Non-native speech recognition is highly challenging, because the wide range of languages causes heavily accented speech, and the pronunciation closely depends on the first language (L1). In current automatic speech recognition (ASR) systems, the candidates’ L1 will have significant influence on its performance. The speaker adaptation then becomes an essential component of ASR systems, which is to take an initial, well-trained model set and use data from a new speaker, the adaptation data, to improve the performance on the new speaker.

Correspondingly, this project aims to examine adapting acoustic models and language models to better reflect the impact of L1 on the system, and improve the deep neural network (DNN) adaptation of ASR systems to non-native speakers.

The current ASR framework is a combination system of tandem and stacked hybrid system with joint decoding. My work focuses on improving the adaptation performance of the hybrid system on non-native speakers with unsupervised datasets.

**Dataset**

- Section C-E from BULATS
- 6 different first languages (L1s)
- Training data from Gujarati Indian speakers
- Small amount of crowd-sourced test data (approximately 25 hours)

**DNN Adaptation**

**Baseline:**
- Hybrid-SAT system with Hybrid-SI supervision
- BN features (39 dim) from DNN trained by AMI dataset
- PLP features

**Adaptation Methods:**

- Consider a layer of a network with 1000 x 1000 connections
  - weights: 1,000,000 parameters to adjust
  - activation functions: 2,000 functions (output and input)

- Take the example of a sigmoid activation function

\[
\phi(a_1, a_2, a_3) = \frac{a_3}{1 + \exp(-a_2 W_1 x + a_3)}
\]

- \(a_1\): scale of the input
- \(a_2\): scale of the output
- \(a_3\): offset on the activation
- train these (or subset) parameters to be speaker specific

**Learning Hidden Unit Contributions (LHUC)**

- Sigmoid function constraint scaling factor
- Total number of adaptation parameters ≤ Total number of hidden units

**Parameterised Sigmoid Activation Functions (P-Sigmoid)**

- Linear scaling factor
- Extra flexibility
- More easily jointly learned with other DNN parameters

**Speaker-aware Training (SaT)**

- Standard i-vector method

\[
t_l = \phi(W_l h_{l-1} + b_l)
\]

Where \(t_l = (t_l^{(s)} + b_l^{(s)})\) is the speaker representation (i-vector) and \(U_l\) is the speaker representation transformation weight matrix for layer \(l\).

- Factorised Feature Transforms

\[
t_l = \phi(W_l h_{l-1} + U_l D^{(s)} U_2 h_{l-1} + b_l)
\]

Where \(D^{(s)} = \text{diag}(v^{(s)})\) and \(U_1\) and \(U_2\) are weight matrices for SD transformation.

**Current Progress**

- Initial Estimation of the ASR performance
- Lower WER to European languages
- Higher WER to Asian languages

**Comparison between adaptation performance to different L1s**

- LHUC, P-sigmoid and CMLLR show 2.64% - 4.57% reduction in WER

**References**

