OPENFEAT: Improving Speaker Identification by Open-set Few-shot Embedding Adaptation with Transformer

Kishan K C, Zhenning Tan, Long Chen, Minho Jin, Eunjung Han, Andreas Stolcke, Chul Lee
Amazon Alexa, USA
IEEE ICASSP, 2022

Speaker identification

- Speaker identification is key to enable personalization for voice assistants, such as Alexa, and Google Home
- Speaker identification in households is challenging because of their
  - Similar voice characteristics
  - Acoustic conditions
- For real-word data, mean cosine similarity within a household is about 10% greater than the similarity with utterances outside the household.

Current approach

- Three stage solution:
  1. Train universal speaker encoder on a large number of speakers
  2. Compute distance between test utterances and each of the speaker embeddings
  3. Identify speaker that is closest to the test utterance based on similarity

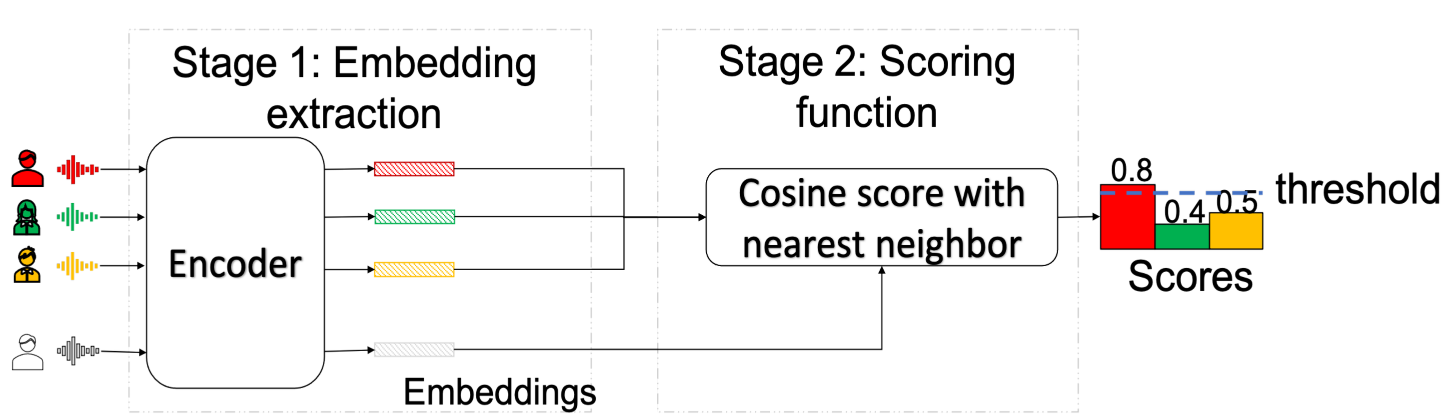


Fig 1: Block diagram of current approach to identify speaker for a test utterance

Limitations

- Embeddings learned from the universal speaker encoder are not necessarily optimal to discriminate specific set of speakers in a household.
  - Household speakers are more difficult to distinguish compared to arbitrary speakers because they typically share similar accent, acoustic conditions.
  - Current approach doesn't consider the similarities between household speakers when making individual comparison at scoring stage.
- Training with classification loss or contrastive loss divides embedding space using class boundaries. Such decision boundaries are not optimal for unseen guest utterances.

Proposed method

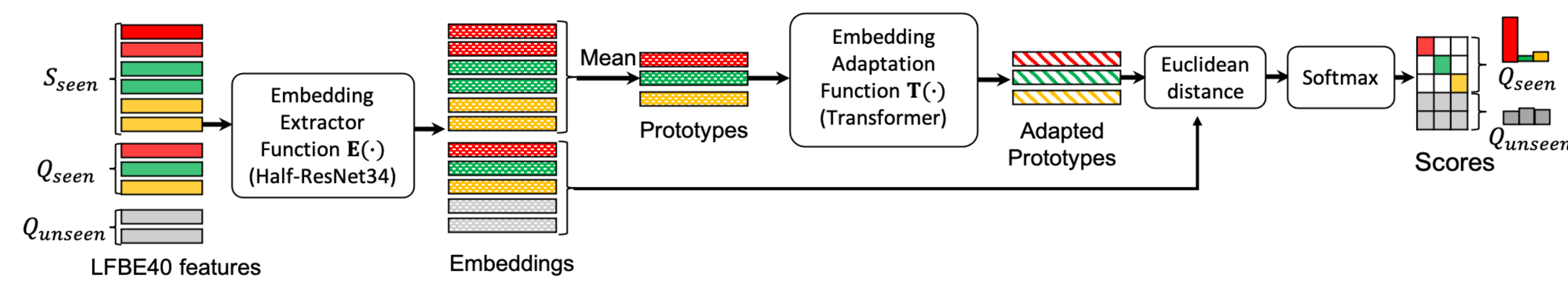


Fig 2: Architecture of OPENFEAT

Few-shot Learning (FSL)

- Learn from few examples (support set) to make predictions on novel cases (query set)
- Episodes are generated by selecting N speakers with K labeled utterances per speaker, represented as an N-way K-shot problem.
- Each episode has support set $S_{seen} = \{x_i^S, y_i^S\}_{i=1}^{NK}$ and a query set $Q_{seen} = \{x_i^Q, y_i^Q\}_{i=1}^{NM}$
- Prototypes $P = \{p_1, p_2, \ldots, p_n\}$ are computed using support set and a distance-based scoring function is used to predict speakers for test utterances.

Few-shot embedding adaptation with Transformer

- To adapt prototypes to be more distinguishable in a household-specific space, a set-to-set function such as transformer is trained.

$$P' = \mathrm{Transformer}(P)$$

- Adapted prototype P' is used to compute FSL Classification loss.

$$L_{query} = \sum_{(x,y)\in Q_{seen}} \mathcal{L}_{CE}(y, f(x, P'))$$

- To ensure instance embeddings after adaptation are closer to their class neighbors and far away from other classes, contrastive loss is computed using both support and query set to compute prototypes C

$$\mathcal{L}_{contrastive} = \sum_{(x,y)\in Q_{seen} \cup S_{seen}} \mathcal{L}_{CE}(y, f(x, C))$$

FSL with Open-set

- For each episode, R speakers with T utterances per speaker are randomly sampled and denoted as unseen query set $Q_{unseen} = \{x_i^U, y_i^U\}_{i=1}^{RT}$
- The open-set loss calculated based on the posterior entropy

$$\mathcal{L}_{open-set} = -\sum_{x \in Q_{unseen}} \mathcal{L}_{entropy}(f(x, P'))$$

- Finally, the total loss for openFEAT is

$$L_{\mathrm{openFEAT}} = L_{query} + \alpha\mathcal{L}_{contrastive} + \beta L_{open-set}$$

Experimental setup

- VoxCeleb2 to train the encoder and embedding adaptation module
- Voxceleb1 to evaluate models
  - Select hard-to-discriminate speakers based on 85th percentile among cosine similarity between speaker profiles
  - Average 4 utterances to generate enrollment utterances
  - Randomly sample 50 * household size as guest utterances

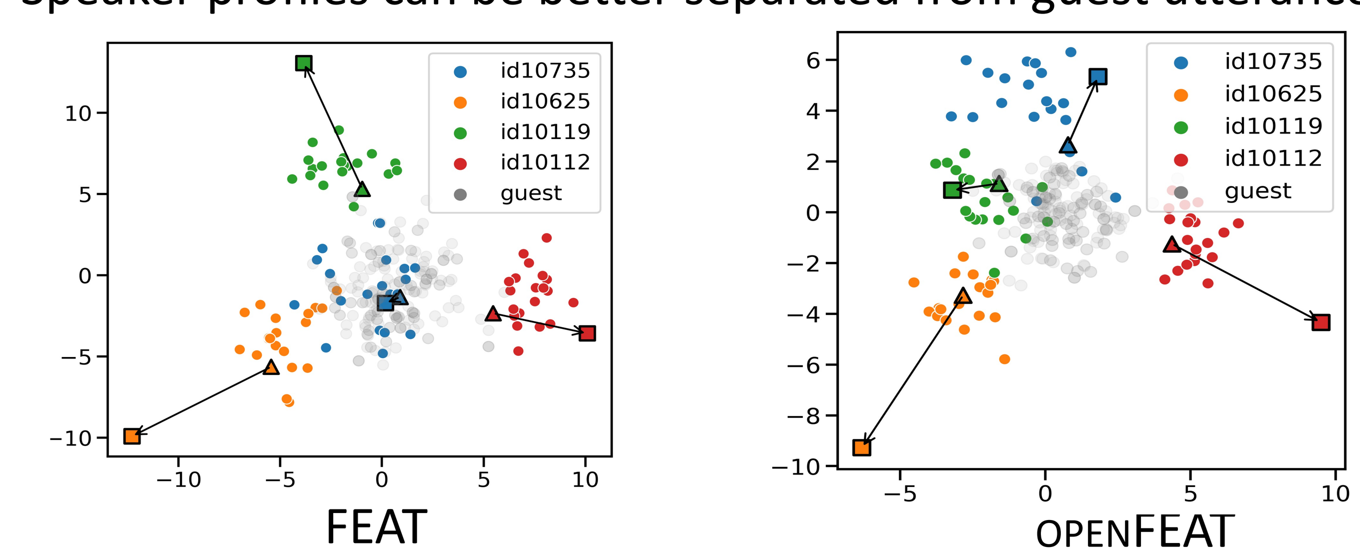
Performance evaluation

- Define identification equal error rate (IEER) as a point where FAR equals FNIR.
- Baseline IEER increases with household size
- IEER reduced by 22.8% to 30.75 relative

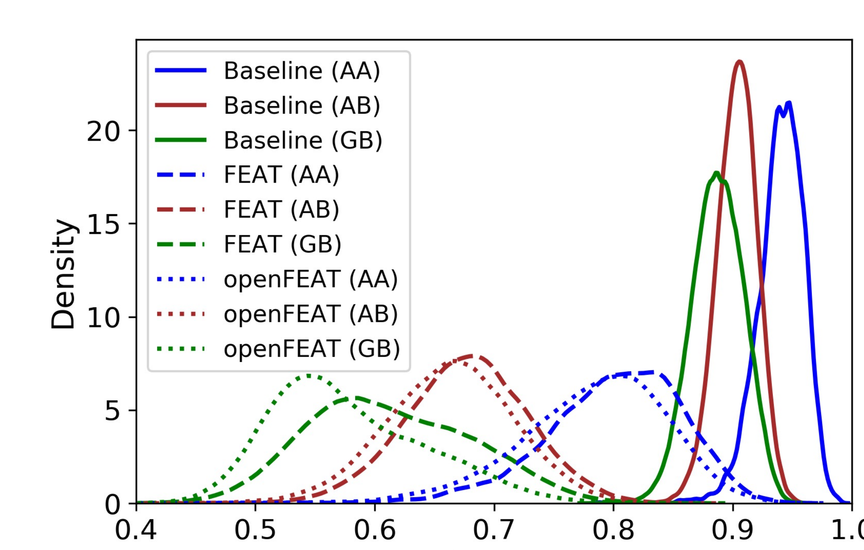
n	Baseline	FEAT	Open-set	openFEAT
2	6.48±0.29	4.91±0.31 (24.3%)	5.16±0.15 (20.4%)	**4.49±0.20 (30.7%)**
3	8.65±0.14	6.75±0.21 (22.0%)	7.06±0.12 (18.4%)	**6.06±0.18 (30.0%)**
4	10.56±0.26	8.56±0.15 (18.9%)	8.73±0.12 (17.4%)	**7.67±0.21 (27.4%)**
5	11.98±0.18	10.01±0.26 (16.5%)	10.04±0.23 (16.2%)	**9.02±0.24 (24.8%)**
6	13.46±0.12	11.37±0.18 (15.5%)	11.23±0.12 (16.5%)	**10.30±0.23 (23.5%)**
7	14.69±0.37	12.45±0.35 (15.3%)	12.35±0.38 (16.0%)	**11.35±0.37(22.8%)**

Embedding visualization

- With openFEAT, adapted speaker profiles are further apart from each other based on PCA projection.
- Speaker profiles can be better separated from guest utterances.



Score distribution



After adaptation of speaker centroid, the margin between the query utterance to its corresponding speaker vs other speakers in the household increases.

Conclusion

- openFEAT enables better separation of speaker profiles and also reduce speaker confusability with unseen speakers.
- openFEAT achieves relative IEER reduction of 23% to 31% for simulated households of hard-to-discriminate speakers.