ALTA Project - Spoken Language Assessment and Learning

Improve Adaptation Performance of ASR to Non-Native Speakers

Supervisor: Prof. Mark Gales

Junjie Pan

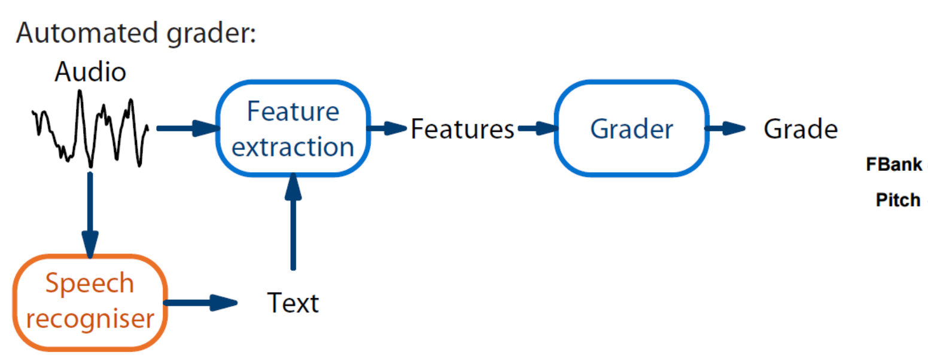
ALTA Institute / Department of Engineering, University of Cambridge

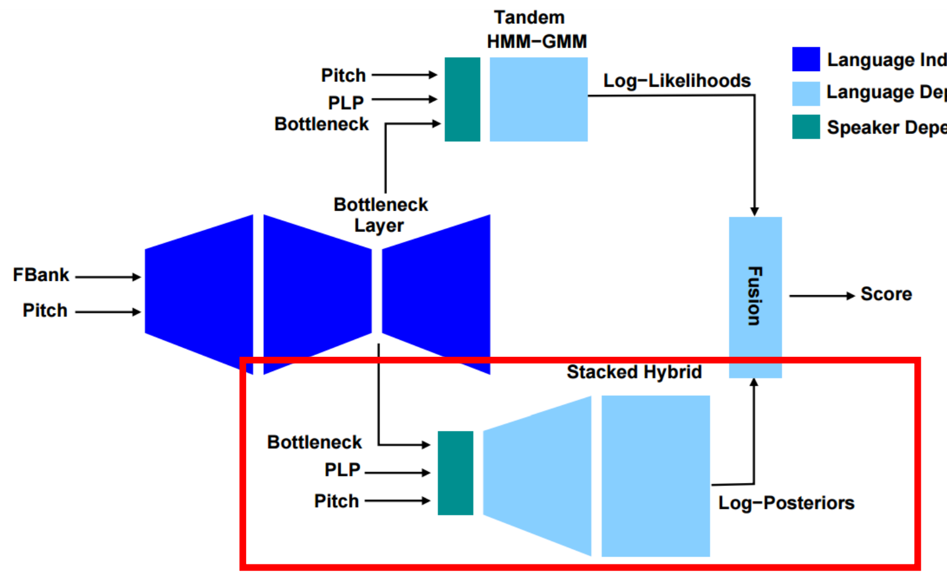
Overview

Non-native speech recognition is highly challenging, because the wide range of languages skills cause heavily accented speech, and the pronunciation closely depends on first language (L1). In current automatic speech recognition (ASR) system, 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 focus on improving the adaptation performance of the hybrid system on non-native speakers with unsupervised datasets.

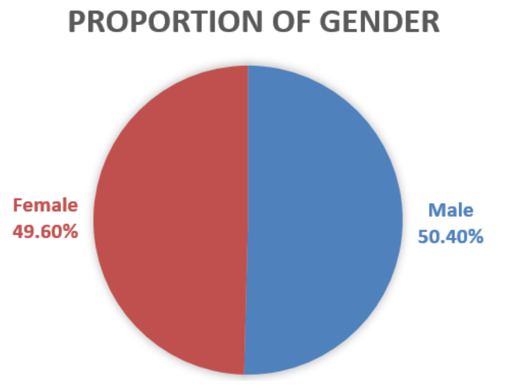


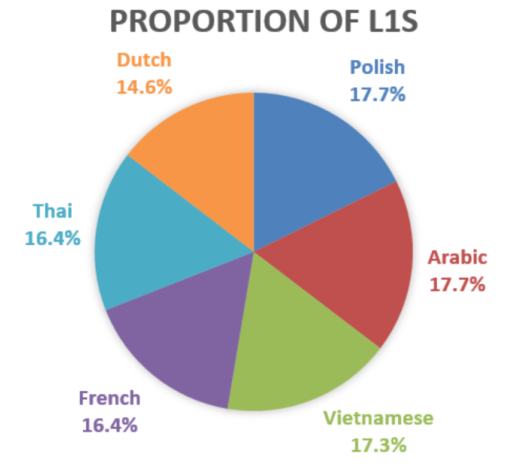


Dataset

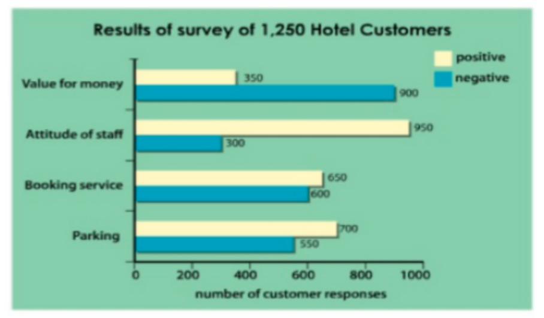
Dataset:

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





A. Introductory Questions: where you are from
B. Read Aloud: read specific sentences
C. Topic Discussion: discuss a company that you admire



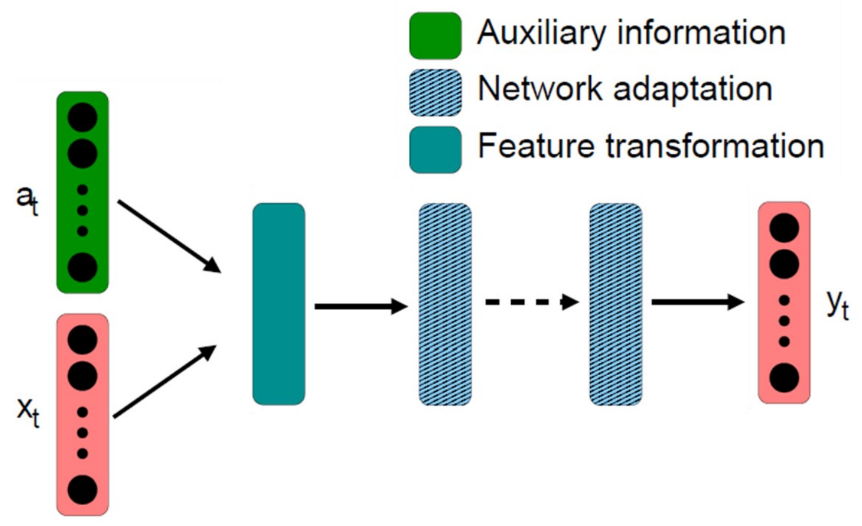
D. Interpret and Discuss Chart/Slide: example above
E. Answer Topic Questions: 5 questions about organising a meeting

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 × 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

$$\phi_i(\alpha_{\rm i},\alpha_{\rm o},\alpha_{\rm b}) = \frac{\alpha_{\rm o}}{1+\exp\left(\alpha_{\rm i}\mathbf{w}_i\boldsymbol{x}_t+\alpha_{\rm b}\right)}$$

- $\alpha_{\rm i}$: scaling of the input
- $\alpha_{\rm o}$: scaling of the output
- $\alpha_{\rm b}$: 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 parameter ≤ 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**

$$h^l = \phi(W_l h_{l-1} + b_l^s)$$

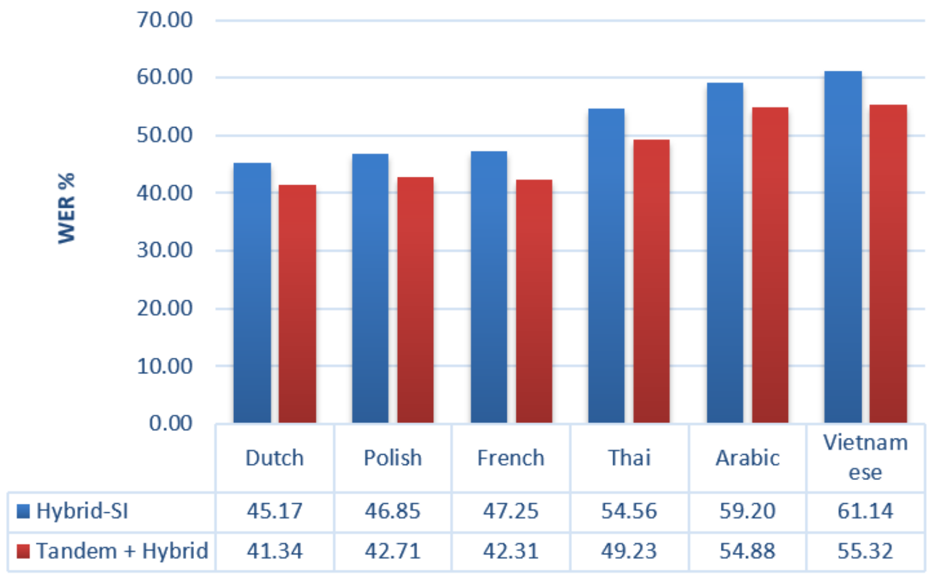
Where $b_l^s = U_l v^{(s)} + b_l$. $v^{(s)}$ is the speaker representation (i-vector) and U_l is the speaker representation transformation weight matrix for layer *l*.

- **Factorised Feature Transforms**

$$h^l = \phi(W_l h_{l-1} + U_1 D^{(s)} U_2 h_{l-1} + b_l^s)$$

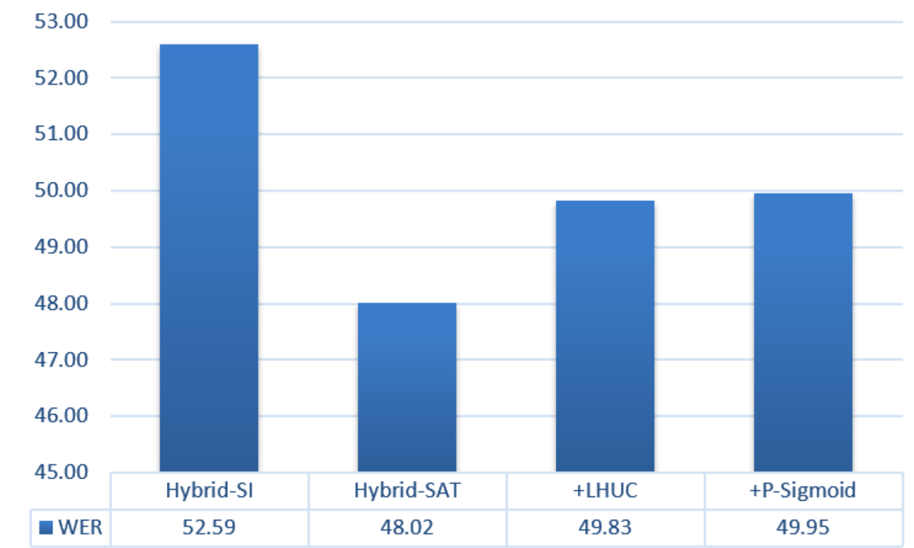
Where $\mathbf{D}^{(s)} = diag(\mathbf{v}^{(s)})$ and U_1 & U_2 are weight matrices for SD transformation.

Current Progress

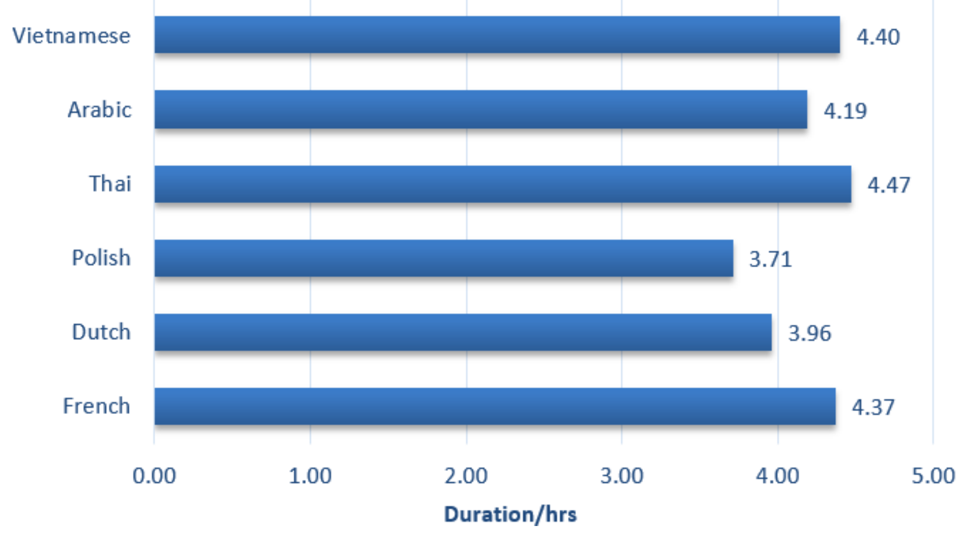


	Dutch	Polish	French	Thai	Arabic	Vietnamese
Hybrid-SI	45.17	46.85	47.25	54.56	59.20	61.14
Tandem + Hybrid	41.34	42.71	42.31	49.23	54.88	55.32

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



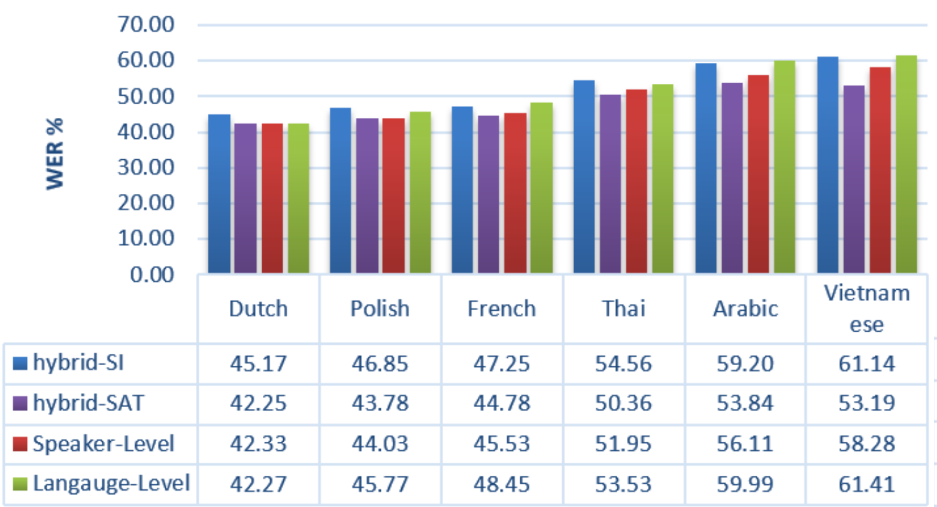
	Hybrid-SI	Hybrid-SAT	+LHUC	+P-Sigmoid
WER	52.59	48.02	49.83	49.95



- LHUC, P-sigmoid and CMLLR show 2. 64% - 4.57% reduction in WER

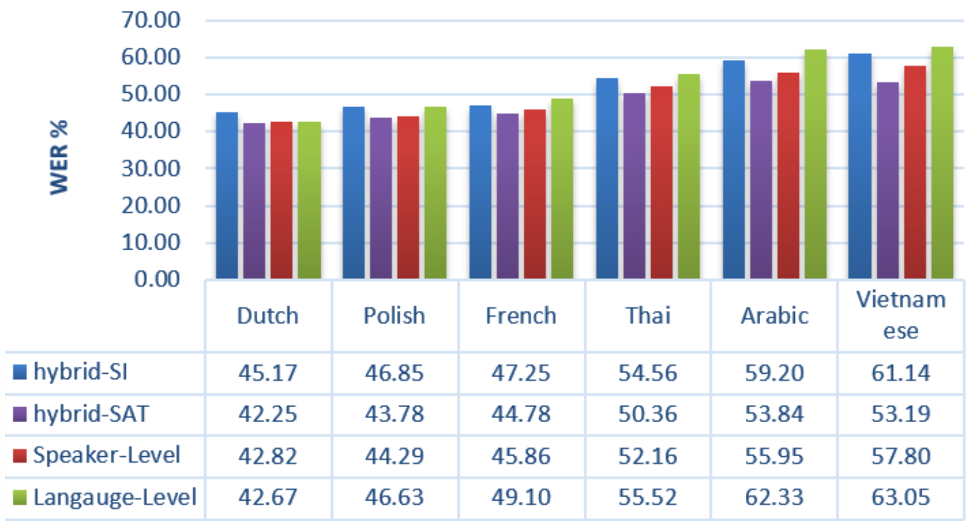
Comparison between adaptation performance to different L1s

LHUC Adaptation



	Dutch	Polish	French	Thai	Arabic	Vietnamese
hybrid-SI	45.17	46.85	47.25	54.56	59.20	61.14
hybrid-SAT	42.25	43.78	44.78	50.36	53.84	53.19
Speaker-Level	42.33	44.03	45.53	51.95	56.11	58.28
Langauge-Level	42.27	45.77	48.45	53.53	59.99	61.41

P-Sigmoid Adaptation



	Dutch	Polish	French	Thai	Arabic	Vietnamese
hybrid-SI	45.17	46.85	47.25	54.56	59.20	61.14
hybrid-SAT	42.25	43.78	44.78	50.36	53.84	53.19
Speaker-Level	42.82	44.29	45.86	52.16	55.95	57.80
Langauge-Level	42.67	46.63	49.10	55.52	62.33	63.05

- Unsupervised adaptation can show similar WER reduction to CMLLR for some L1s
- Speaker level and language level adaptation show consistent improvement
- Adaptation of other L1s shows no obvious improvement

References

[1] Samarakoon, L. and Sim, K. C. (2015). Learning factorized feature transforms for speaker normalization. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 145 - 152.

[2] Swietojanski, P. and Renals, S. (2014). Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models. In *Spoken Language Technology Workshop (SLT)*, 2014 IEEE, pages 171-176.

[3] Zhang, C. and Woodland, P. (2016). DNN speaker adaptation using parameterised sigmoid and ReLU hidden activation functions. *ICASSP*.