DNN based Embeddings for Language Recognition

Alicia Lozano-Diez\(^{1,2}\), Oldrich Plchot\(^2\), Pavel Matejka\(^2\), Joaquin Gonzalez-Rodriguez\(^1\)

\(^1\)Audias-UAM, Universidad Autonoma de Madrid, Madrid, Spain
\(^2\)Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Czechia

**Abstract**

- **Bottleneck features**: frame-wise representation
  - From DNN trained for speech recognition (ASR)
  - Variable-length sequence of features \(\rightarrow\) fixed-length i-vector

- **Embeddings**: utterance-level representation (fixed-length)
  - From DNN trained for language recognition (LID, target task)
  - Similar approach previously applied to speaker recognition

**Proposed DNN-Embeddings**

- **DNN** trained to discriminate between languages
- Also provides fixed-length embeddings that summarize the whole utterance with useful information about the language
- Proposed architecture:
  - **Frame level**: input + BLSTM + fully connected
  - **Pooling** layer (mean and std over time)
  - **Utterance level**: fully connected (embeddings) + softmax output layer

- **Input**: 30-dimensional stacked bottleneck features (SBN)
- Trained to optimize multi-class cross-entropy loss function
- 3s sequences for training, no length constraints for embedding extraction

**Experiments and Results**

**Different size of DNN-embeddings** (concatenation \(\text{emb}_a + \text{emb}_b\)):
- Halving size of the embedding layers:
  - DNN\(_1\): 512 + 300 = 812
  - DNN\(_2\): 256 + 150 = 406
  - DNN\(_3\): 128 + 75 = 203
- Comparison of performance:
  - Embeddings vs. posterior probabilities (directly from the DNN)

**Study of PCA post-processing of embeddings**, motivated by:
- Better results with smaller embeddings (DNN\(_2\) and DNN\(_3\) vs. DNN\(_1\))
- First architecture (DNN\(_1\)) inspired by success on speaker recognition with larger number of classes (thousands of speaker vs. 20 languages)

**Score-level fusion of embedding and reference i-vector systems**:
- Reference i-vector (GLC)
- Best DNN-embeddings: DNN\(_2\) (concatenated \(\text{emb}_a + \text{emb}_b\) + PCA 25)
- Best DNN posterior probabilities: DNN\(_2\)

**LID Backend**

- Gaussian Linear Classifier (GLC):
  - On top of i-vectors or embeddings
  - Outputs the vector of 20 class-conditional log-likelihoods for each segment
  - Model of each language: Gaussian distribution
  - Mean: over i-vectors of given language
  - Covariance matrix: shared across all models
- **Calibration** and score level fusion:
  - Multi-class logistic regression trained on top of the development scores

**Conclusions**

- Proposed DNN-embedding system for LID
- Embedding: fixed-length utterance-level representation, provided by a DNN trained for the target task (LID)
- **Novel** approach for LID (in line with research in speaker ID)
- Results comparable with state-of-the-art i-vector system
  - Up to 7.3\% relative improvement with simple fusion
  - Better results with embeddings than posteriors from the DNN, possibility for more general DNNs usable across LID tasks

**Selected References**