Language identification and Emotion recognition

Reporter: 齐诏娣
Email: zdqi0707@163.com
Data: 2018.11.21

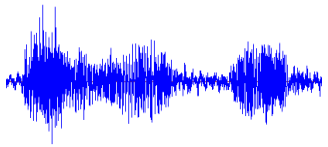
outline

- 语种识别（LID）概述
- 基于音素识别器的语种识别
- 声学特征的语种识别
- 神经网络的语种识别

- 情感识别探讨

语种识别概述

- 是指利用计算机自动判定给定语音片段所属语言种类过程。



基于音素识别器的语种识别

- PRLM

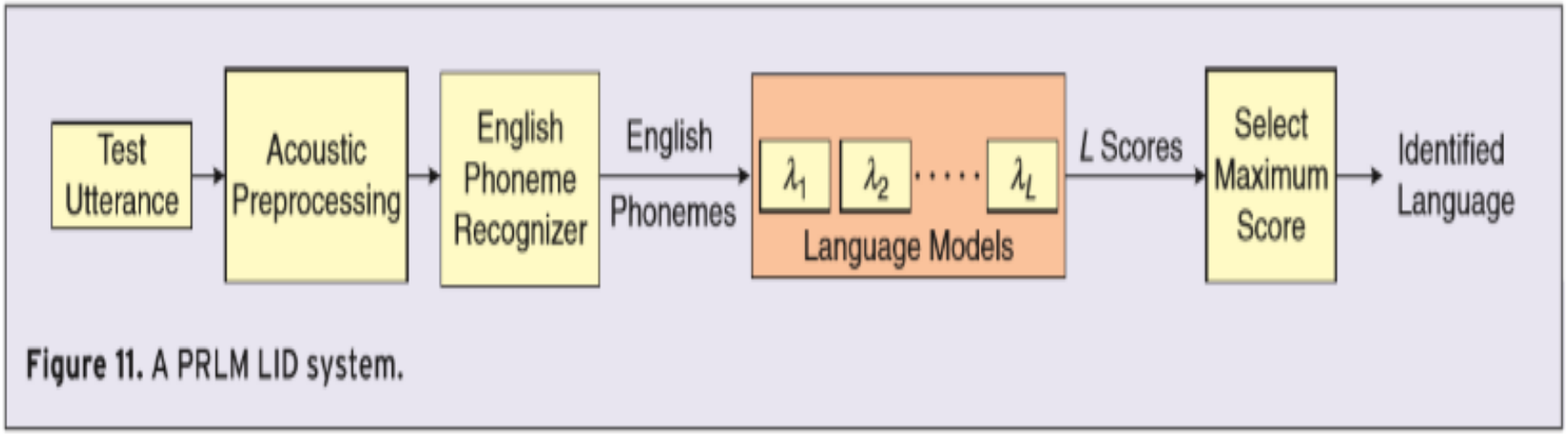


Figure 11. A PRLM LID system.

- PPRLM

基于声学特征的语种识别

- 基本操作

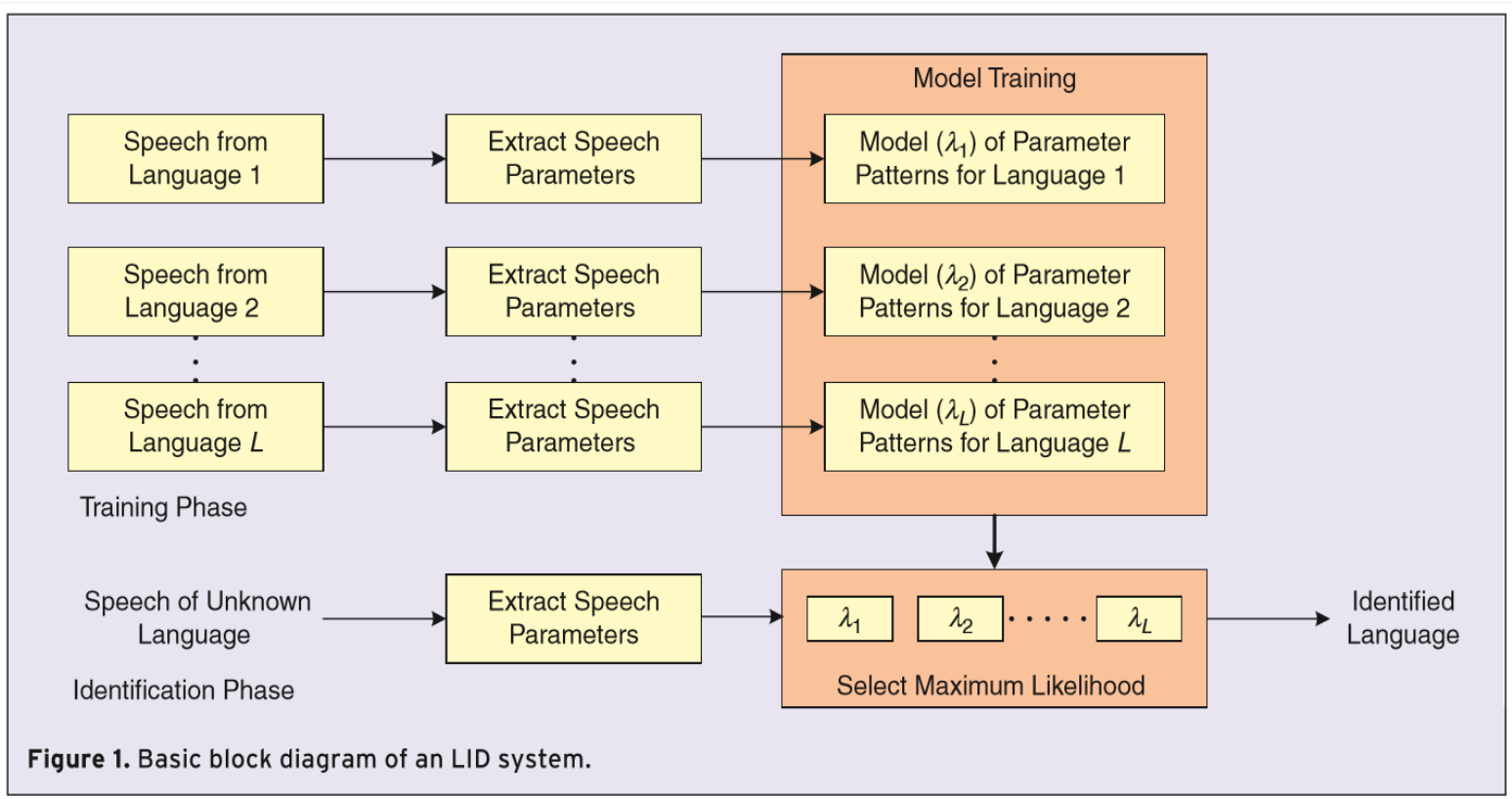


Figure 1. Basic block diagram of an LID system.

- 可用特征

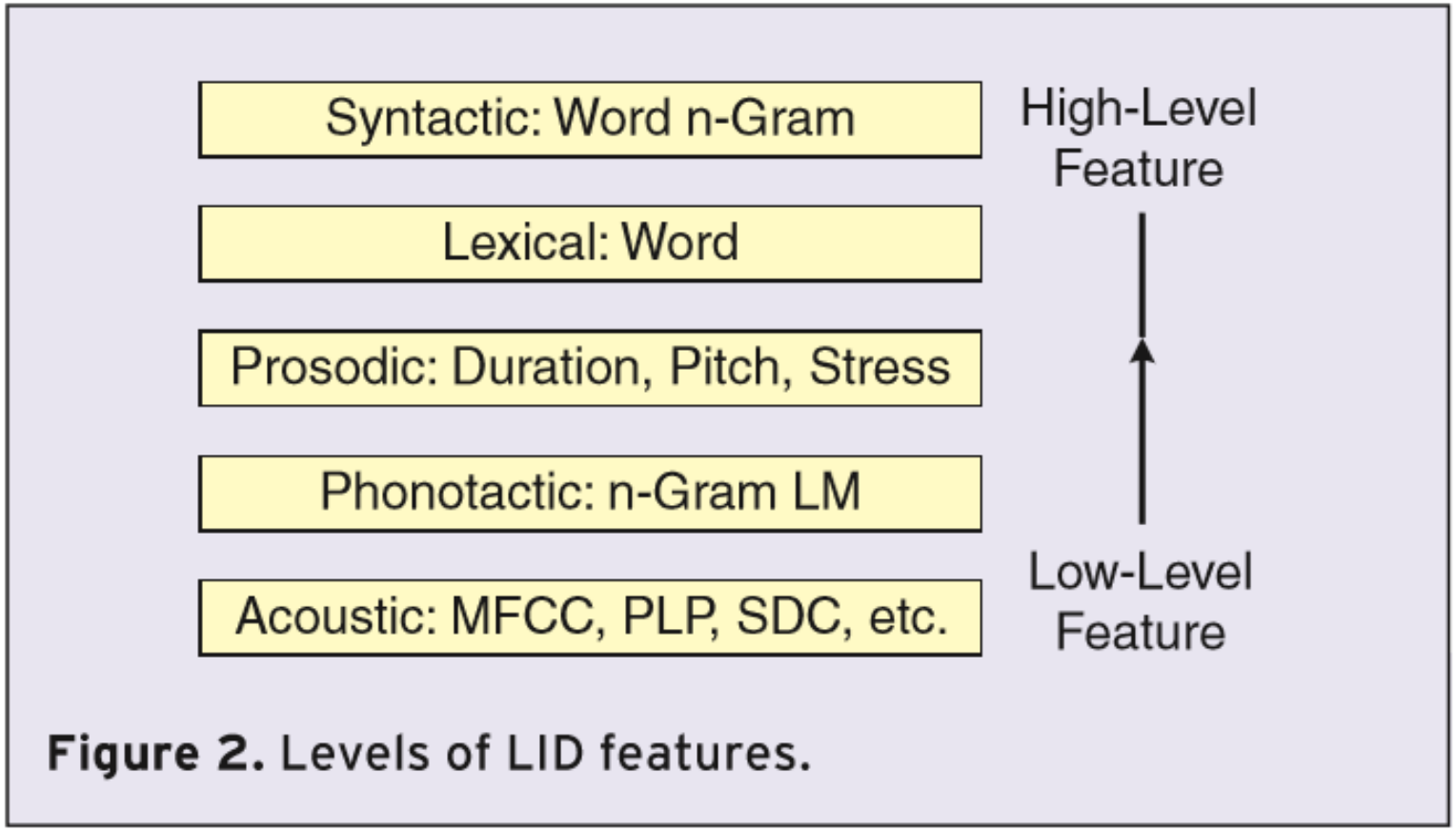
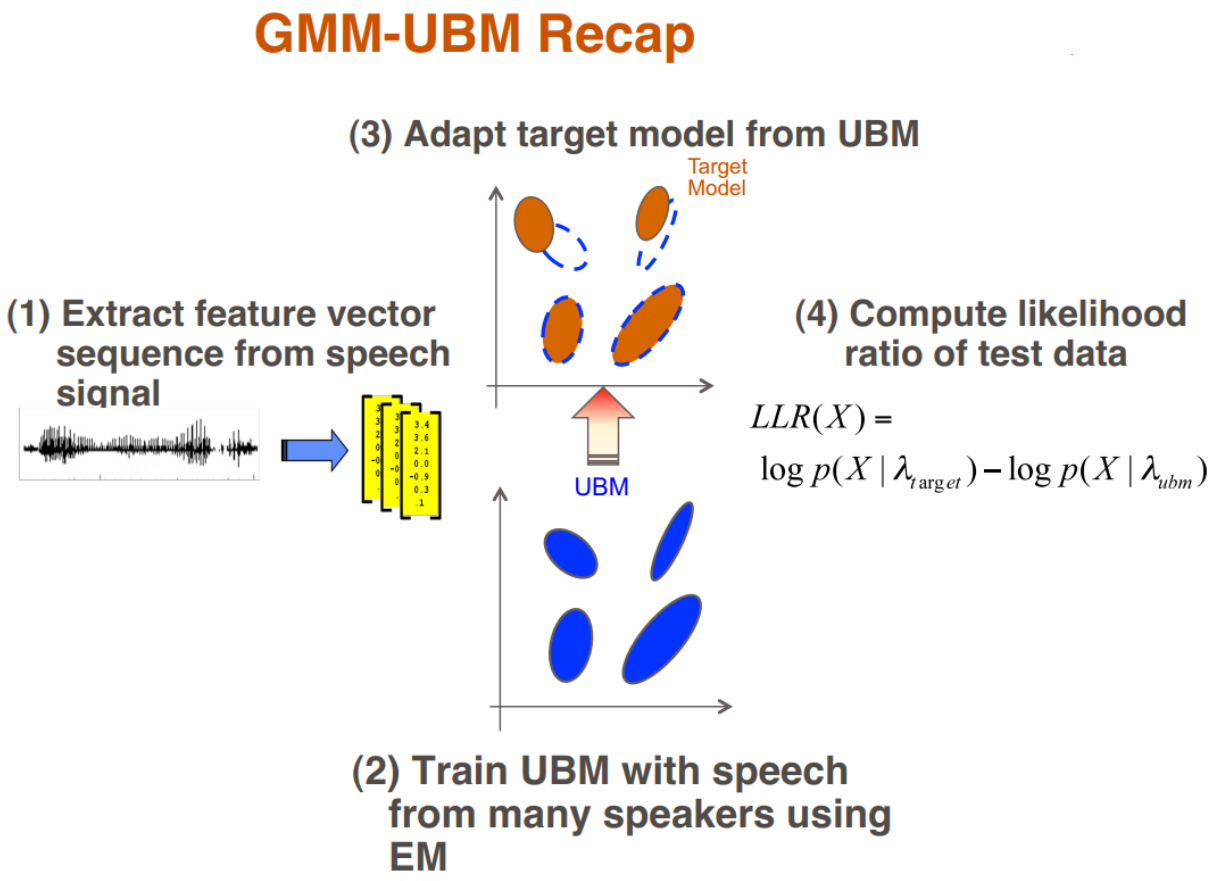


Figure 2. Levels of LID features.

- i-vector


GMM-UBM Recap
(3) Adapt target model from UBM
Target Model
(1) Extract feature vector sequence from speech signal
(4) Compute likelihood ratio of test data
$LLR(X) = \log p(X \mid \lambda_{target}) - \log p(X \mid \lambda_{ubm})$
UBM
(2) Train UBM with speech from many speakers using EM

Total variability

- Factor analysis as feature extractor
- Joint factor analysis

$$\mathbf{M} = \boldsymbol{m} + \boldsymbol{Vy} + \boldsymbol{Dz} + \boldsymbol{Ux}$$

- Speaker and channel dependent supervector

$$\mathbf{M} = \boldsymbol{m} + \boldsymbol{Tw}$$

  - T is rectangular, low rank (total variability matrix)
  - w standard Normal random (total factors – intermediate vector or i-vector)

- 子带包络特征

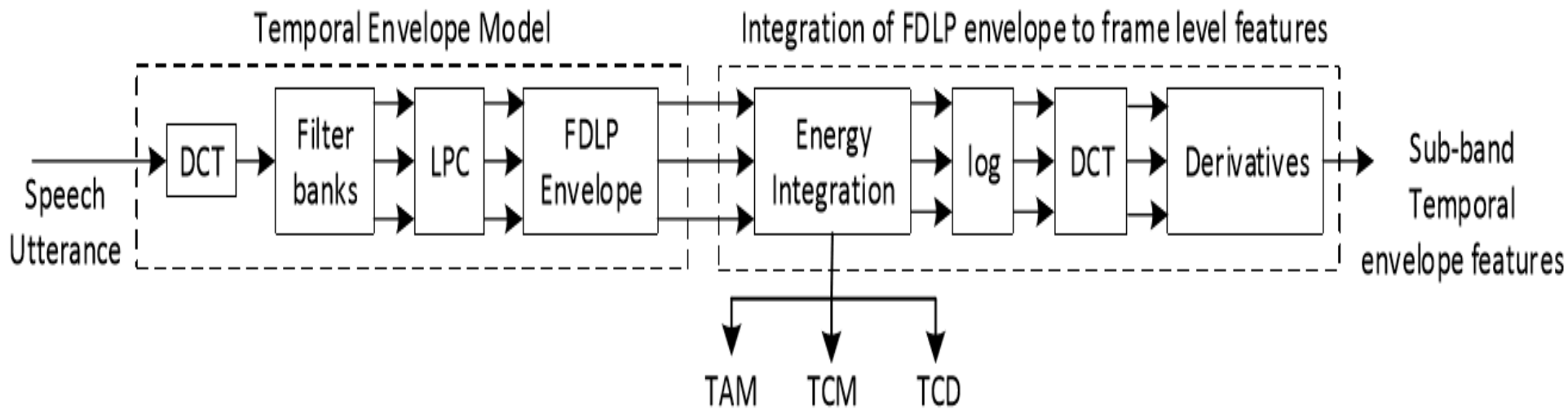


Figure 1: *Proposed feature extraction schematic with three types of energy integration methods to compute envelope features.*

- IFCC

Table 2. Evaluation of acoustic features for SLR in terms of percentage of EER and $minC_{avg}$ (reported within parenthesis).

Features	DEV17		EVAL15
	MLS14	VS	MLS14
SDCC	10.22 (0.359)	6.49 (0.216)	11.82 (0.421)
IFCC	11.41 (0.374)	12.58 (0.421)	15.51 (0.501)
DBN	5.97 (0.218)	4.08 (0.143)	6.75 (0.249)
SDCC+IFCC	**7.15 (0.251)**	**5.32 (0.188)**	**9.44 (0.340)**
DBN+IFCC	**4.60 (0.166)**	**3.42 (0.129)**	**5.97 (0.222)**

2018, on the importance of analytic phase of speech signals in spoken language recognition

基于神经网络的语种识别

- ivector → DNN

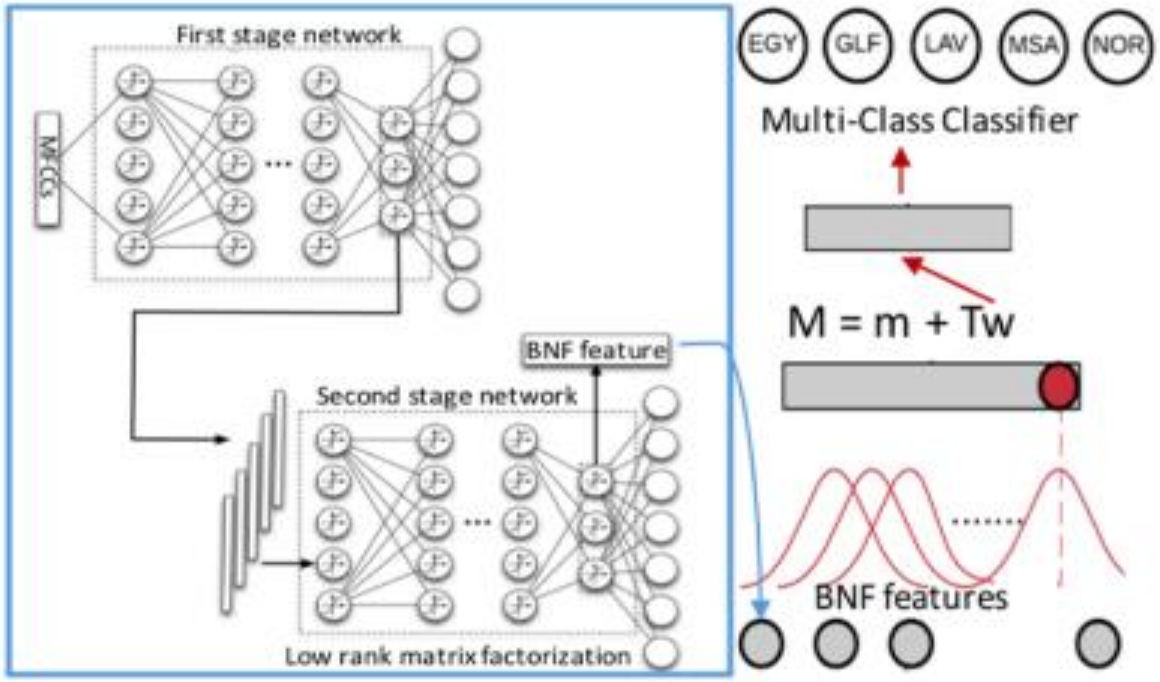


Figure 3: *I-vector based DID system*

2018, iccasp, Exploiting convolutional neural networks for phonotactic based dialect identification

- RNN→LM

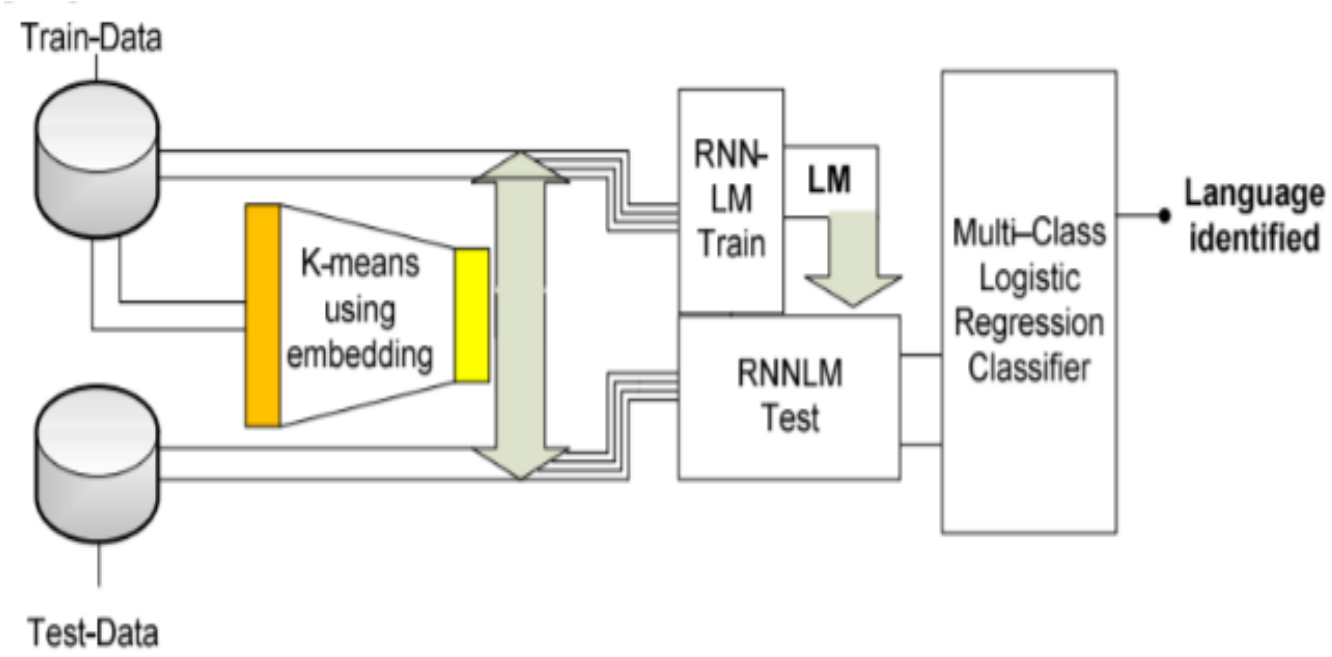
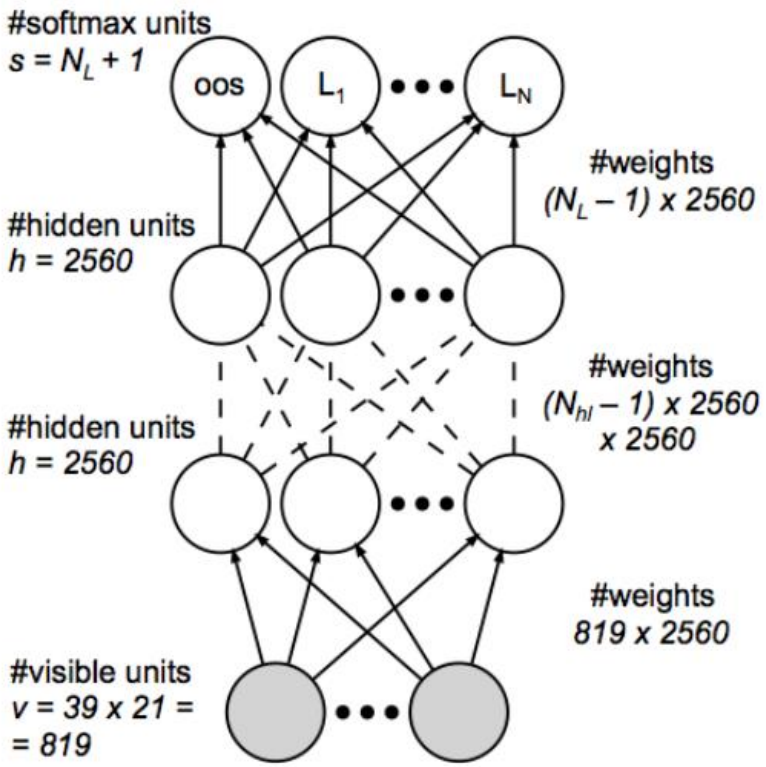


Figure 3: *LID system used to reduce the vocabulary size of phone-grams.*

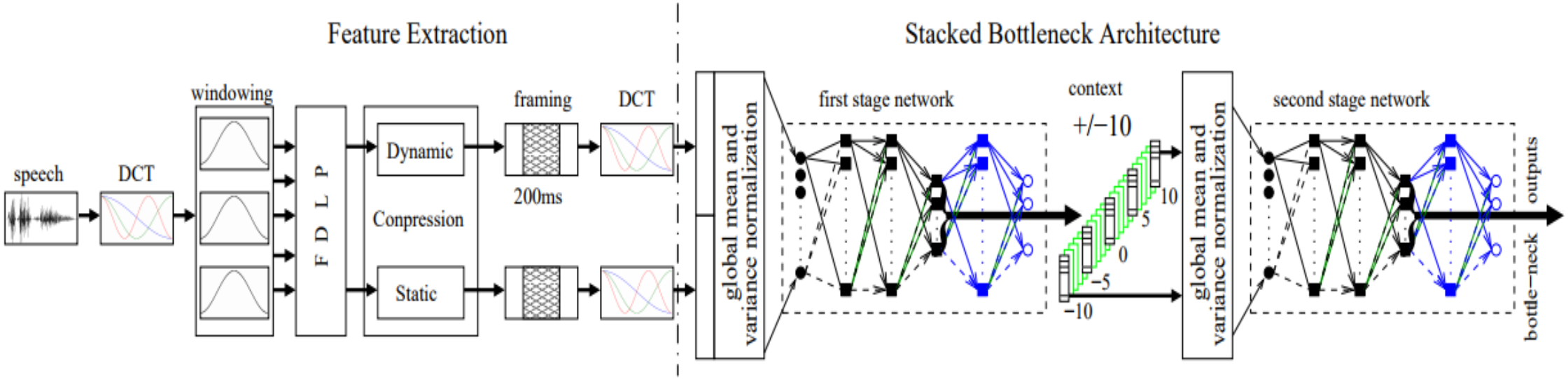
2016, On the use of phone-gram units in recurrent neural networks for language identification

- DNN



2014,iccasp,Lopez-Moreno, Ignacio, et al. "Automatic language identification using deep neural networks."

- BNF



2014, Matejka, Pavel, et al. "Neural network bottleneck features for language identification."

- RNN

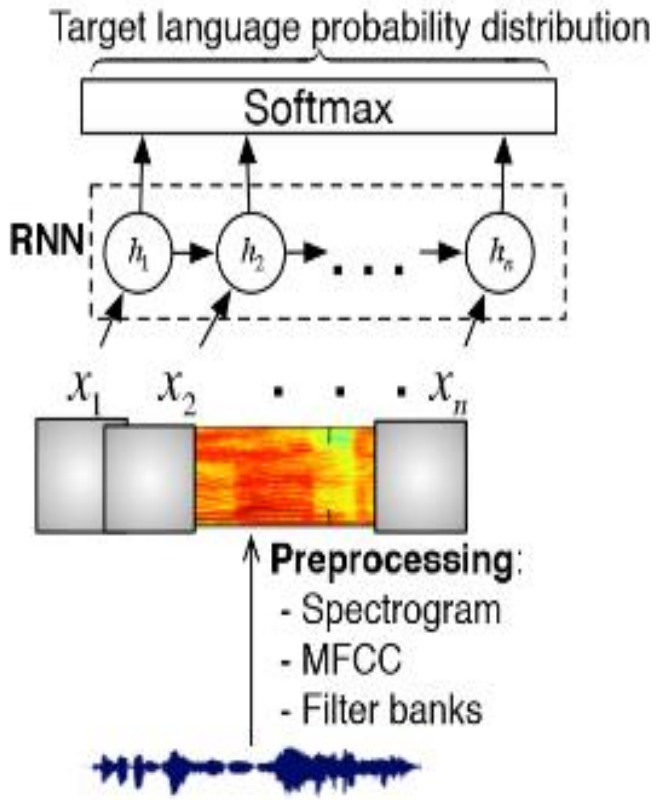


Figure 1: *Design for an end-to-end LID system using RNN.*

2016, a comprehensive deeplearning approach to end-to-end language recognition

- LSTM

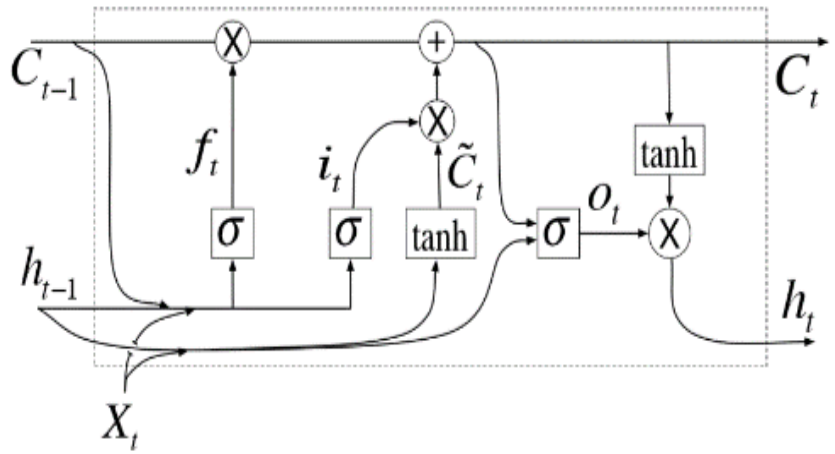


Figure 2: *A LSTM architecture, as a flow of information through memory block which controlled by input gate* $\boldsymbol{i}_t$*, forget gate* $\boldsymbol{f}_t$ *and output gate* $\boldsymbol{o}_t$

- Attention-based RNN

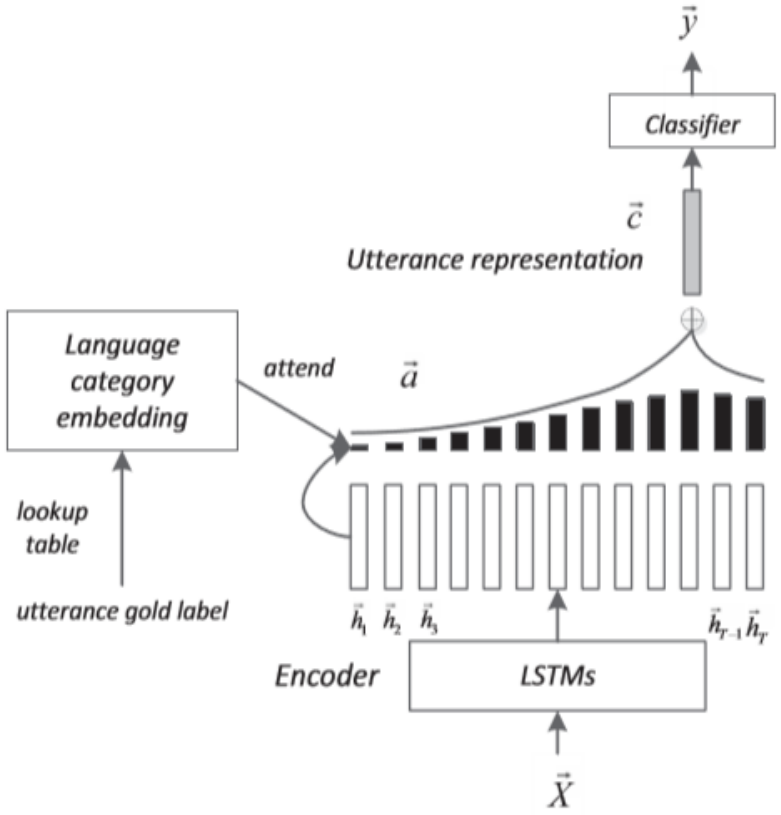


Figure 1: *The Architecture of attention-based recurrent neural network.*

2016, interspeech End to-end language identification using attention-based recurrent neural networks

Table 1: *System performance of different models (EER %) on LRE 2007 (3s segments).*

model	EER(%)
i-Vector	20.39
LSTM RNNs	16.03
Attention model	14.72

- CNN

- TDNN

- GAN

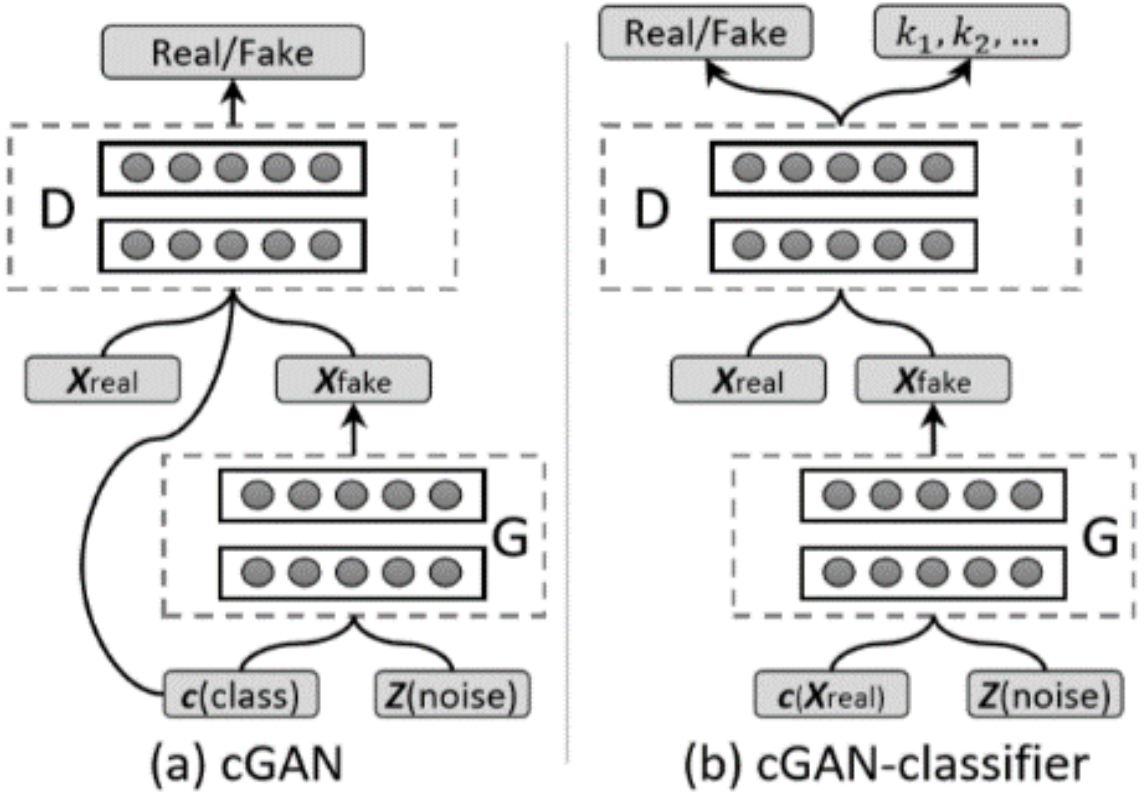


Figure 1: *cGAN (a): Conditional GAN and cGAN-classifier (b): conditional GAN-based classifier.*

2017, Conditional Generative Adversarial Nets Classifier for Spoken Language Identification

- Siamese

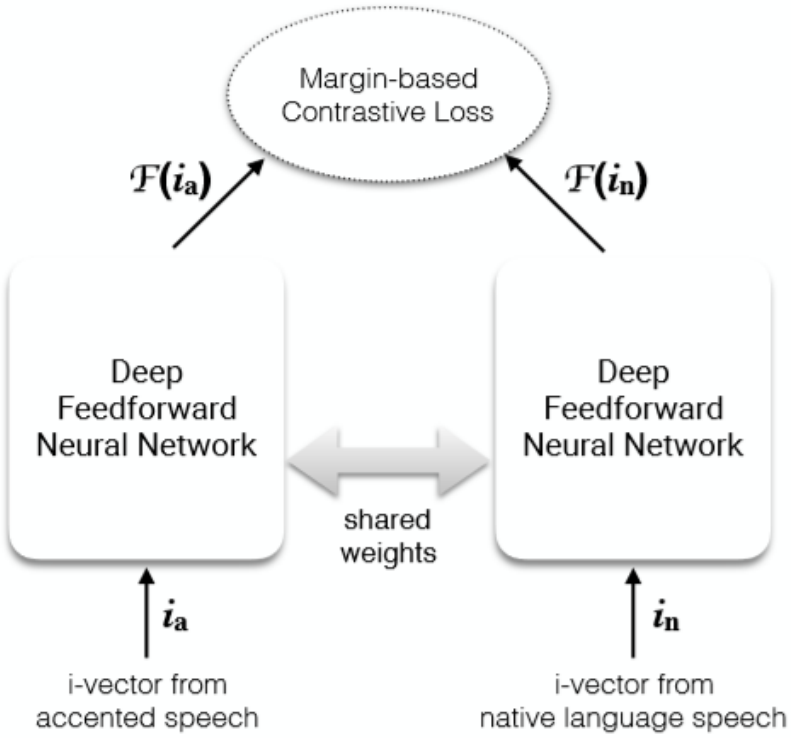


Fig. 1. Siamese network architecture for accent identification

2017, Leveraging Native Language Speech or Accent identification using deep Siamese networks

- 层次架构

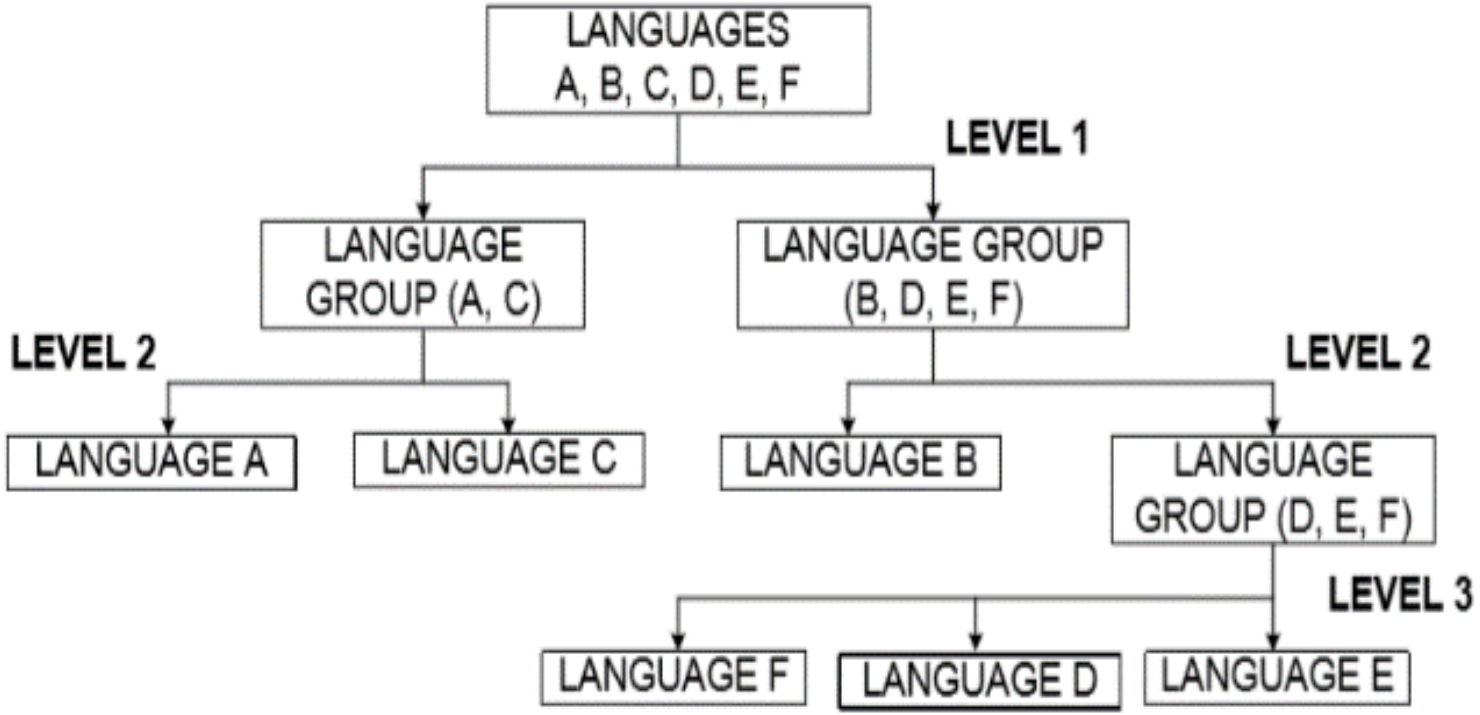


Figure 1: Hierarchical Framework for Language Identification

2016 iccasp, Hierarchical framework for language identification

- embeddings

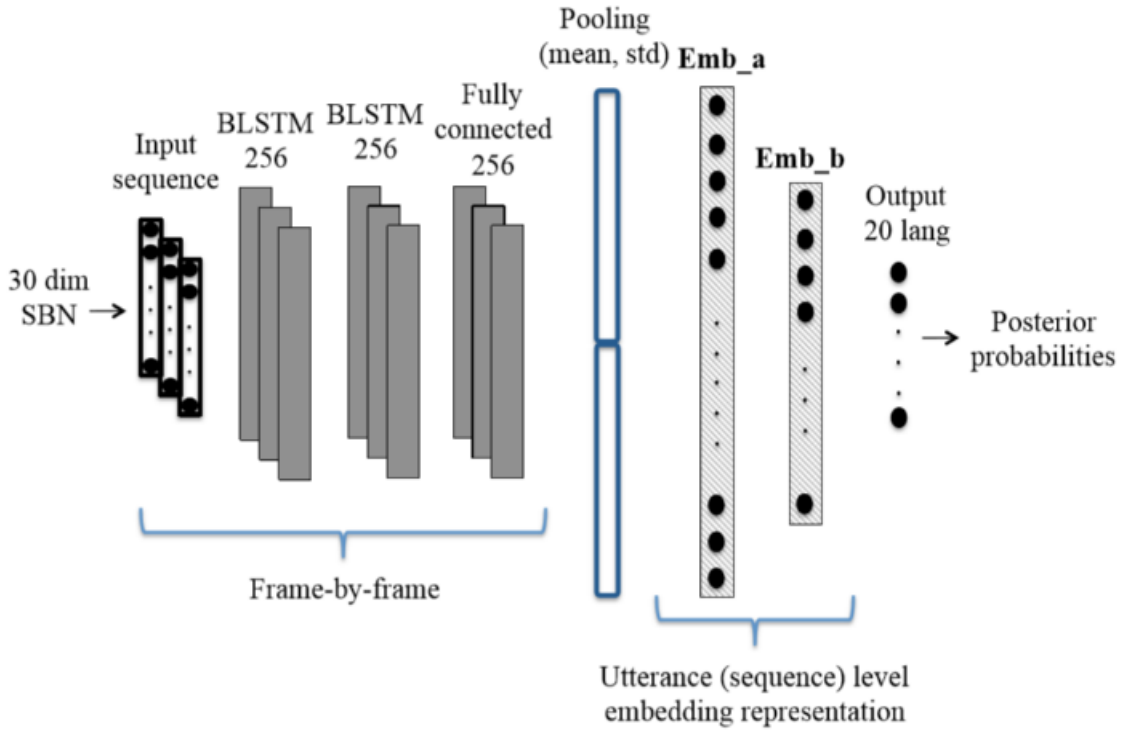


Fig. 1. Architecture of the proposed DNN for language recognition with embeddings.

2018,iccasp, dnn based embeddings for language recognition

- PTN

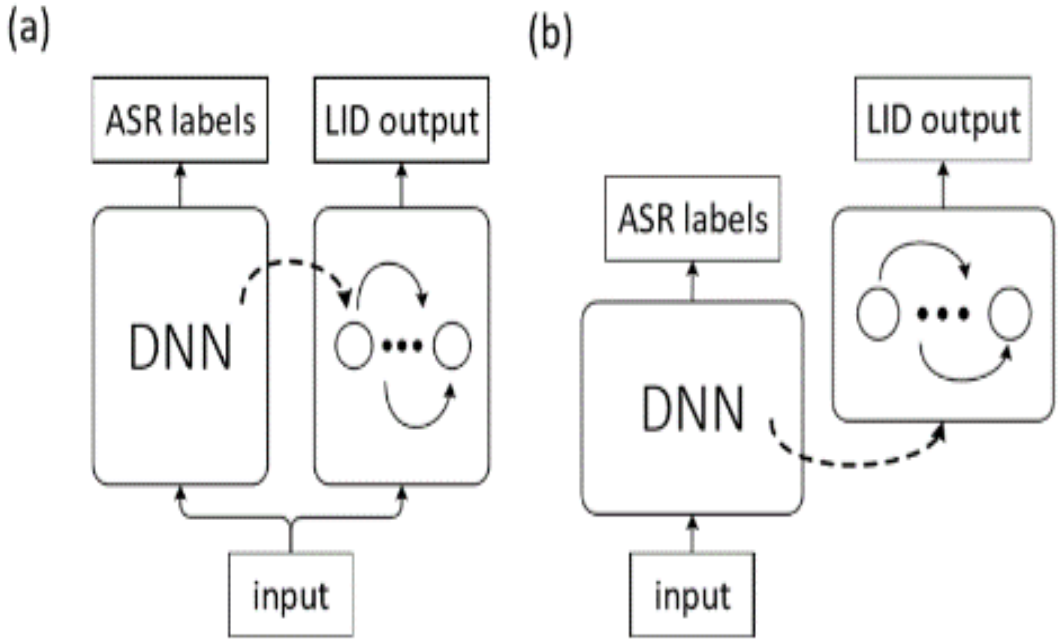


Fig. 1: LID models employing phonetic information: (a) the phonetically aware model; (b) the PTN model. Both models consist of a phonetic DNN (left) to produce phonetic features and an LID RNN (right) to make LID decisions.

2017, Phonetic temporal neural model for language identification

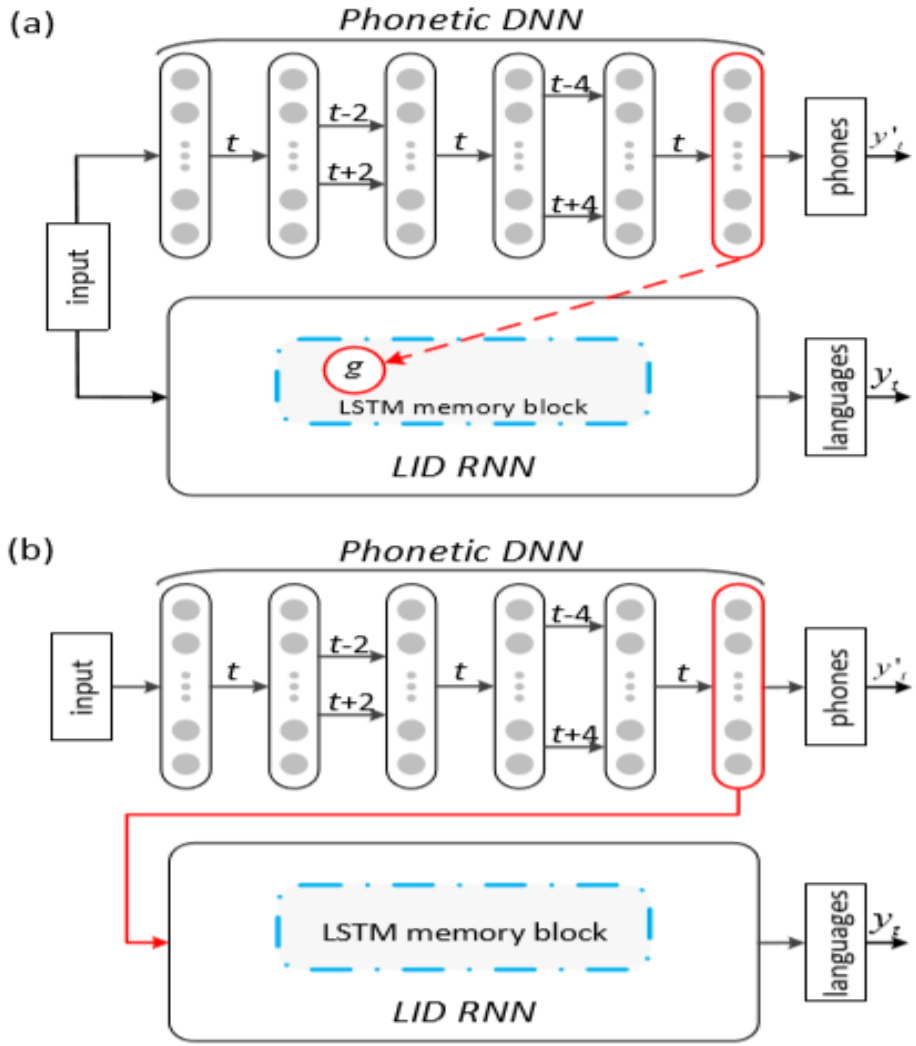


Fig. 3: The phonetically aware RNN LID system (top) and the PTN LID system (bottom). The phonetic feature is read from the last hidden layer of the phonetic DNN which is a TDNN. The phonetic feature is then propagated to the g function for the phonetically aware RNN LID system, and is the only input for the PTN LID system.

- PPRLM

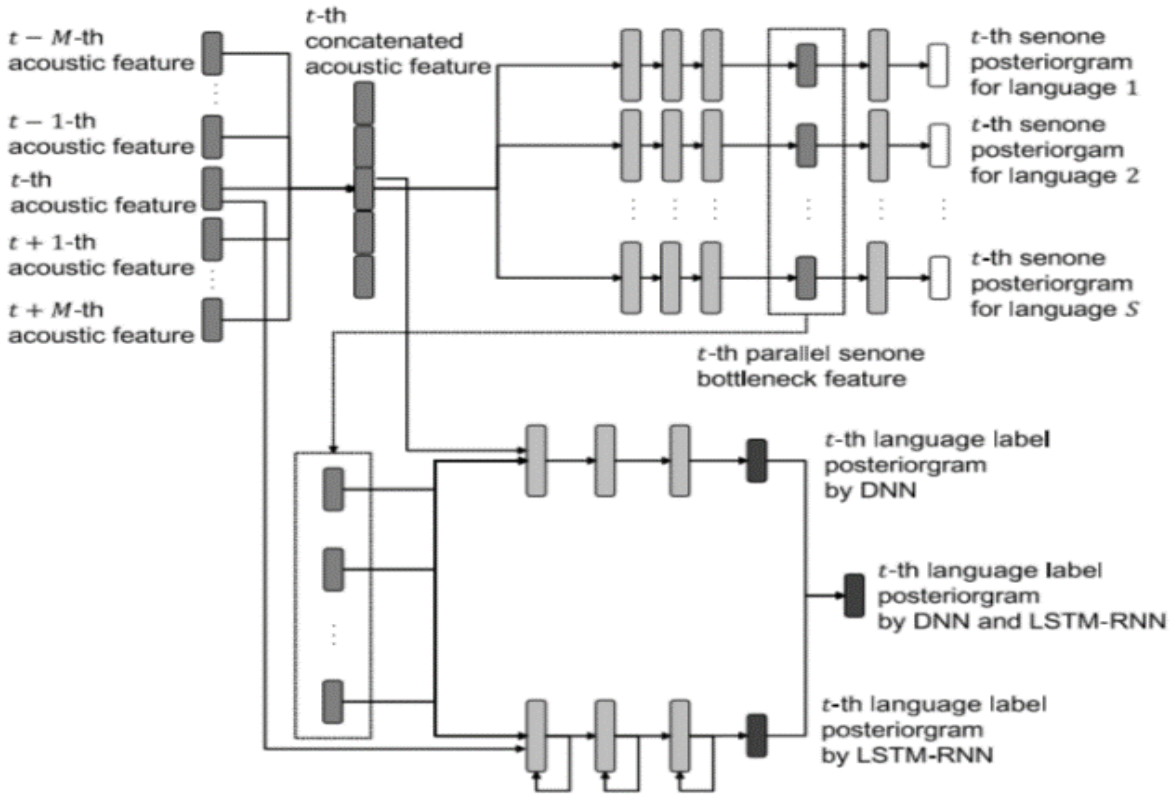
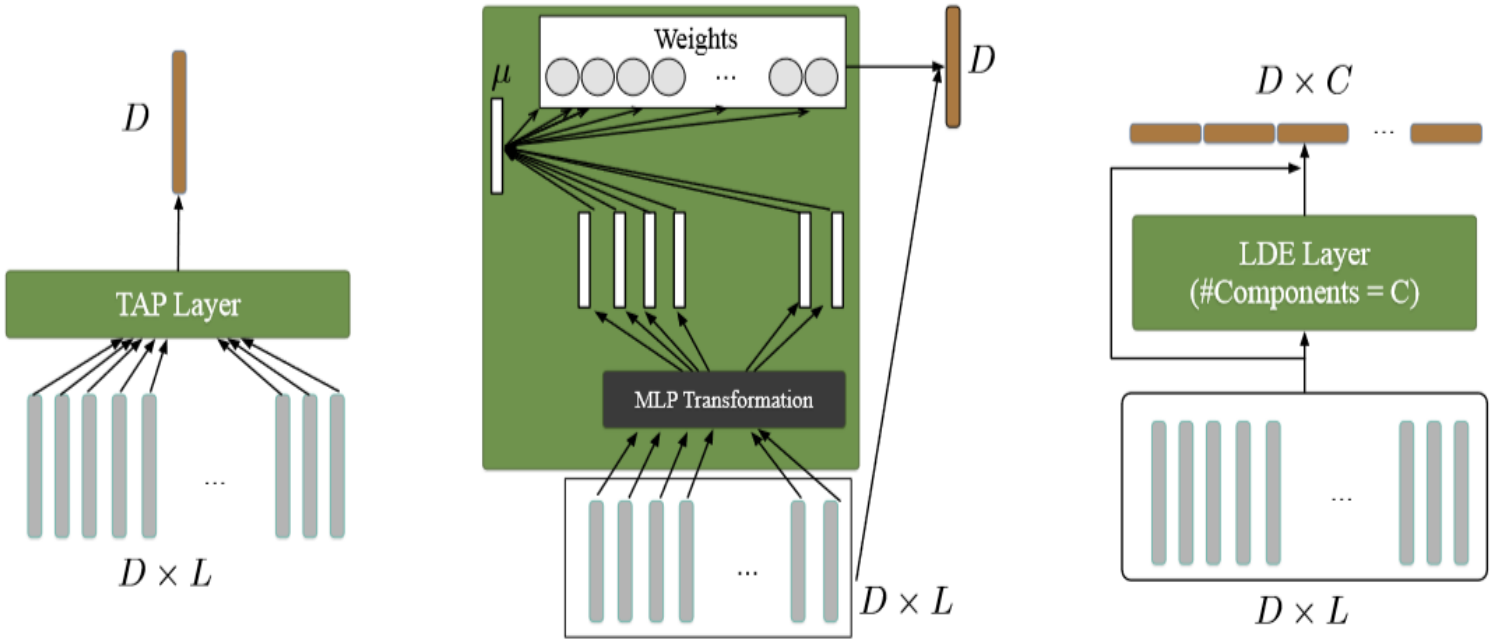


Fig. 1. *PPA-DNNs and PPA-LSTM-RNNs based on parallel senone bottleneck feature extraction.*

2017,iccasp,PARALLEL PHONETICALLY AWARE DNNS AND LSTM-RNNS FOR FRAME-BY-FRAME DISCRIMINATIVE MODELINGOF SPOKEN LANGUAGE IDENTIFICATION

- 不定长？

- 补0
- TAP
- SAP
- Recurrent encoding layer
- LDE



(a) TAP layer (b) SAP layer (c) LDE layer

Figure 5: Comparison of different encoding procedures

2018,iccasp, Exploring the Encoding Layer and Loss Functionin End-to-End Speaker and Language Recognition System

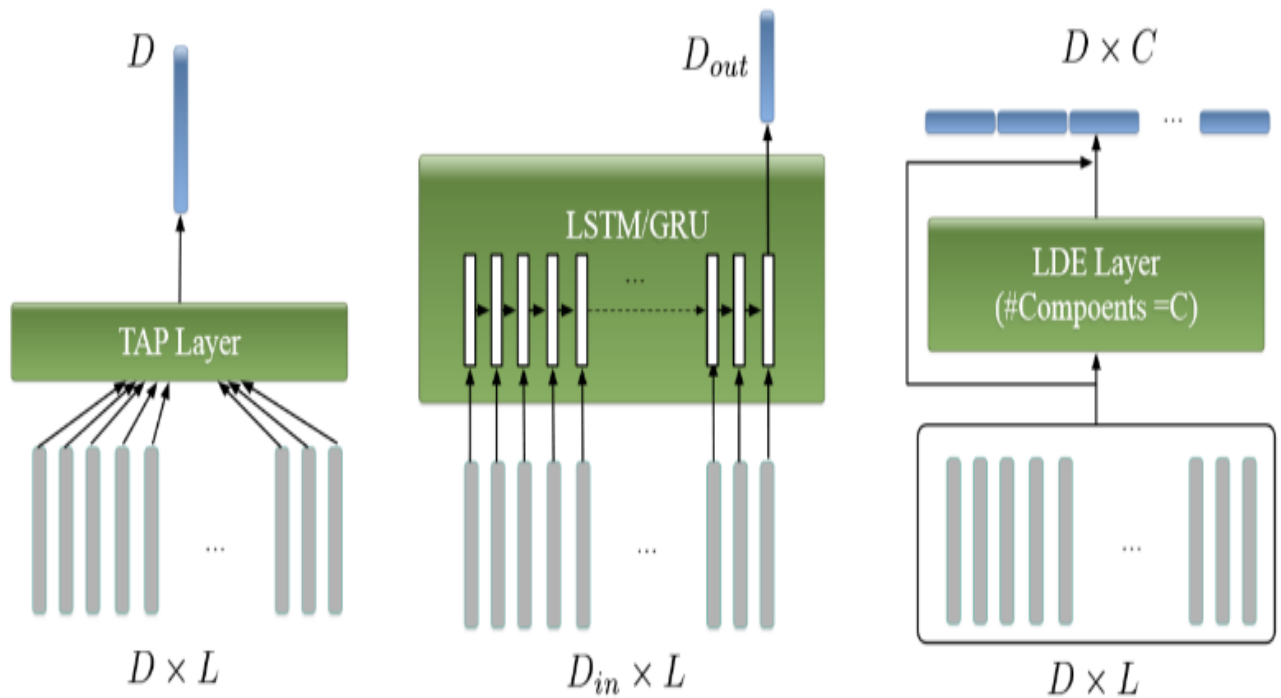


Fig. 5. Typical encoding layers. They all receive variable-length sequence, produce encoded utterance level vector with fixed dimension

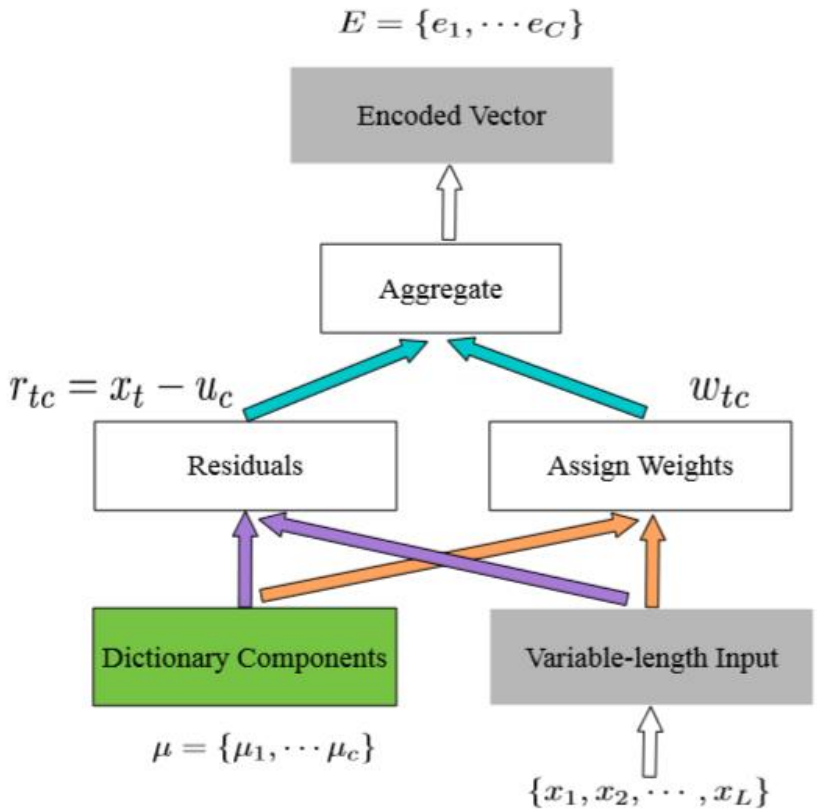


Figure 4: The forward diagram within the LDE layer

Table 1. Performance on the 2007 NIST LRE closed-set task

System ID	System Description	$C_{avg}(\%)/EER(\%)$		
		3s	10s	30s
1	GMM i-vector	20.46/17.71	8.29/7.00	3.02/2.27
2	DNN i-vector	14.64/12.04	6.20/3.74	2.601.29
3	DNN PPP Feature	**8.00/6.90**	**2.20/1.43**	**0.61/0.32**
4	DNN Tandem Feature	9.85/7.96	3.161.95	0.97/0.51
5	DNN Phonotactic[22]	18.59/12.79	6.28/4.21	1.34/0.79
6	RNN D&C[22]	22.67/15.57	9.45/6.81	3.28/3.25
7	LSTM-Attention[21]	-/14.72	-/-	-/-
8	**CNN-TAP**	9.98/11.28	3.24/5.76	1.73/3.96
9	**CNN-GRU**	11.31/10.74	5.49/6.40	-/-
10	**CNN-LSTM**	10.17/9.80	4.66/4.26	-/-
11	**CNN-LDE**	**8.25/7.75**	**2.61/2.31**	**1.13/0.96**

- 可扩展性？

- ？重新训练
- 层次架构
- 仅训练新类

- 分层语言架构

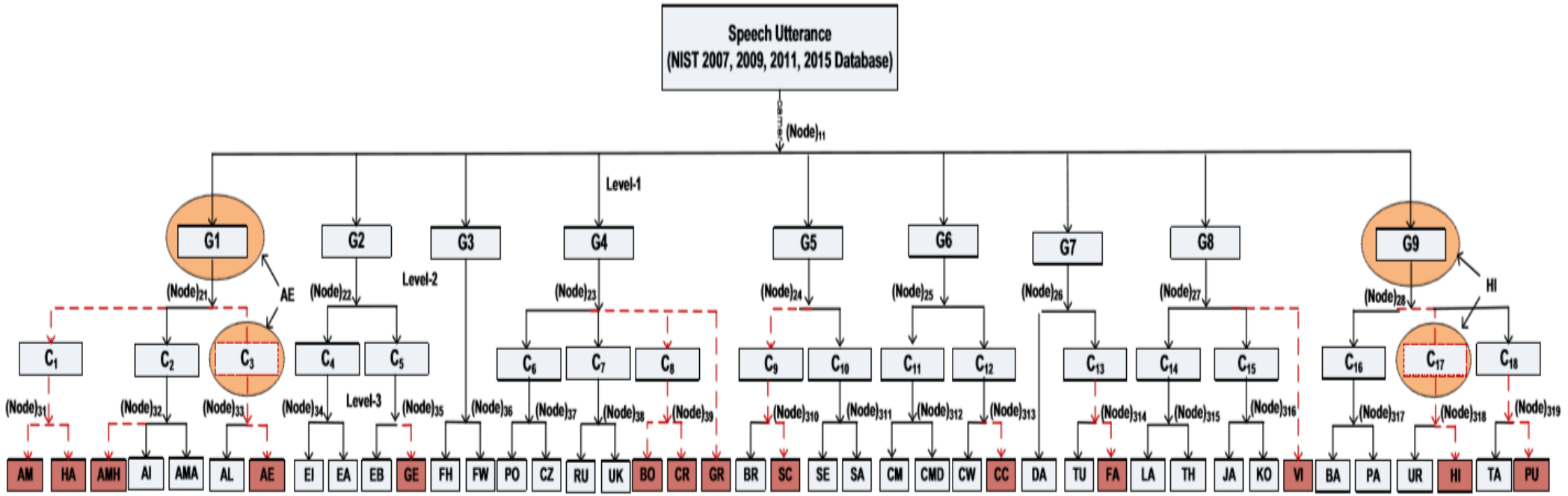


Figure 2: *Hierarchical Language Identification Framework on NIST 2007, 2009, 2011 and 2015 Database.*

2017, Investigating Scalability in Hierarchical Language Identification System

- 仅训练新类

Table 1: *Adding Classes Without Original Data (ACWOD) algorithm.*

Input:

- A network that predicts classes in A:

$$p(y = i|x) = \frac{\exp(w_i^\top h(x))}{\sum_{j \in A} \exp(w_j^\top h(x))}, \qquad i \in A.$$

- Training data $x_1, ..., x_n$ with corresponding labels $y_1, ..., y_n$ from a disjoint class-set B.
- No training data with labels from A are available.

Goal: Extend the network to predicts all classes in $A \cup B$:

$$\tilde{p}(y = i|x) = \frac{\exp(w_i^\top h(x))}{\sum_{j \in A \cup B} \exp(w_j^\top h(x))}, \qquad i \in A \cup B.$$

Algorithm: Fix the model parameters and find the parameters $\{w_i | i \in B\}$ that maximize the following concave objective function:

$$S = \sum_{t=1}^{n} \left((1-\epsilon) \log \tilde{p}(y_t|x_t) + \frac{\epsilon}{|A|} \sum_{i \in A} \log \tilde{p}(y = i|x_t) \right)$$

2018, Adding New Classes without Access to the Original Training Data with Applications to Language Identification

- 数据有限？

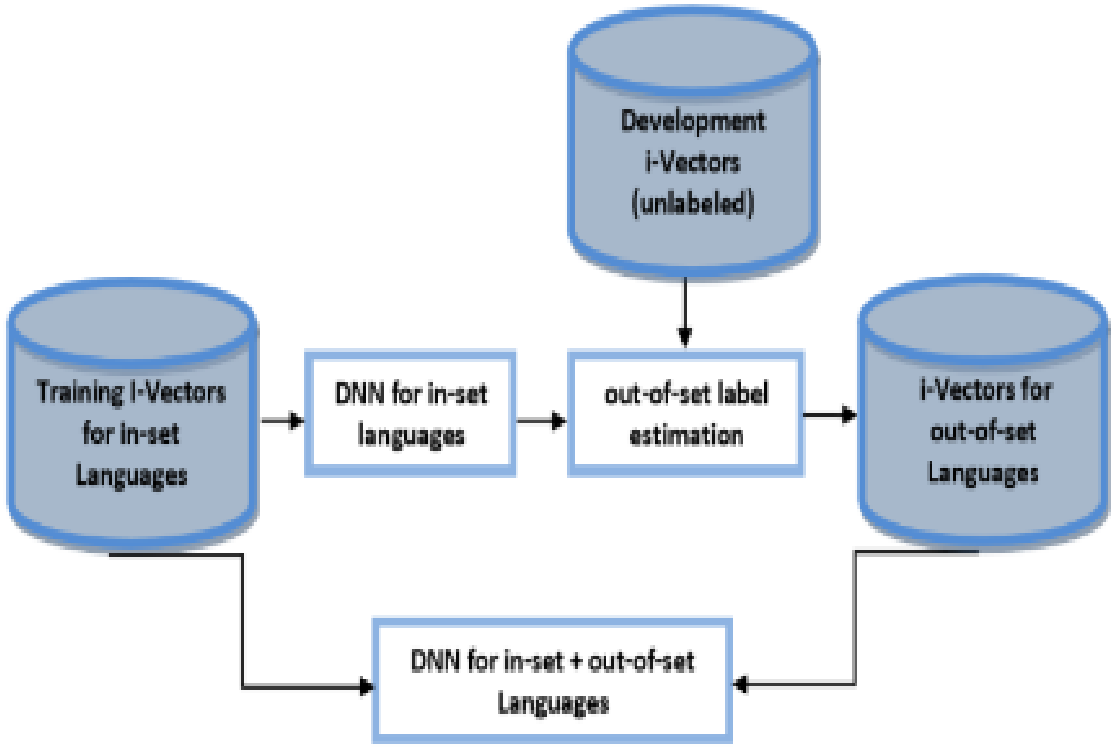


Fig. 1. *2-step DNN training for LID using i-Vectors.*

2016,iccasp, LANGUAGE RECOGNITION USING DEEP NEURAL NETWORKS WITH VERY LIMITED TRAINING DATA

情感识别概述

- 情感识别面临的挑战？

  - 1. 什么样的特征在分辨情感中最有用？
  - 2. 一段发音可以包含多种情感，不同情感的边界也难以界定，那么，哪个情绪占主导？
  - 3. 情绪可能有瞬间的变化，比如被炒鱿鱼，会悲伤很久，但这期间吃了顿大餐，吃的时候是很开心的，但人还处于伤心的状态中，那么该判定为悲伤还是开心呢？
  - 4. 情感？如何定义？比如喜极而泣 那哭声是开心还是伤心？

情感中的特征

- Continuous speech features 连续语音特征

  - pitch-related features
  - formants features
  - energy-related features
  - timing features
  - articulation features
  - 常用的有F0，Energy，Duration，Formants。另外在特征的提取中，除了使用特征还对特征进行一些转换，比如平均，最大最小等。

- Voice quality features

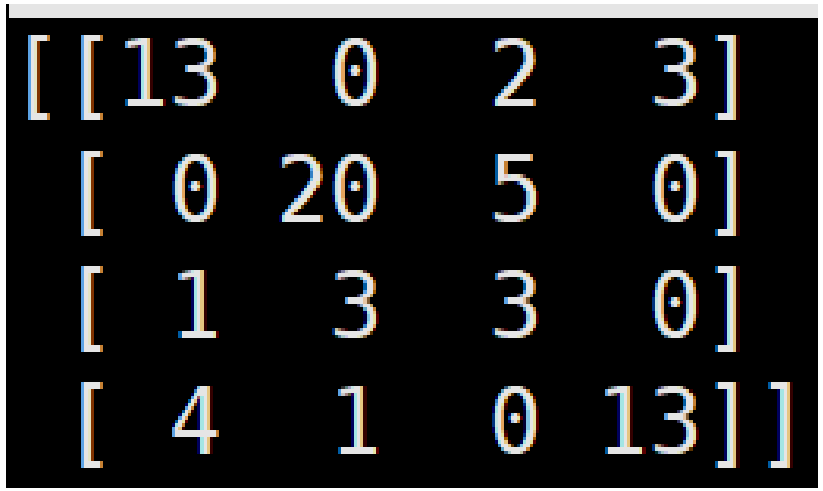
  - voice quality
  - harsh
  - tense
  - breathy

- Spectral-based speech features
    - LPC
    - MFCC
    - LFPC
- TEO-based features

Demo

数据：4种，Angry 、Happy、Neutral、Sad

SVM下混淆矩阵：



网络结构	ACC
SVM	0.72
Random Forest	0.56
MLP	0.78
LSTM	0.90
CNN	0.93

THANK YOU!