Scalable Identity-Oriented Speech Retrieval

Chaotao Chen, Di Jiang, Jinhua Peng, Rongzhong Lian, Yawen Li,
Chen Zhang, *Member, IEEE,* Lei Chen, *Member, IEEE,* and Lixin Fan

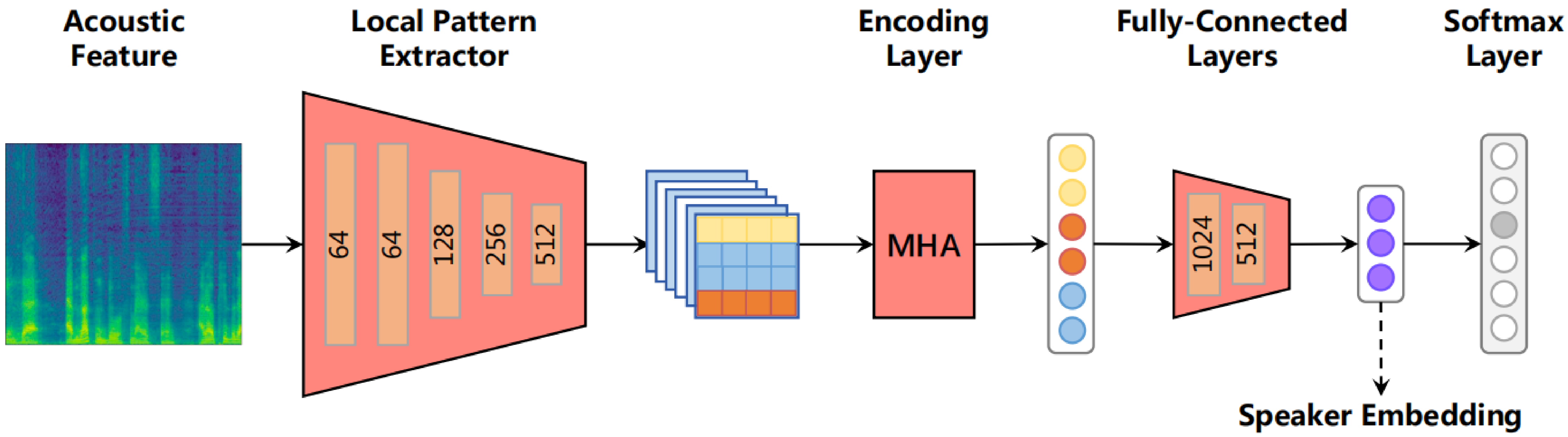
Ruihai Hou

2022/04/01

Identity-Oriented Speech Retrieval

- Because most of speech data is collected without identity annotation, how to efficiently retrieve the speech snippets uttered by a given person has become the main challenge in the application of speech data to security surveillance and financial risk management. This task is named as s **Identity-Oriented Speech Retrieval** (IO-SR).

Speaker Embedding Model(SEM)



Deep Hashing

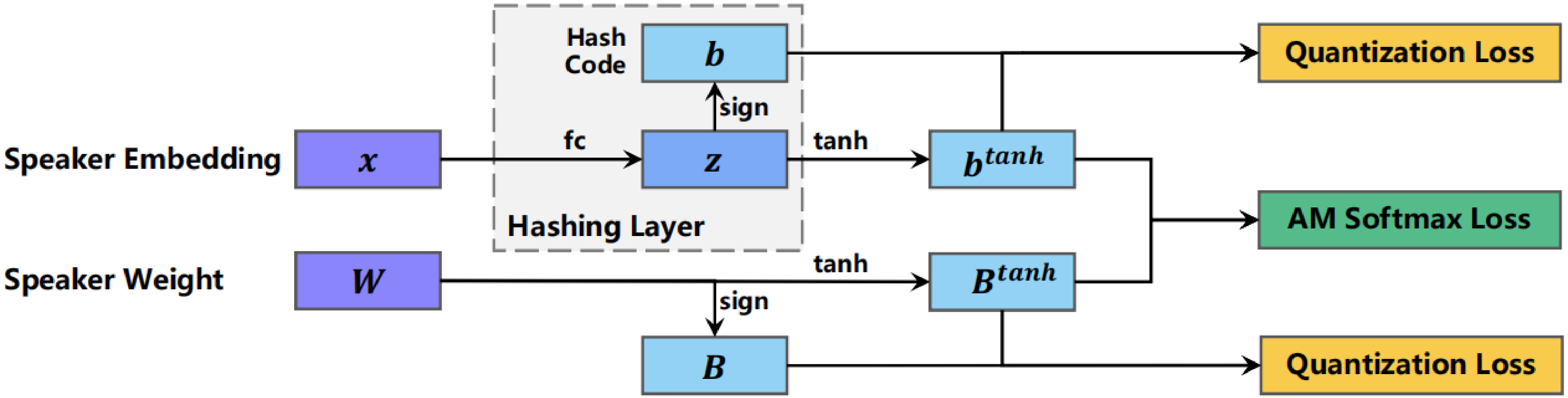


Fig. 2. Diagram of our proposed Deep Hashing.

$$\mathcal{L}_{speaker} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{i,y_i})-m)}}{e^{s(\cos(\theta_{i,y_i})-m)}+\sum_{r\neq y_i}e^{s(\cos(\theta_{i,r}))}}$$

Combining loss function

$$\mathcal{L}_{hash} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}(||\boldsymbol{b}_i - \boldsymbol{b}_i^{tanh}||_2^2 + ||\boldsymbol{B} - \boldsymbol{B}^{tanh}||_2^2)$$

$$\mathcal{L}_{speaker} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{i,y_i})-m)}}{e^{s(\cos(\theta_{i,y_i})-m)} + \sum_{r\neq y_i} e^{s(\cos(\theta_{i,r}))}}$$

$$\mathcal{L} = \mathcal{L}_{speaker} + \lambda\mathcal{L}_{hash}$$

Inverted index

- Indexing process
  - Perform k-means algorithm on the training set of binary hash codes to find the centroids of each inverted list
  - A speech snippet is assigned to its closest cluster in hamming distance and its speech ID is appended to the corresponding inverted list.

Inverted index

- Search process
  - The query speech is first compared with the clusters with the binary hash codes
  - Then only the candidates in the most similar clusters are retrieved for linear scan

Experiments

- Datasets
  - Different public speech datasets for evaluation, including VoxCeleb1 , VoxCeleb2 , Aishell-1 , Aishell-2 and MAGIDATA1.
  - Only the speech snippets longer than 3 seconds are reserved and the speakers with less than 70 snippets are removed.
  - In total, 1,638,983 speech snippets from 8,923 speakers.

Experimental Results

- SEM achieves a high Acc@1 (98.0%) and a high MAP (97.7%)
- Deep Hash method achieves comparable Acc@1 and MAP to PQ and comparable query speed to LSH
- All methods with inverted indexing achieve significant speedup

TABLE 2
Evaluation results of Acc@1 (%), MAP (%), database memory consumption (in MB) and query time (in second).

Method	Bytes	Acc@1	MAP	Memory	Time
SEM	2048	98.0	97.7	2156	16.508 (1.00x)
LSH [10]	16	90.4	89.4	18	1.053 (15.7x)
PQ [28]	16	95.2	94.3	18	6.650 (2.48x)
Deep Hash	16	95.1	94.1	17	1.181 (14.0x)
LSH [10]	32	95.5	94.8	35	1.711 (9.65x)
PQ [28]	32	96.6	95.7	35	12.483 (1.32x)
Deep Hash	32	96.3	95.6	34	2.358 (7.00x)
Index	2048	97.9	97.5	2173	2.225 (7.42x)
Index + PQ	16	93.9	93.1	27	1.410 (11.7x)
Deep Index	16	93.4	92.8	26	0.042 (393.x)
Index + PQ	32	96.0	95.3	44	1.515 (10.9x)
Deep Index	32	95.7	95.1	43	0.052 (317.x)

Conclusion

- Propose a novel system for large-scale identity-oriented speech retrieval by seamlessly combining techniques from DNN-based speaker recognition, deep hashing and indexing methods

- Quantitative experiments on a one-million speech database demonstrate the effectiveness and scalability of our proposed system

Thank you!