Deep learning-based end-to-end spoken language identification system for domain-mismatched scenario

Woo Hyun Kang, Jahangir Alam, Abderrahim Fathan
Computer Research Institute of Montreal (CRIM)

Introduction

•Spoken language identification (LID)
- •The task of identifying the uttered language given a speech sample
- •In real life applications, **numerous factors can contribute to the mismatches in LID**
  - •Speech signals can be collected from different devices
  - •Speech signals can be recorded from various environments

➔Therefore, the LID system should be robust against such adverse conditions

Oriental language recognition (OLR) challenge

•The OLR challenge provides a standard benchmark for LID systems on various mismatched conditions
- •The following problems should be considered:
  - •**No in-domain data is provided** for training or validating the LID system
  - •The **primary performance metric is the C_avg** which considers the language-dependent false acceptance ratio (FAR) and false rejection ratio (FRR)
    - •**The typical identification metrics, such as accuracy, will not reflect the systems C_avg performance**

$$C_{avg} = \frac{1}{N_L}\{[C_{Miss} * P_{Target} * \sum_{L_T} P_{Miss}(L_T)] + \frac{1}{N_L - 1}[C_{FA} * (1 - P_{Target}) * \sum_{L_T}\sum_{L_N} P_{FA}(L_T, L_N)]\}$$

Deep learning-based end-to-end LID framework

- Composed of a frame-level network, pooling layer, and a classifier network
  - •**Frame-level network**: takes the acoustic feature extracted from the input speech and outputs a sequence of frame-level representations
  - •**Pooling layer**: aggregates the deep representations into an utterance-level fixed-dimensional feature (embedding vector)
  - •**Classifier**: takes the embedding vector and outputs the language probability

Backbone architectures

•In our experiments, we have adopted 3 different architectures
- •**ResNetSE34**: More specifically the Fast ResNet, which follows the same general structure as the original ResNet with 34 layers (ResNet-34) with squeeze-and-excitation. But unlike the standard ResNet-34, Fast ResNet uses only one-quarter of the channels in each residual block to reduce computational cost.
- •**ECAPA-TDNN**: An architecture that achieved state-of-the-art performance in text-independent speaker verification. The ECAPA-TDNN uses squeeze-and-excitation as in the SE-ResNet, but also employs channel- and context-dependent statistics pooling and multi-layer aggregation.
- •**Hybrid** network: A CNN-LSTM-TDNN hybrid architecture with multi-level global-local statistics pooling, which demonstrated good performance in various speaker verification tasks.

•Input acoustic features
- •**MFB**: 40 dimensional mel-filterbank energy features
- •**MFCC+pitch**: concatenation of 40 dimensional MFCC and 3 dimensional pitch features

Training objective

•In our experiments, we have used 2 different training objectives
- •**Softmax**
  - •The Hybrid system was trained using the standard softmax objective function
- •**Angular additive margin softmax (AAMSoftmax)**
  - •The ECAPA-TDNN and ResNet systems were trained using the AAMSoftmax objective function

$$L_{AAMSoftmax} = -\frac{1}{N}\sum_{i=1}^{N} log(\frac{e^{s(cos(\theta_{y_i,i}+m))}}{K_1}),$$

$$K_1 = e^{s(cos(\theta_{y_i,i}+m))} + \sum_{j=1,j\neq i}^{C} e^{scos\theta_{j,i}}.$$

Flow-based embedding regularization (Flow-ER)

•In addition, we have adopted the recently proposed Flow-ER strategy to tackle the cross-domain problem
- •The **Flow-ER** framework **regularizes the embedding network according to the information bottleneck scheme**
  - •The mutual information between the embedding and the label is maximized
  - •The mutual information between the embedding and the input representation is minimized

$$L_{IB} = -L_{xent} + \beta L_{redundancy},$$

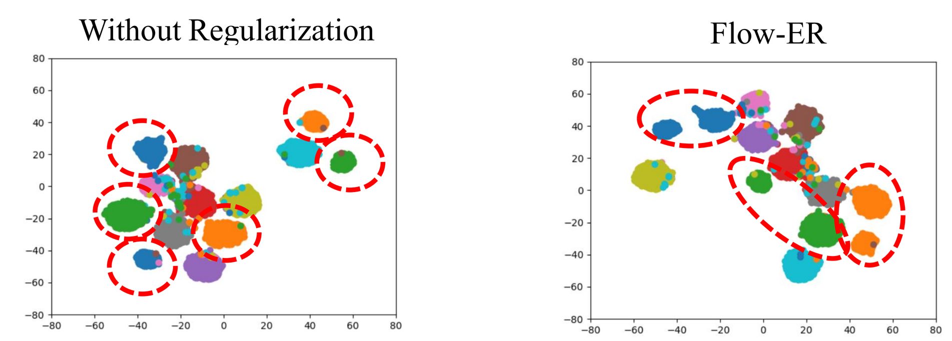
- •The regularization term is the upper-bound mutual information, which is estimated according to the contrastive log-ratio upperbound (CLUB) method

$$L_{redundancy} = E_{p(X,\omega)}[\log p_X(X|\omega)] - E_{p(X)p(\omega)}[\log p_X(X|\omega)],$$

- •The conditional likelihood is estimated using a normalizing flow model (i.e., MelFlow)

Experiments

•Embedding analysis
- From the embeddings trained without regularization, we could observe that some clusters are far away from each other if they have the same language identity
- From the Flow-ER embeddings, the clusters with the same language identity are relatively much closer to each other, and the general distribution of the embeddings is more spread out



•LID performance comparison
- The submitted systems generally performed well in terms of C_avg, but the EER was very high in some systems
  - •Such disparity between EER and C_avg is attributed to the different score statistics they consider
- The ECAPA-TDNN-based systems showed better performance than the Baseline
  - •**The best performance was achieved by the ECAPA-TDNN system trained with Flow-ER**
  - •This indicates that the Flow-ER strategy can effectively minimize the non-language information from the LID system

#	Architecture	Objective	Input	C_{avg}	EER [%]
0	Baseline			0.0826	9.038
1	Hybrid + LDA (20-dim.) + PLDA	Softmax	MFB	0.1364	47.220
2	Hybrid + LDA (30-dim.) + PLDA	Softmax	MFB	0.1360	47.290
3	Hybrid + LDA (20-dim.) + PLDA	Softmax	MFCC+pitch	0.1447	46.860
4	ResNetSE34	AAM	MFB	0.0951	10.180
5	ECAPA-TDNN	AAM	MFCC+pitch	0.0671	8.0940
6	ECAPA-TDNN	AAM + Flow-ER	MFB	0.0639	7.4370
7	**ECAPA-TDNN**	**AAM + Flow-ER**	**MFCC+pitch**	**0.0631**	**7.3340**
8	ECAPA-TDNN + LDA (20-dim.) + PLDA	AAM + Flow-ER	MFCC+pitch	0.4981	8.9400

Conclusions

•From our results, we could notice a huge disparity between the C_avg and the EER metrics, due to the different statistics they consider

•Among the experimented methods, the best performance was achieved by the ECAPA-TDNN system which takes MFCC and pitch features as input and trained using AAMSoftmax and Flow-ER strategy

•Our future research will include new methods for training the system to jointly minimize the C_avg and other metrics (e.g., EER, accuracy)

•Moreover, we will experiment with various fusion models to exploit the potential complementarity between the different LID systems