 16ms. We splice features of 15 consecutive frames, and subsequently subsample the spliced feature matrix by a factor of 16 along the time dimension. Therefore, we compute a 345-dimensional feature vector every 256ms.

The second tested feature vector consists of a learned filterbank applied to waveform samples. This method is based on the encoder in [19], and consists of 12 blocks of Conv1D with zero-padding, followed by LeakyReLU activation, then Decimation Pooling which halves the time dimension. At a waveform sampling rate of 16000Hz, this feature extractor produces a 288-dimensional feature vector every 256 ms.

Embedding F_{embed}: Two types of neural embeddings were tested.

The first tested neural embedding is a Bi-Directional LSTM (BLSTM). Similar to [6], we use 5 layers with 256 hidden units for both forward and backward LSTM.

The second tested neural embedding is a multi-headed self-attention model. As in [6], we linearly transform each frame's feature vector, then apply two encoder layers. As in [20], each encoder layer includes layer normalization, then multi-headed self-attention, then another layer norm, then a two-layer position-wise fully-connected network. After both encoder layers, another layer normalization is applied. For self attention, we set both input and output dimensions to 256, and for the position-wise fully-connected network, we set the hidden layer size to 1024.

Classifier F_{cls}: To map the speaker embedding of each frame to a binary label for each speaker class, we use either a linear predictor, or a two layer fully connected network with ReLU activation (denoted as MLP), with the first hidden layer the same size as speaker embedding.

Focal Loss for Imbalanced Binary Labels: We use the Adam optimizer to train our networks. The majority of frames are silence: without considering voice overlap, for a label tensor of size $\mathbb{R}^{C \times T}$, only $\frac{T_{\text{on}}}{C \times T}$ of target entries will be 1, where T_{on} is the total number of frames where someone is speaking. Therefore, we treat this as a class-imbalanced classification problem, and use focal loss [21] to balance our training. As with the best configuration in [21], we use $\alpha = 0.25, \gamma = 2$ as our hyper-parameters. When running on test data, we set the prediction threshold to 0.5.

3. PRE-TRAINING WITH MULTIPLE INSTANCE LEARNING

Because the time resolution of our diarization system is 256ms per frame, we require high-resolution labels to train the system. However, infant recordings mostly consist of non-speech vocalizations, so the average speaker turns are extremely short [16] compared with normal diarization datasets. In addition, overlapping vocalizations are frequent in infant recording. Therefore, it is relatively costly to acquire accurate labels; our core training and test datasets are precisely labeled, but relatively small.

Due to the limited size of our own training set, we use the Braunwald [23] and Providence [24] Corpora from the CHILDES project as a transfer learning dataset. During manual inspection, we noticed that in both datasets, although the

Config	F_{feat}	F_{embed}	F_{cls}	pre train	param loaded	freeze F_{feat} 10epoch	freeze F_{embed} 10epoch	input len (s)	loss	DER
1	Conv	BLSTM	linear	None	-			20	focal	0.486
2	Conv	MHA								0.490
3	LogMel	BLSTM								0.535
4	LogMel	MHA								0.665
5	Conv	BLSTM	MLP 2layer	MIL1	$F_{\text{feat}}, F_{\text{embed}}$	Yes	Yes	20	focal	**0.438**
6	Conv	MHA			$F_{\text{feat}}, F_{\text{embed}}$	Yes				0.509
7	LogMel	BLSTM			F_{embed}	-				0.509
8	LogMel	MHA			F_{embed}	-				0.638
9	Conv	BLSTM	MLP 2layer	MIL1	$F_{\text{feat}}, F_{\text{embed}}$	No	No	20	focal	0.461
10	Conv	MHA			$F_{\text{feat}}, F_{\text{embed}}$	No				0.446
11	LogMel	BLSTM			F_{embed}	-				0.533
12	LogMel	MHA			F_{embed}	-				0.628
13	Conv	BLSTM	linear	MIL2	$F_{\text{feat}}, F_{\text{embed}}, F_{\text{cls}}$	Yes	Yes	20	focal	0.465
14	Conv	MHA			$F_{\text{feat}}, F_{\text{embed}}, F_{\text{cls}}$	Yes				0.449
15	LogMel	BLSTM			$F_{\text{embed}}, F_{\text{cls}}$	-				0.521
16	LogMel	MHA			$F_{\text{embed}}, F_{\text{cls}}$	-				0.656
17	Conv	BLSTM	linear	MIL2	$F_{\text{feat}}, F_{\text{embed}}, F_{\text{cls}}$	No	No	20	focal	0.514
18	Conv	MHA			$F_{\text{feat}}, F_{\text{embed}}, F_{\text{cls}}$	No				0.444
19	LogMel	BLSTM			$F_{\text{embed}}, F_{\text{cls}}$	-				0.523
20	LogMel	MHA			$F_{\text{embed}}, F_{\text{cls}}$	-				0.631
LENA	-			-				-	-	0.554
Lavechin et al. [17]	Sinc-Net[22]	BLSTM	MLP 3layer	-				2	MSE	0.586
Ablation1	Conv	BLSTM	linear	None	-			20	BCE	0.509
Ablation2	Conv	BLSTM	linear	None	-			2	focal	0.470
Ablation3	Conv	BLSTM	MLP 2layer	MIL1	$F_{\text{feat}}, F_{\text{embed}}$	Yes	Yes	2	focal	0.451

Table 1. Performance of different system configurations

labelled speaker turns are usually correct in terms of speaker class, turn boundaries are relatively imprecise: most speaker turns contain silence at both start and end.

Because we do not have the true binary speaker class label $\mathcal{Y}$ for each frame in our transfer learning dataset, we cannot directly train the same system described in 2. Therefore, we re-formulate the problem as a multiple-instance learning problem, described below:

Given $(\mathbf{x}, s) \in \mathcal{D}_{\text{pretrain}}$ as our transfer learning datum, where each pair of $\mathbf{x}$ and $0 < s < C$ are respectively the audio samples and the speaker class label for a segment, we wish to find a mapping G_θ that shares the same parameter θ with F_θ. We wish to learn the parameters θ to maximize the accuracy of Eq. 1, but the true value of $\mathcal{Y}$ is unknown (time alignment of the utterance within the segment is unknown), therefore we design a classifier G_θ to maximize the accuracy of

$$\mathbf{x} \xrightarrow{G_\theta} \mathcal{Y}_{MIL} \in \{\mathbf{1}[s], \mathbf{0}\} \tag{6}$$

where $\mathcal{Y}_{MIL}$, the multiple-instance learning target, is either $\mathbf{1}[s]$ (a one-hot vector for s) or $\mathbf{0}$ (the zero vector). Eq. 6 can be computed using a G_θ that computes the maximum over a segment, and compares the maximum to $\mathbf{1}[s]$, thus

$$G_\theta(\mathbf{x}) = (\text{SoftMax} \circ \text{MaxPool} \circ F_{\text{cls}} \circ F_{\text{embed}} \circ F_{\text{feat}})(\mathbf{x}) \tag{7}$$

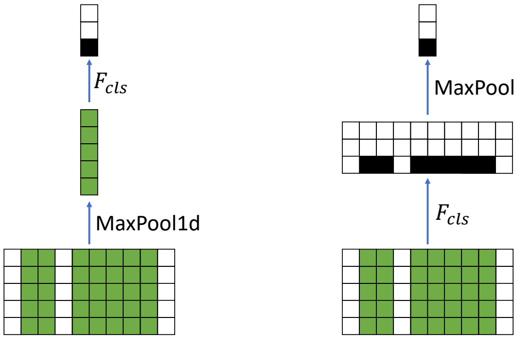
or

$$G_\theta(\mathbf{x}) = (\text{Softmax} \circ F_{\text{cls}} \circ \text{MaxPool} \circ F_{\text{embed}} \circ F_{\text{feat}})(\mathbf{x}) \tag{8}$$

where MaxPool denotes global max-pooling over time. In other words, we add global max-pooling to our original network to make it a classifier over the whole segment.

4. EXPERIMENTAL METHODS AND RESULTS

Primary training, validation and test data were drawn from two studies of socioemotional development. Families with typically developing infants between 3 and 24 months of age were recruited from the community. Infants wore the LENA recorder for a total of 16 hours in the home. Reference labels were manually coded for 107 10-minute segments that,



(a) MIL1 (b) MIL2

Fig. 2. Two configurations for multiple-instance learning. Colored blocks are speaker embedding. Black and white boxes are speaker class.

according to LENA segmentation, had the highest frequency of voice activity (23 at 3 months, 20 at 6 months, 22 at 9 months, 22 at 12 months, and 20 at 13-24 months). We split those into 87 for train, 10 for validation, and 10 for test. For each 10-minute segment, we then manually labelled four tiers, with cross-labeler validation at a precision of 0.2 seconds: CHN (key child), CXN (other child), FAN (female adult), and MAN (male adult). Neural nets were trained using audio waveforms normalized to $[-1, 1]$, and divided into 20-second segments with 0.256 second frames. We consider CHN and CXN as the same speaker class, and FAN and MAN as separate speaker classes.

We use Diarization Error Rate(DER) as our primary metric. It can be computed as

$$\frac{\sum_{s=1}^{S} \text{dur}(s) \cdot (\max(N_{ref}(s), N_{hyp}(s)) - N_{correct}(s))}{\sum_{s=1}^{S} \text{dur}(s) \cdot N_{ref}} \tag{9}$$

Note that DER in the case of infant speech would be overall much higher than normal case, since the denominator of DER is smaller when less voice activity is present.

Networks were pre-trained using the transfer learning datasets, Braunwald and Providence, using two different MIL frameworks (MIL1=Eq. 7 and MIL2=Eq. 8). Variable-length segments were used, based on the start and end times in the labels. We only keep segments with durations between 1.28s and 10.24s. The best configurations using MIL1 and MIL2 achieved respectively 82.9% and 83.3% classification accuracy on validation set.

For both MIL1 and MIL2, we pre-train with all combinations of F_{feat} and F_{embed}, and use a linear classifier with 3 classes(Child, Female, Male). We train with Adam optimizer for 5 epochs, with learning rate of 0.0005 and decay of 0.5 per epoch.

Configurations and results are listed in Table 1, using the notation introduced in Sec. 2. Configurations 1-4 and 9-12 usee learning rate 0.001 with decay of 0.98 per epoch. Configurations 5-8 used learning rate 0.0005 with decay 0.94 per epoch. Configurations 13-20 used learning rate 0.0005 with decay 0.98.

We ran two baselines developed by others on our test set: LENA [7, 8] and the system of Lavechin et al. [17]. We ran three additional ablation studies on our baseline, each studying the effect of loss function, input segment size, and whether to use an additional speaker class for non-key child, each using the same learning rate as the configuration it ablates.

Results[1] are shown in 1. Ablation studies 1, 2 and 3, compared to configurations 1, 1, and 5, respectively, show that focal loss improves performance and that shorter chunk size has uncertain effect on performance. Our best system is based on configuration 5, which achieves 43.8% DER, compared to 55.4% and 58.6% DER achieved by baselines from LENA software and Lavechin et al.[17]. We also note that convolutional feature extractors work better than logmel features in most cases, which contradicts prior practice in diarization systems. This could be due to the various forms of vocalizations in the infant speaker class, which include extremely high-pitched vocalizations that may be ill-adapted to logmel features.

Our best system (configuration 5) was also trained on a task with 4 speaker-class targets, in which the key child and other children are counted as separate classes. DER and frame error rate (percentage of frames with prediction error in at least one class) are reported in Table 2. DER of LENA and our system were not affected much, but DER of [17] increased.

System	DER	Frame Error Rate
Ours, config 5	0.497	0.338
LENA	0.581	0.353
Lavechin et al., 2020[17]	0.762	0.454

Table 2. DER and Frame Error Rate of each system on 4-speaker case

5. CONCLUSIONS

This paper offers two key contributions. First, we decomposed the child-speech diarization systems into three separately analyzed components: F_{embed}, F_{feat}, F_{cls}, and provided results for two configurations of each component. Second, we developed a pre-training procedure to enable transfer learning from datasets with coarse speaker segment labels. We found that convolutional features, combined with focal loss training and transfer learning, together achieve the most accurate system.

[1] code can be found at https://github.com/JunzheJosephZhu/Child_Speech_Diarization

References

[1] H.L. Egger and A. Angold, "Common emotional and behavioral disorders in preschool children: presentation, nosology, and epidemiology," *Journal of child psychology and psychiatry*, vol. 47, no. 3-4, pp. 313–337, 2006.

[2] B. Wright and E. Edgington, "Evidence-based parenting interventions to promote secure attachment: Findings from a systematic review and meta-analysis," *Global Pediatric Health*, vol. 3, pp. 1–14, 2014.

[3] K. Hoagwood, B.J. Burns, L. Kiser, H. Ringeisen, and S.K. Scheonwald, "Evidence-based practice in child and adolescent mental health services," *Psychiatric Services*, vol. 52, no. 9, pp. 1179–1189, 2001.

[4] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain, "Speaker diarization from speech transcripts," in *Interspeech*, 2004.

[5] J.M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences," in *Interspeech*, 2006, pp. 1337:1–4.

[6] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *ASRU*, 2019, pp. 296–303.

[7] M. Canault, M-T. Normand, S. Foudil, N. Loundon, and H. Thai-Van, "Reliability of the language environment analysis system (LENA) in european french," *Behavior research methods*, vol. 48, 07 2015.

[8] 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.

[9] A. Le Franc, E. Riebling, J. Karadayi, Y. Wang, C. Scaff, F. Metze, and A. Cristia, "The ACLEW DiViMe: An easy-to-use diarization tool," in *INTERSPEECH*, 2018.

[10] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The second DIHARD diarization challenge: Dataset, task, and baselines," in *INTERSPEECH*, 2019.

[11] E. Bergelson and R.N. Aslin, "Nature and origins of the lexicon in 6-mo-olds," *PNAS*, vol. 114, no. 49, pp. 12916–12921, 2017.

[12] L. Sun, J. Du, T. Gao, Y. Lu, Y. Tsao, C. Lee, and N. Ryant, "A novel LSTM-based speech preprocessor for speaker diarization in realistic mismatch conditions," in *ICASSP*, 2018, pp. 5234–5238.

[13] Z. Zajíc, M. Kunesová, J. Zelinka, and M. Hrúz, "ZCU-NTIS speaker diarization system for the DIHARD 2018 challenge," in *INTERSPEECH*, 2018.

[14] J. Xie, L. P. García-Perera, D. Povey, and S. Khudanpur, "Multi-PLDA diarization on children's speech," in *INTERSPEECH*, 2019.

[15] P. García, Jesús Villalba, and H. Bredin et al., "Speaker detection in the wild: Lessons learned from JSALT 2019," 2019.

[16] M. Najafian and J. H. L. Hansen, "Speaker independent diarization for child language environment analysis using deep neural networks," in *SLT*, 2016, pp. 114–120.

[17] M. Lavechin, R. Bousbib, H. Bredin, E. Dupoux, and A. Cristia, "An open-source voice type classifier for child-centered daylong recordings," 2020.

[18] D. Yu, M. Kolbæk, Z-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP*, 2017, pp. 241–245.

[19] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *ISMIR*, 2018.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *ArXiv*, vol. abs/1706.03762, 2017.

[21] T.-Y. Lin, P. Goyal, R.B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2999–3007.

[22] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," 2018.

[23] S.R. Braunwald, "Mother-child communication: The function of maternal-language input," *WORD*, vol. 27, no. 1-3, pp. 28–50, 1971.

[24] J. Yung Song, K. Demuth, K. Evans, and S. Shattuck-Hufnagel, "Durational cues to fricative codas in 2-year-olds' American English: Voicing and morphemic factors," *JASA*, vol. 133, no. 5, pp. 2931–2946, 2013.