Language Recognition in ICASSP 2019

Dong Wang

2019/09/09

Language recognition

- 全世界至少有7102种语言
- 近10个语系
- 母语使用人数超过5千万的有23种

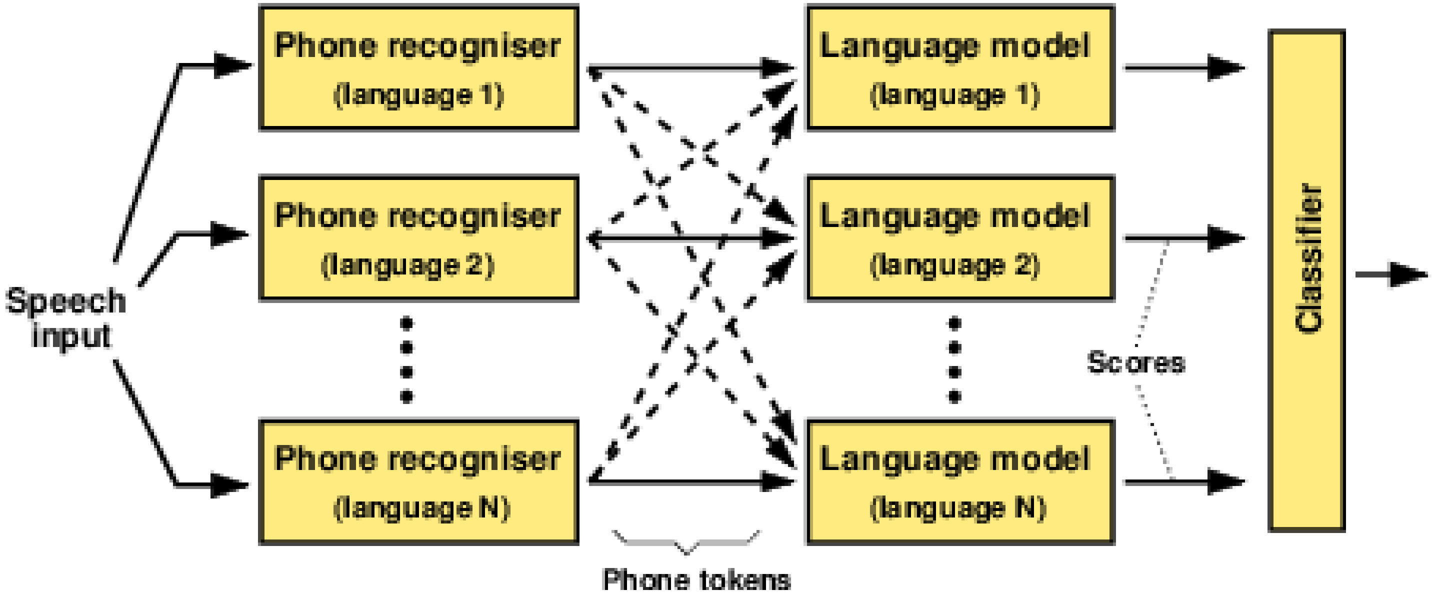
语言识别技术

- PPLM/PPRLM
- GMM/ivector
- DNN/RNN/PTN/x-vector
- Generative model, knowledge enrichement…

困难与挑战

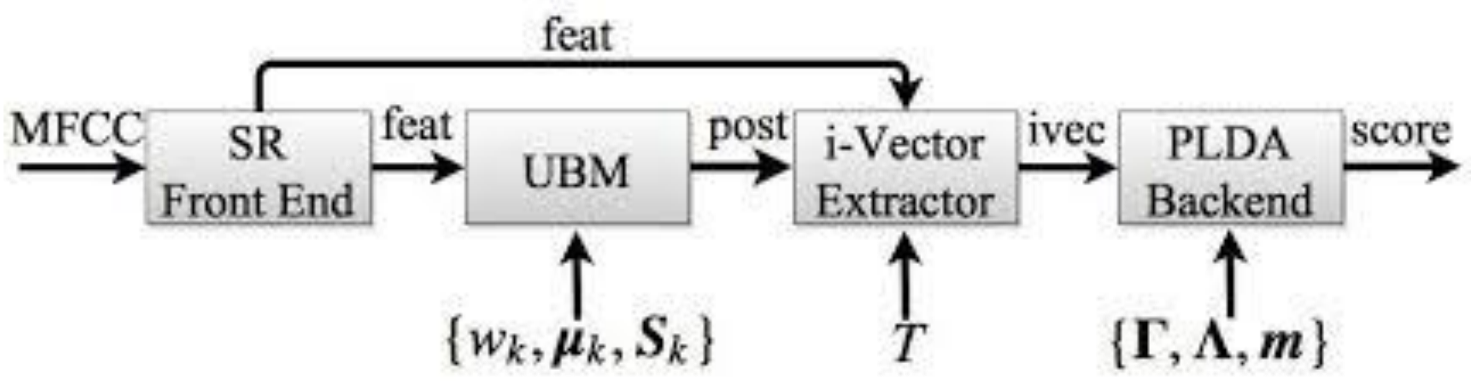
- 多场景，跨信道
- 短语音

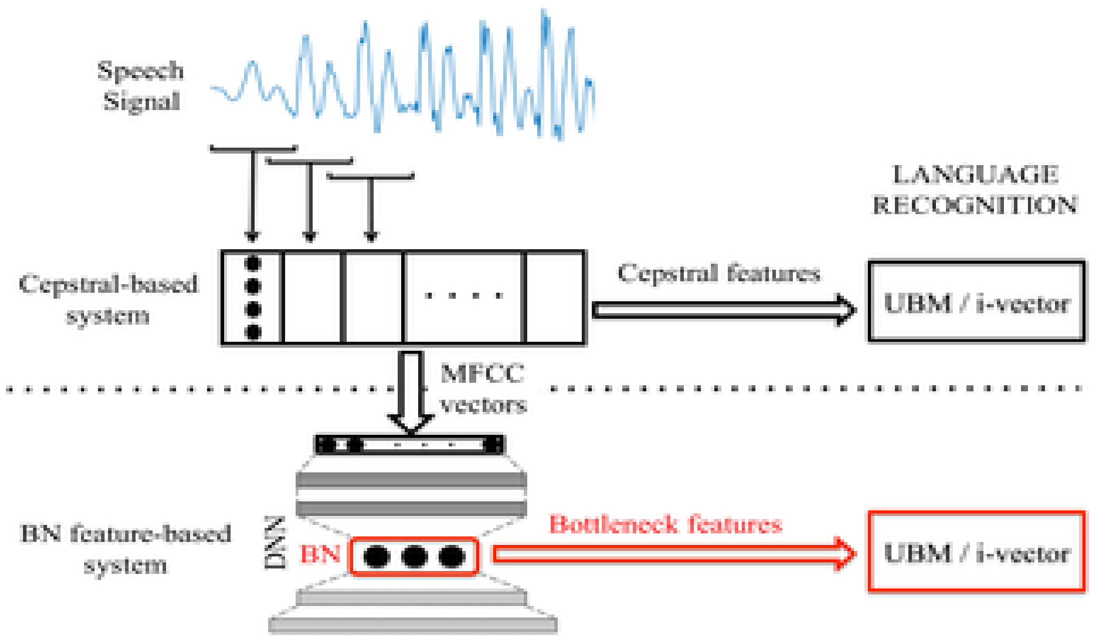
PPLM/PPRLM



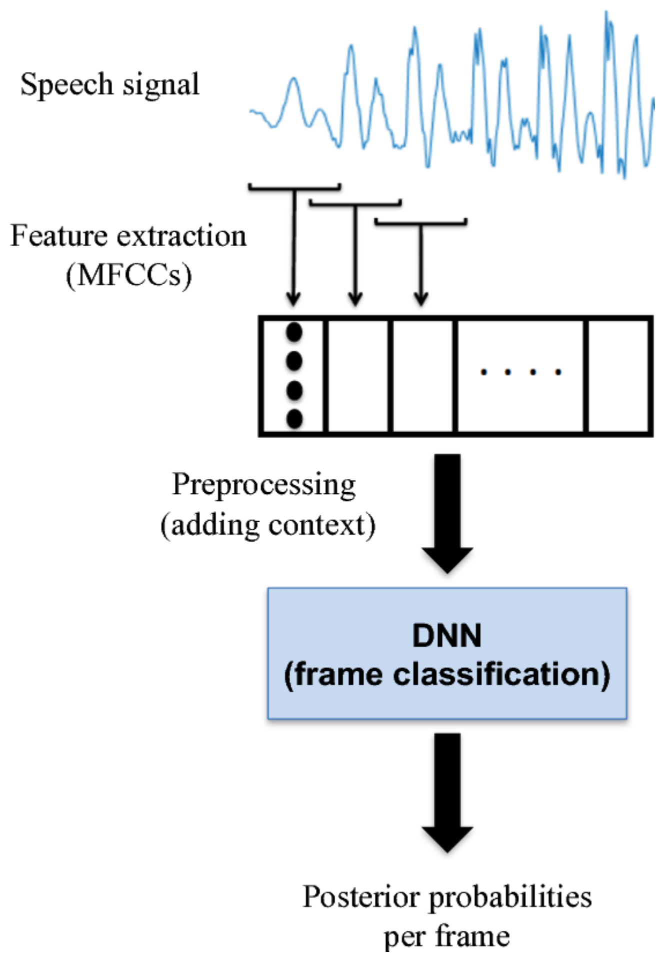
Speech input
Phone recogniser (language 1)
Phone recogniser (language 2)
Phone recogniser (language N)
Phone tokens
Language model (language 1)
Language model (language 2)
Language model (language N)
Scores
Classifier

i-vector system

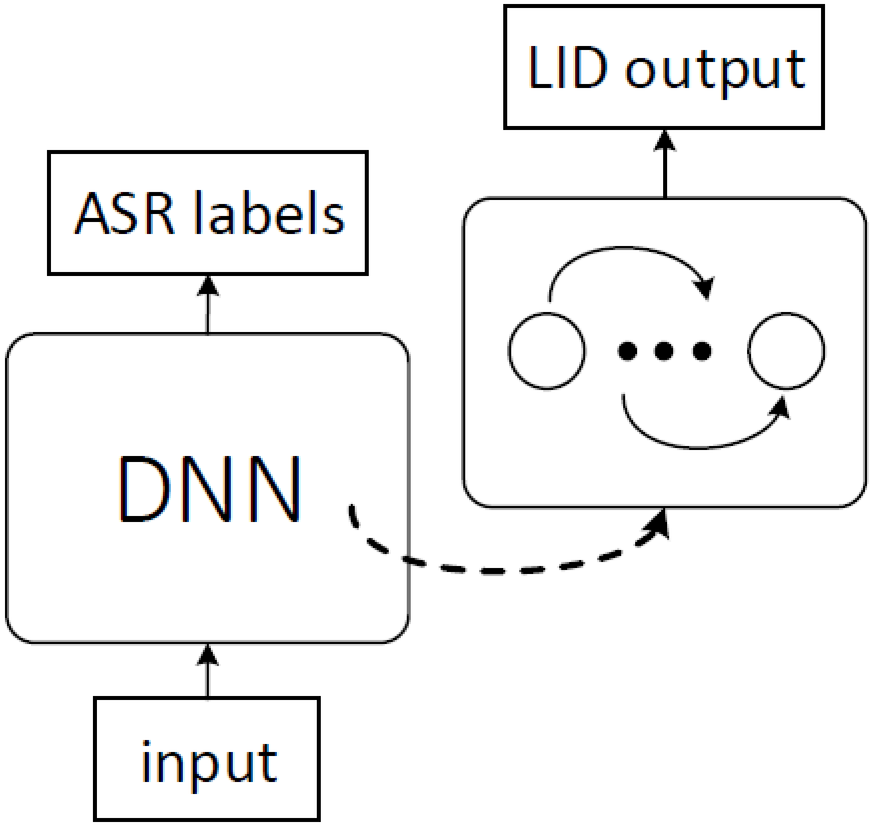




DNN system

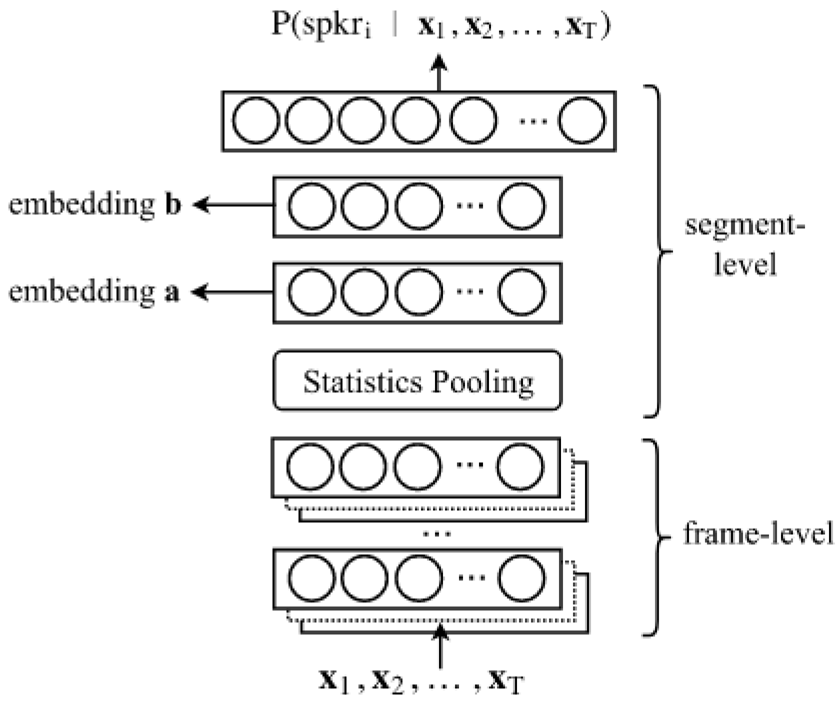


Speech signal
Feature extraction
(MFCCs)
Preprocessing
(adding context)
DNN
(frame classification)
Posterior probabilities
per frame



LID output
ASR labels
DNN
input

Statistical pooling



Attentive statistics

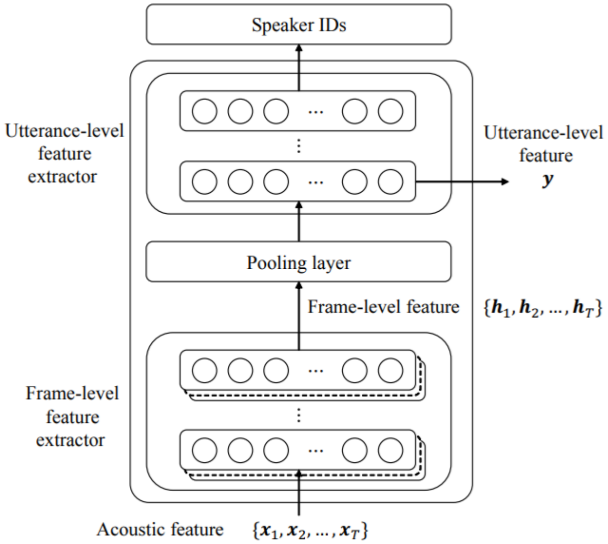


Figure 1: *DNNs for extracting utterance-level speaker features*

$$e_t = \boldsymbol{v}^T f(\boldsymbol{W}\boldsymbol{h}_t + \boldsymbol{b}) + k,$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_\tau^T \exp(e_\tau)}.$$

$$\tilde{\boldsymbol{\mu}} = \sum_t^T \alpha_t \boldsymbol{h}_t. \qquad \tilde{\boldsymbol{\sigma}} = \sqrt{\sum_t^T \alpha_t \boldsymbol{h}_t \odot \boldsymbol{h}_t - \tilde{\boldsymbol{\mu}} \odot \tilde{\boldsymbol{\mu}}},$$

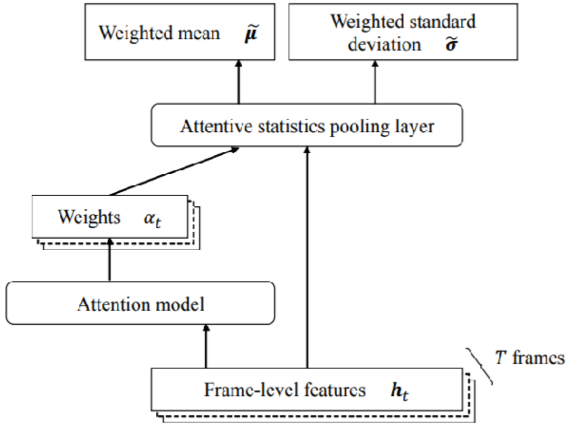


Figure 2: *Attentive statistics pooling*

Table 1: *Performance on NIST SRE 2012 Common Condition 2.* ***Boldface*** *denotes the best performance for each column.*

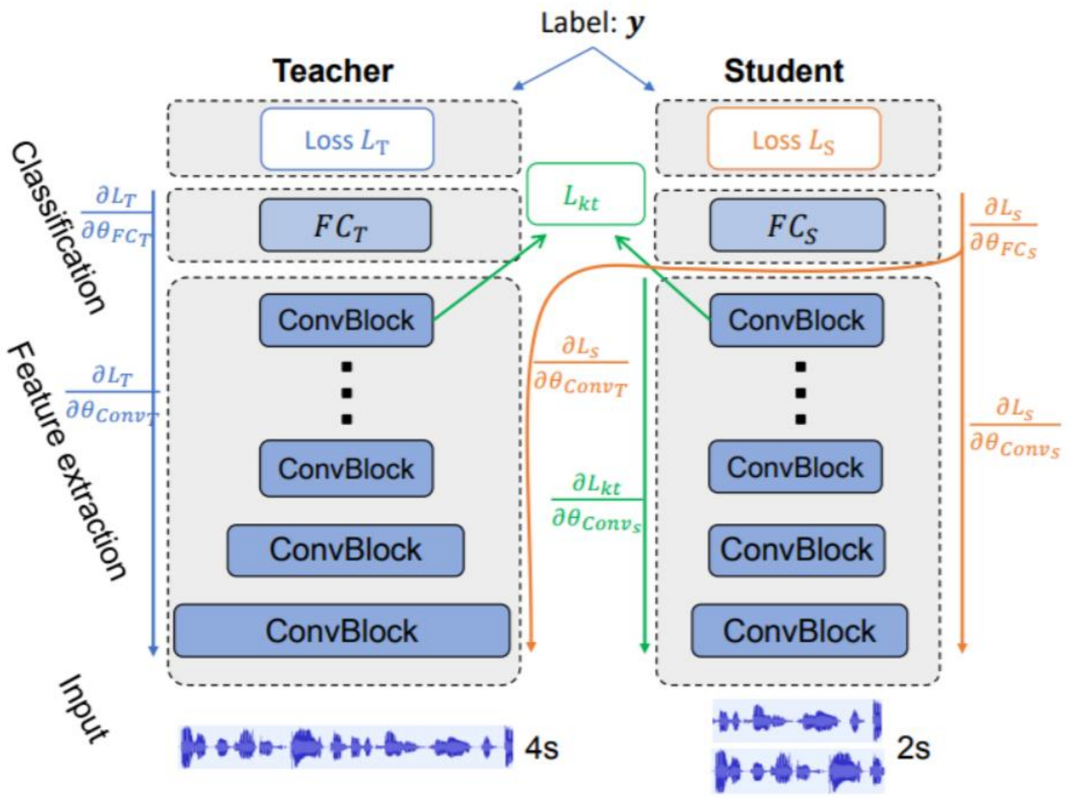
Embedding	**DCF10^{-2}**	**DCF10^{-3}**	**EER (%)**
i-vector	**0.169**	**0.291**	1.50
average [7, 8]	0.290	0.484	2.57
attention [10, 11]	0.228	0.399	1.99
statistics [9]	0.183	0.331	1.58
attentive statistics	0.170	0.309	**1.47**

Table 3: *Performance on VoxCeleb.* ***Boldface*** *denotes the best performance for each column.*

Embedding	**DCF10^{-2}**	**DCF10^{-3}**	**EER (%)**
i-vector	0.479	0.595	5.39
average [7, 8]	0.464	0.550	4.70
attention [10, 11]	0.443	0.598	4.52
statistics [9]	0.413	0.530	4.19
attentive statistics	**0.406**	**0.513**	**3.85**

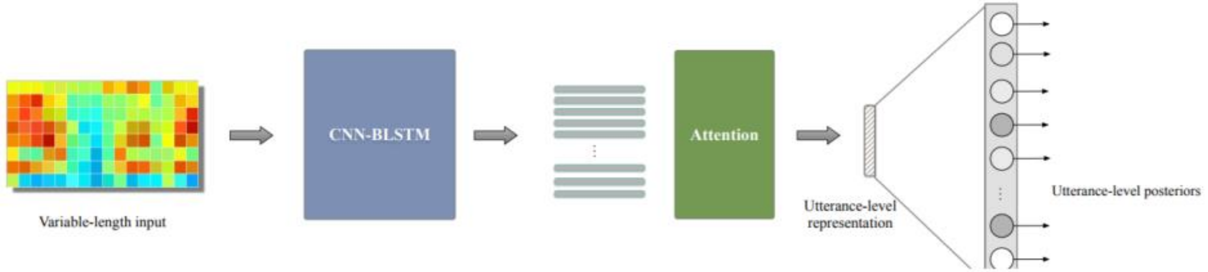
Attentive Statistics Pooling for Deep Speaker Embedding, 2018.

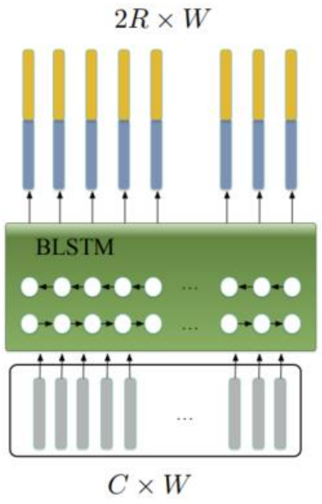
Interactive transfer learning

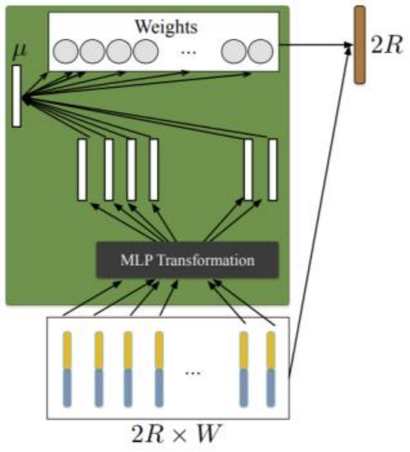


INTERACTIVE LEARNING OF TEACHER-STUDENT MODEL FOR SHORT UTTERANCE SPOKEN LANGUAGE IDENTIFICATION, ICASSP 2019.

Attention pooling

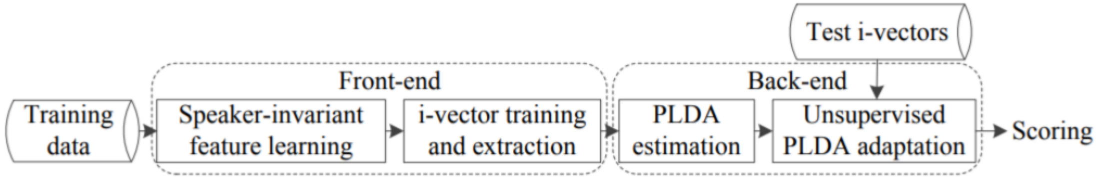


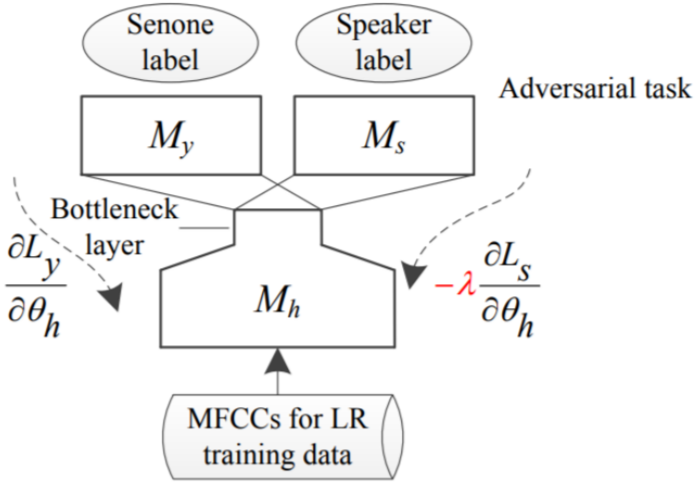




UTTERANCE-LEVEL END-TO-END LANGUAGE IDENTIFICATION USING ATTENTION-BASED CNN-BLSTM, ICASSP 2019.

Adversarial training





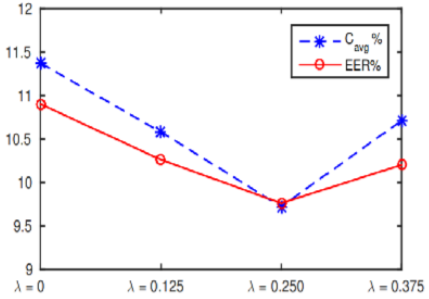


Fig. 3. C_{avg}/EER% results by employing speaker AMTL-DNN BNFs on dev_1s. Back-end is cosine scoring.

Table 4. EER% results with/without unsupervised PLDA adaptation. Back-end is PLDA.

	No Adapt.	Adapt. with cluster number					SOTA [41]
		10	50	100	200	500	
Dev_1s	7.56	6.84	6.65	**6.49**	6.99	7.26	N/A
Test_1s	8.78	—	—	7.53	—	—	7.91

ADVERSARIAL MULTI-TASK DEEP FEATURES AND UNSUPERVISED BACK-END ADAPTATION FOR LANGUAGE RECOGNITION, CUHK

Multi-time scale modeling

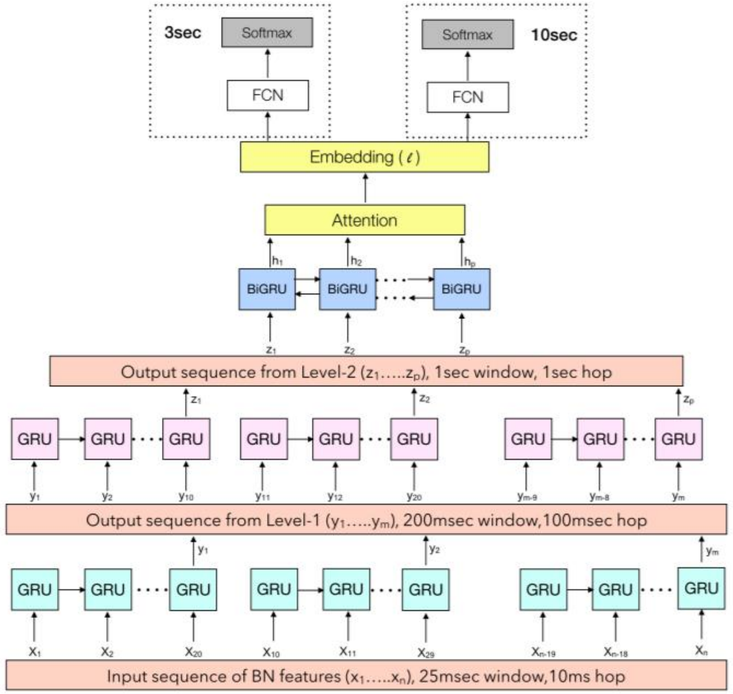


Fig. 1. End-to-end Hierarchical GRU RNN with attention module and duration dependent target layers.

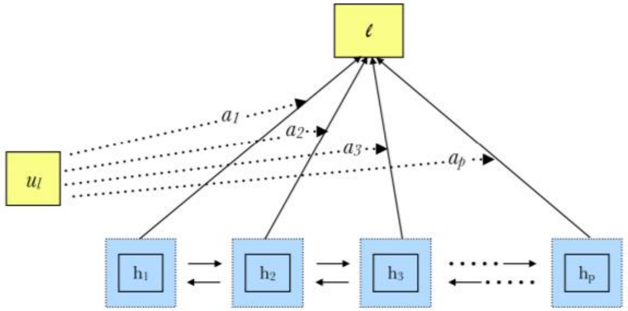


Fig. 2. Attention Mechanism in HGRU.

Table 1. LRE2017 evaluation results on clean evaluation data in terms of accuracy in % (and Cavg in parenthesis) for baseline system [20], LSTM model [16] and the proposed HGRU model.

Dur. (sec)	ivec [20]	LSTM [16]	HGRU
3	53.8 (0.53)	54.7 (0.55)	**55.1** (0.55)
10	72.3 (0.27)	72.1 (0.35)	**74.1** (0.32)
30	83.0 (0.13)	76.1 (0.28)	**83.0** (0.23)
1000	**56.2** (0.54)	42.8 (0.79)	53.5 (0.62)
overall	67.9 (0.37)	64.3 (0.48)	**68.5** (0.42)

END-TO-END LANGUAGE RECOGNITION USING ATTENTION BASED HIERARCHICAL GATED RECURRENT UNIT MODELS, ICASSP 2019.

Unsuperivsed feature & semi-learned GAN

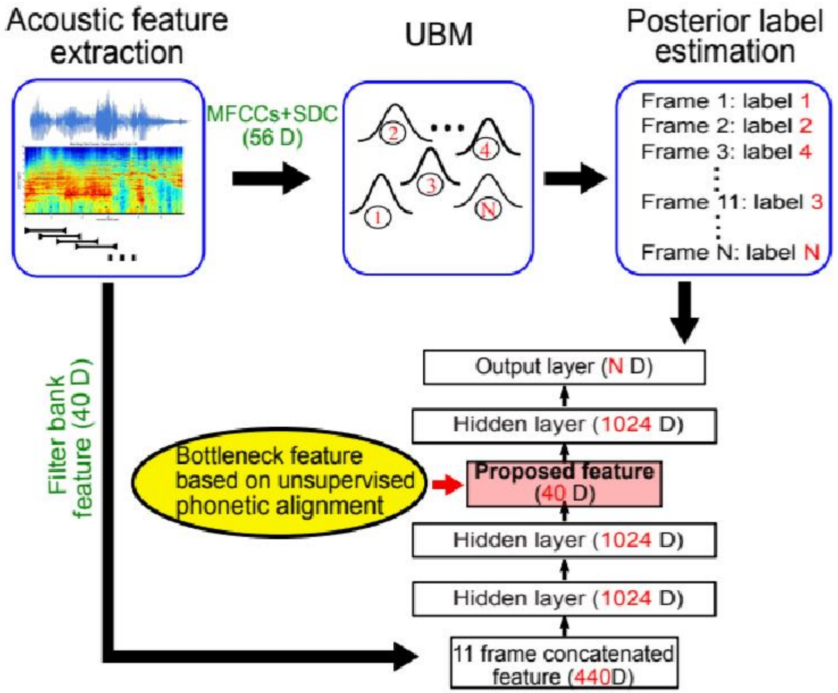


Fig. 1. The UBNF feature extraction diagram. *Sigmoid* non-linearity is used with softmax normalization for the output layer of DNN.

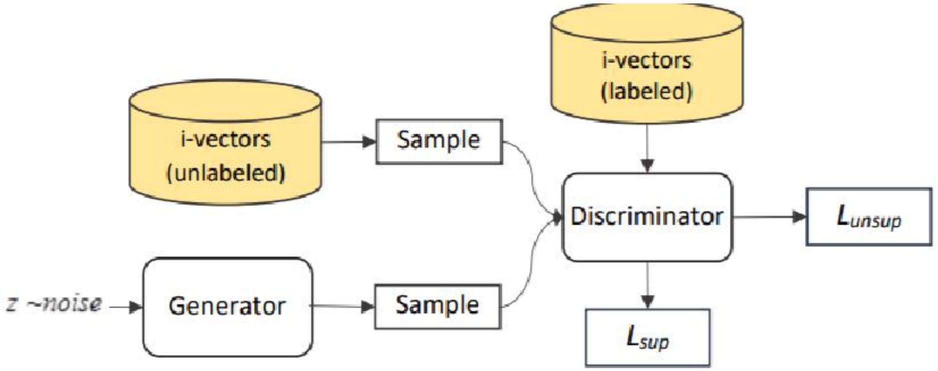


Fig. 2. A conceptual semi-supervised learning framework with GANs. The "feature matching" trick is also employed to construct the generator loss, as proposed in [22].

$$
\begin{aligned}
L = &- \mathbb{E}_{\mathbf{x},y\sim p_d(\mathbf{x},y)}[\log p_m(y|\mathbf{x})] \\
&- \mathbb{E}_{\mathbf{x}\sim G}[\log p_m(y = K+1|\mathbf{x})],
\end{aligned}
$$

SEMI-SUPERVISED LEARNING WITH GENERATIVE ADVERSARIAL NETWORKS FOR ARABIC DIALECT IDENTIFICATION, ICASSP 2019.

Highlights

- High-level feature extraction by CNN
- Temporal feature extraction by LSTM
- Attentive statistical pooling
- Adverserial training to purify other factors