A COMPARISON STUDY ON INFANT-PARENT VOICE DIARIZATION

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

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

Challenges

- Extremely noisy environment
- Lack of training data(due to privacy reasons)
- Difficulty in labelling

Prior Art

- Mostly from DiHARD challenge, based on oracle voice activity detection(VAD), i.e. the coverage of speech events is known a priori
- LENA system
- Speaker diarization has achieved great advancement recently due to the development of neural networks. End-to-end(E2E) methods have been developed.

Task & training data

- In home recordings, >12 hours long for each family
- Out of 12 hours, mostly silence
- Picked 107 segments of 10-minute length (23 at 3 months, 20 at 6 months, 22 at 9 months, 22 at 12 months, and 20 at 13-24 months) with high voice activity
- At most 4 speakers present in each recording: an infant, an older child, a mother, a father

Data Examination

- A team of around >5 labelers

2.12%
8.57%
9.47%
7.65%
12.84%
19.14%
17.30%
18.92%
16.02%
6.32%

- Average disagreement between labelers on the same file(by mismatching % of frames): 19.77%

Our work

- A survey of neural network architectures on this particular task
- Pre-training techniques

Neural Network Architectures: Feature

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

- Features:
- Downsample in time dimension(by samples/frame), upsample feature dimension
- 2 methods: Convolutional neural network Encoder, LogMel spectra

Neural Network Architectures: Embedding

- $$F_{\text{embed}} : \mathbb{R}^{H \times L} \to \mathbb{R}^{E \times L}$$
- Backbone: map from feature to speak embedding
- 2 architectures: Bi-LSTM and Transformer
- One-to-one correspondence in time dimension

Neural Network Architectures: Output

- $F_{\text{cls}} : \mathbb{R}^{E \times L} \rightarrow \mathbb{R}^{C \times L} \equiv \mathbb{R}^{E} \rightarrow \mathbb{R}^{C}$
- Classifier: Speaker embedding to classes
- Linear/two layer neural network
- Operated independently on each frame

Output & Loss

- Treat diarization as a multi-label problem, one probability for each class at each time index
- Use sigmoid function for output activation
- Use focal loss to handle class imbalance

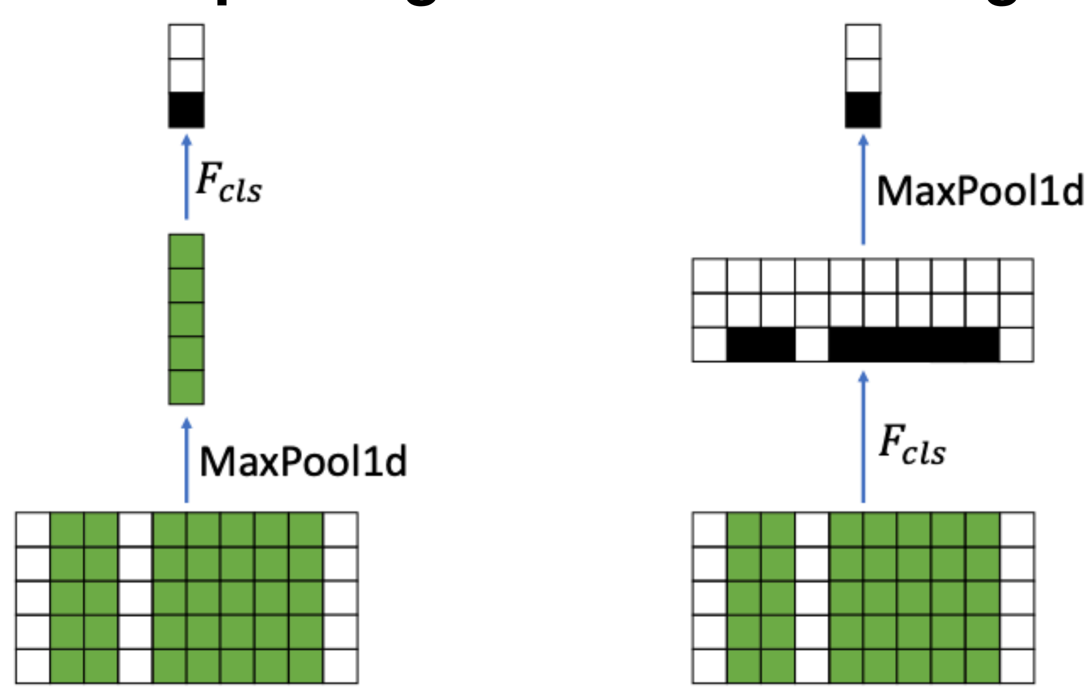
Pre-training

- Few training datasets with accurate labelling
- Make use of public datasets with noisy label
- For example: for a 5s recording, actual speech happens in [2.3, 3.5], but label says [2.3, 4,5] is speech
- Can’t train directly with end-to-end diarization due to false positive labels

Using multiple instance learning to learn from noisy label

- Re-formulate into a speaker classification problem - classify a speech event w/ noisy boundary
- The bottom matrix in the chart denotes speaker embedding in a labelled speech segment, where the horizontal axis is time.
- Each column is a speaker embedding. Green columns indicate actual speech
- White strips denote false positive labels (silence labelled as speech). Max pooling ignores their contribution to the final result.

2 ways of max pooling: above embedding/out layer



(a) MIL1 (b) MIL2

Final Results

- Best results: Convolutional neural network encoder + Bi-LSTM backbone + 2-layer neural network classifier

- Diarization Error Rate(DER) of 0.438 (Baseline is LENA, a proprietary system, with DER of 0.581)

- High DER due to small denominator (low total duration of speech events for in-home environments)

Summary

- Task: Diarize in-home recordings
- Different neural network encoder/backbone/classifier
- Use pre-training to solve few-data problem