END-TO-END HIERARCHICAL LANGUAGE IDENTIFICATION SYSTEM

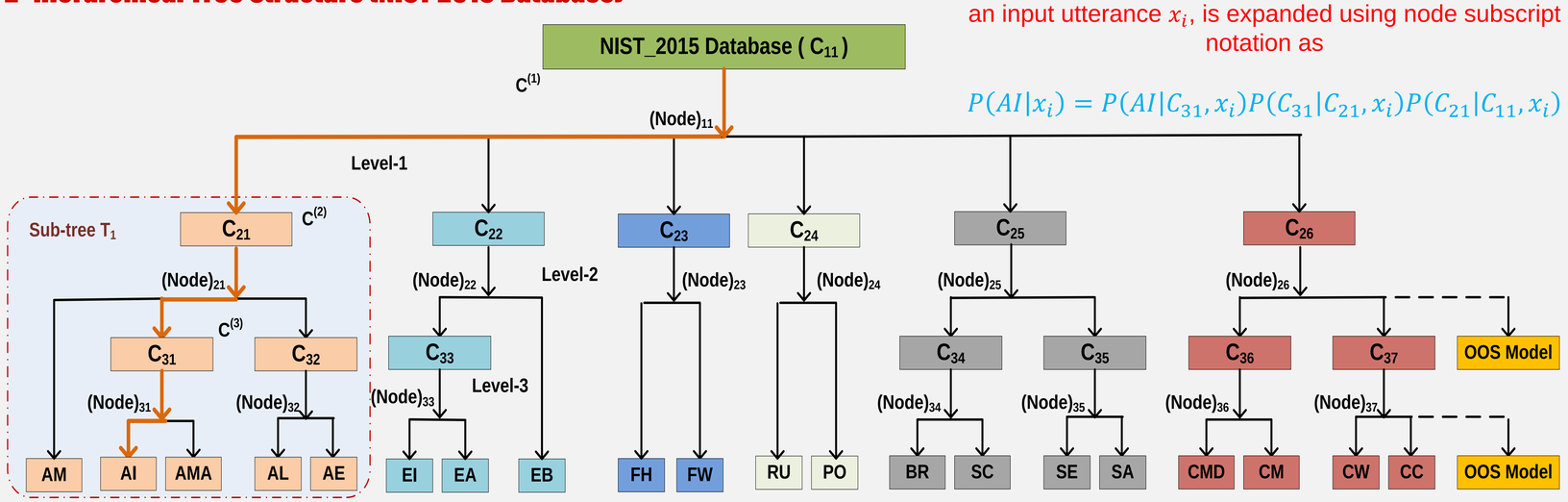
Saad Irtza[1,2]*,Vidhyasaharan Sethu*[1]*, Eliathamby Ambikairajah*[1,2]*,Haizhou Li*[3]
[1] The School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney NSW 2052, Australia
[2] ATP Laboratory, National ICT Australia (NICTA), Australia, Sydney NSW 2015, Australia
[3]National University of Singapore, Singapore

s.irtza@unsw.edu.au

1- Introduction

- ✓ The deep learning approaches have been successfully employed to develop single level end-to-end LID systems.
- ✓ State of art LID system:
  - ❖ Treats all languages equally (in-set or out of set)
  - ❖ May require data from additional languages that are not in the set of target languages in order to model OOS language model
- ✓ Hierarchical LID framework:
  - ❖ Divides the classification problem into simpler set of tasks
  - ❖ Allows target languages to be identified in final layer
  - ❖ Requires significant effort to choose best features and classifier at each node
- ✓ Contributions
  - ❖ Proposes an end-to-end HLID system training to jointly optimize the feature extraction and classification
  - ❖ Demonstrates its in-built ability to enables an OOS model, without using any additional OOS language training data

2- Hierarchical Tree Structure (NIST 2015 Database)



The posterior probability of AI (path highlighted red), given an input utterance x_i, is expanded using node subscript notation as

$$P(AI|x_i) = P(AI|C_{31}, x_i)P(C_{31}|C_{21}, x_i)P(C_{21}|C_{11}, x_i)$$

3- Optimizing Combined Prediction Loss

Approach-I

- ✓ Proposed to combine the prediction loss of each language group's specific network from all nodes in a path from root to leaf of the sub-tree

Approach-II: Proposed to combine the prediction layer of language group specific network

- ✓ In HLID system, language posterior of each target language is the chain product of the conditional probabilities of a target language or group, given its parent group, on the path from root to leaf node as:

$$P(\ell_t|x_i) = P\left(\ell_t\middle|C^{(N)}, x_i\right)\left(\prod_{n=2}^{N} P\left(C^{(n)}\middle|C^{(n-1)}, x_i\right)\right)$$

n denotes the number of level in sub-tree T_j
$C^{(n)}$ denotes the language group at n^{th} level

- ✓ The conditional probability of each target language/group is computed from the prediction layer of each language group's specific network as:

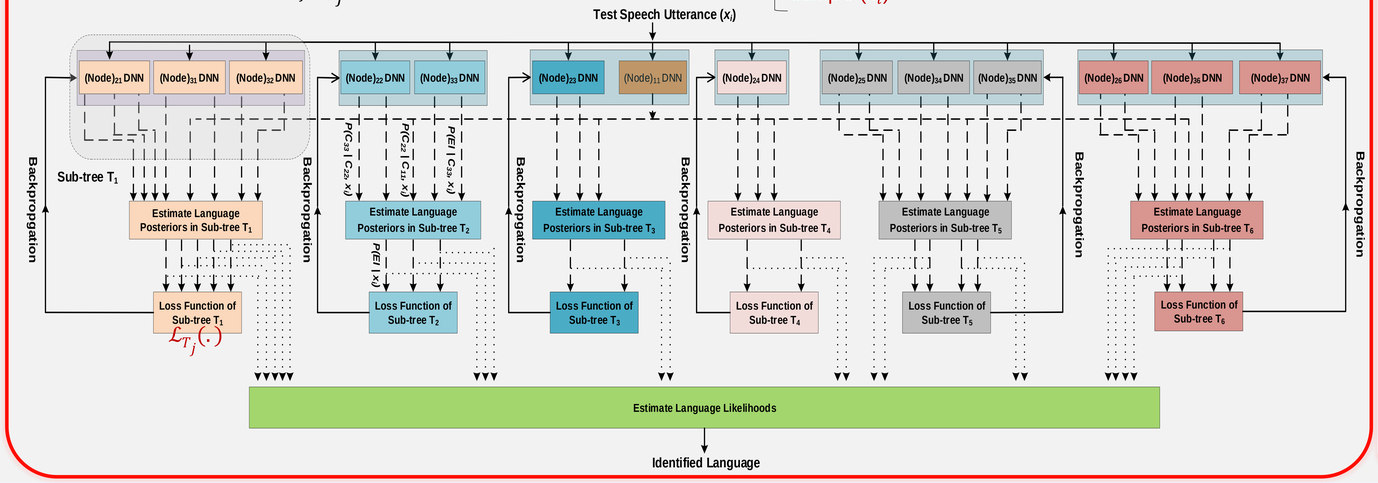
$$P\left(C^{(n)}\middle|C^{(n-1)}, x_i\right) = \left(G^{(n-1)}\left(G_f\left(x_i;\ \theta_f\right); \theta^{(n-1)}\right), x_i\right)$$

G_f is the feature extraction network
G and $\theta^{(n-1)}$ are the language group specific network and associated parameters

- ✓ Network is trained by optimizing each tree objective function computed as:

$$E_{T_j}(\theta_f, \theta_{T_j}) = \min_{\theta_f, \theta_{T_j}} \frac{1}{I}\sum_{i=1}^{I} \mathcal{L}^i_{T_j}(P(\boldsymbol{\ell}|x_i); \theta_f, \theta_{T_j})$$

Here, $\mathcal{L}^i_{T_j}(.)$ is the loss function of sub-tree T_j for the ith speech sample (x_i)



4- Experimental Setup

Features:

- ✓ Spectrogram of 128 frequency bins (30msec with 50% overlap)

End-to-end Network:

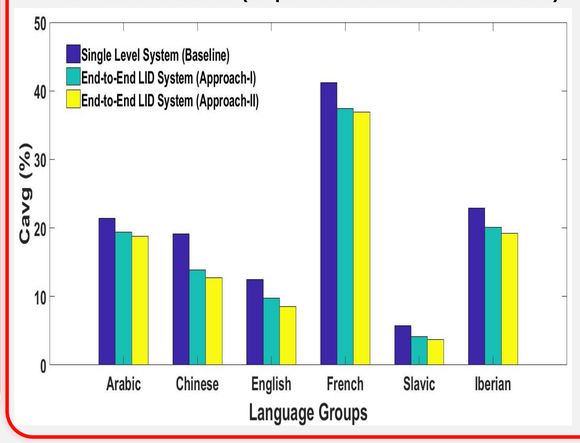
- ✓ Two CNN layers with filter size of 9x9 and 3x5
- ✓ LSTM layer of 256 memory blocks
- ✓ Four fully connected DNN layers of 100 dimensions each
- ✓ Activation function: Rectified linear unit
- ✓ Momentum optimizer with drop out regularization

Out of Set Languages Test Data:

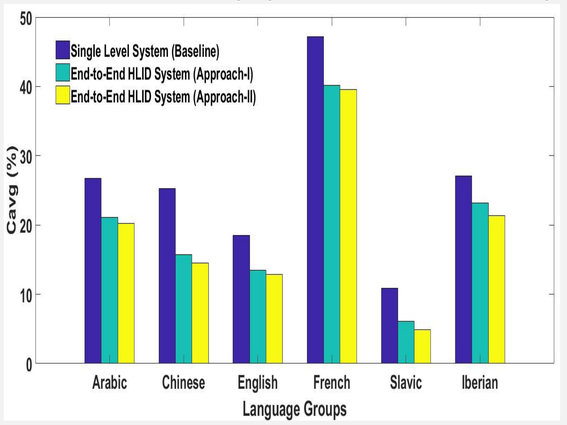
- ✓ 17 different languages from previous NIST LRE datasets (2007 and 2011)

5- Results

Closed Set Detection Results on NIST 2015 Database (as per NIST 2015 LRE Evaluation)



Open Set Detection Results on NIST 2015 Database (as per NIST 2015 LRE Evaluation)



6- Conclusion and Future Work

- ✓ Proposed two novel approaches to train end-to-end hierarchical structure for language identification
- ✓ The proposed hierarchical structure
  - ❖ Jointly optimize the nodes that are under same sub-tree in the hierarchical structure
  - ❖ Automate the feature extraction and classification process at each node
  - ❖ Develop the OOS language model without using any additional non-target languages data
- ✓ Future Work:
  - ❖ Automate the language clustering with hierarchical structure training