A COMPARISON STUDY ON INFANT-PARENT VOICE DIARIZATION

Junzhe Zhu Mark Hasegawa-Johnson Nancy L. McElwain

University of Illinois at Urbana-Champaign

MOTIVATION

Task: In-home speech diarization

- In-home environment, with 3-24 month children
- Detects who speaks when & for how long
- End goal: generate reliable speech event labels for child linguistic studies
- 4 speakers for each family: Infant, Older Child, Mother, Father

INTRODUCTION

Challenges: Noisy Environment, Few Data

Analysis of in home recordings is a challenging task. Usually home environments are very noisy, and the recording devices are worn by the subjects, which add more noises to the recording.

The privacy agreement of these recordings also don't usually allow researchers to publicize the data, which makes it very hard to find training data for the task.

In addition, newborn children make many sounds like crying or babbling, which adds to the difficulty of the labelling process.

Figure 1: Example of labeller disagreement

We recruited a team of >5 labellers. We selected 19 files of each 10-mintues, and had 2 labellers independently annotate each file. Results show that on average, two labellers disagree on 19% of frames.

OUR APPROACH

Dataset: We collected our own dataset from recruited families. Out of all our recordings, we selected 107 segments of 10-minutes each with the highest voice activity.

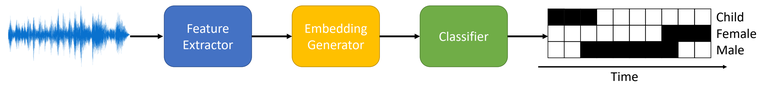

Figure 2: Components of our system include a feature extractor, backbone processor, and classifier

We decompose our system as follows:

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

Encoder: a time-invariant mapping from signal samples to feature space. $F_{\text{feat}} : \mathbb{R}^T \rightarrow \mathbb{R}^{H \times L}$

We use either a 12-layer convolutional neural network or Log-Mel Spectrogram features.

Backbone Network: maps features to speaker embeddings with a one-to-one correspondence, and is conditioned on all frames. $F_{\text{embed}} : \mathbb{R}^{H \times L} \rightarrow \mathbb{R}^{E \times L}$

We use either a bidirectional LSTM network or a transformer.

Classifier: maps each frame in the embedding sequence to a set of logits, and does not depend on other frames. $F_{\text{cls}} : \mathbb{R}^{E \times L} \rightarrow \mathbb{R}^{C \times L} \equiv \mathbb{R}^{E} \rightarrow \mathbb{R}^{C}$

We use either a linear classifier or a 2-layer neural network.

Focal Loss: We treat the problem as a multi-target binary classification problem, since speech from different individuals sometimes overlap. Since we have 4 output tiers, and only around 50% of active speech time, the output target is unbalanced between 0s and 1s. To address the sparsity, we use focal loss.

PRETRAINING

Pretraining: Since it is hard to retrieve data, we developed a pre-training method to train our system on data with noisy boundaries.

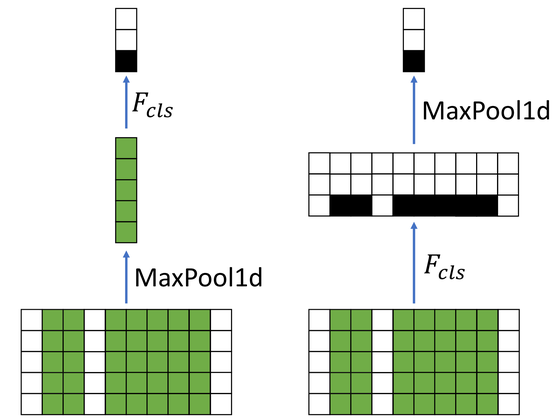

Figure 3: two ways of pre-training the system with multiple instance learning

In order to ignore the effects of noise boundary, we pre-train our system to classify the speaker class of segments cropped based on the noisy boundary. We add a max pooling layer above either the backbone or classifier layer. By using max pooling, the network ignores the parts of the noisy cut segment where no speech is present.

RESULTS

System	DER	Frame Error Rate
Ours(Convolutional Encoder + Bi-LSTM + 2-layer NN)	0.497	0.338
LENA	0.581	0.353
Lavechin et al., 2020[?]	0.762	0.454

Table 1: DER and Frame Error Rate of each system