Deep learning for automatic pronunciation assessment of spontaneous non-native speech based on phone distances

K. Kyriakopoulos, K.M. Knill, M.J.F. Gales
{kk492,kate.knill,mjfg}@eng.cam.ac.uk
ALTA Institute / Department of Engineering, University of Cambridge

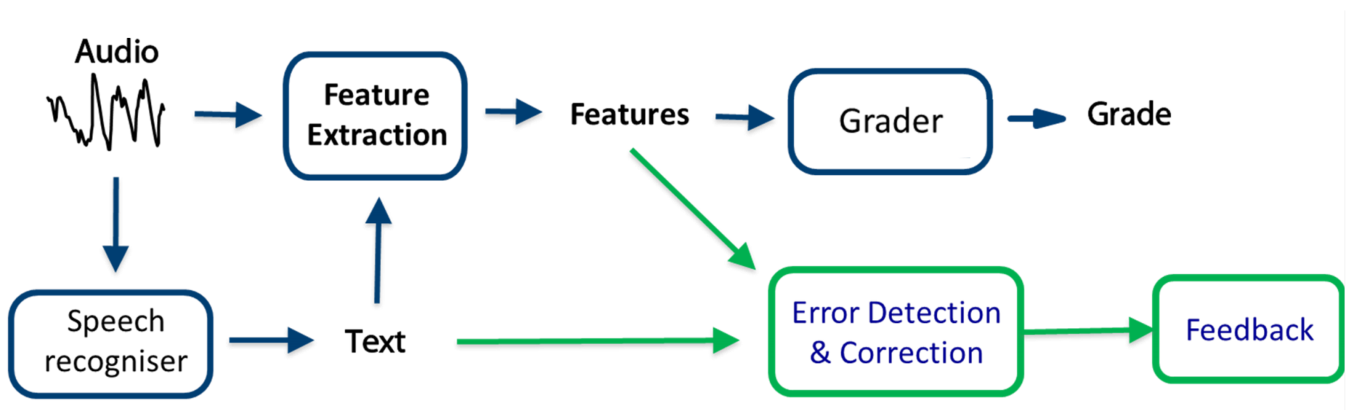
UNIVERSITY OF CAMBRIDGE

1. Introduction

Automatic assessment: How bad is speaker's pronunciation?
Feedback: How is speaker's pronunciation bad?
- Individual mispronunciations
- Overall problem phones

Motivation:
- Computer assisted language learning (CALL)
- Auto-marking of spoken examinations

- Features should be predictive of grade and interpretable
- Features projected to grade through feed-forward neural network
- Extraction and grading can be separate or combined

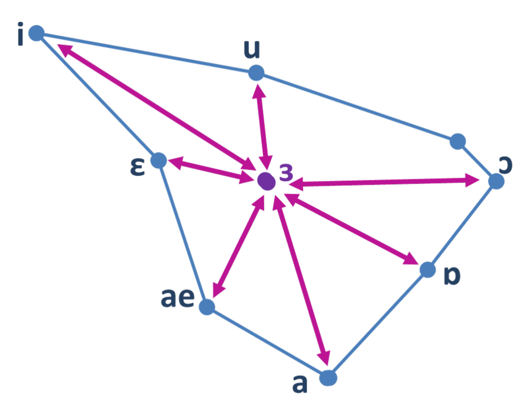


First steps:
1. Pass audio through ASR
2. Viterbi align to get label and boundaries for each phone instance i
3. Extract PLP feature vector $\mathbf{x}_t^{(i)}$ for each frame t of audio within each i

Constraints on features:
- Unstructured, spontaneous speech
  - High ASR work (and phone) error rate (c. 40%)
  - No native models with identical text
- Broad not narrow transcription
- Variability in speaker attributes

2. Model-based phone distance features (baseline)

- Each phone characterised relative to others
- Phone-to-phone distances act as features



- Train Gaussian model $\mathcal{N}(\mathbf{x}_t^{(i)}; \boldsymbol{\mu}_\phi, \Sigma_\phi)$ for all instances i of each phone ϕ
- Features are symmetric K-L divergences between pairs of models:

$$D_{\phi,\psi} = \frac{1}{2}\langle \mathcal{KL}\left(\mathcal{N}(\boldsymbol{\mu}_\phi, \Sigma_\phi)||\mathcal{N}(\boldsymbol{\mu}_\psi, \Sigma_\psi)\right) + \mathcal{KL}\left(\mathcal{N}(\boldsymbol{\mu}_\psi, \Sigma_\psi)||\mathcal{N}(\boldsymbol{\mu}_\phi, \Sigma_\phi)\right)\rangle$$

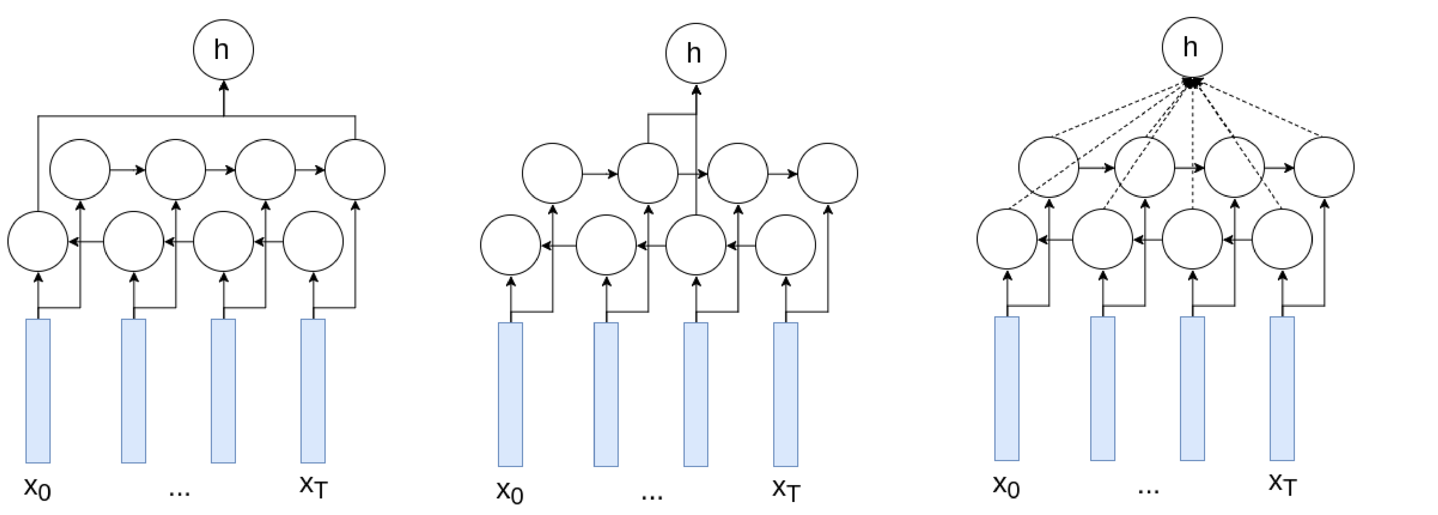
3. Deep representation of phone instances

- Bidirectional LSTM projects sequence of frame vectors $\mathbf{x}_t^{(i)}$:

$$\boldsymbol{h}_t^{(f,i)} = f(\boldsymbol{x}_t^{(i)}, \boldsymbol{h}_{t-1}^{(f,i)}, \boldsymbol{\lambda}^{(f,i)})$$

$$\boldsymbol{h}_t^{(b,i)} = f(\boldsymbol{x}_t^{(i)}, \boldsymbol{h}_{t+1}^{(b,i)}, \boldsymbol{\lambda}^{(b,i)})$$

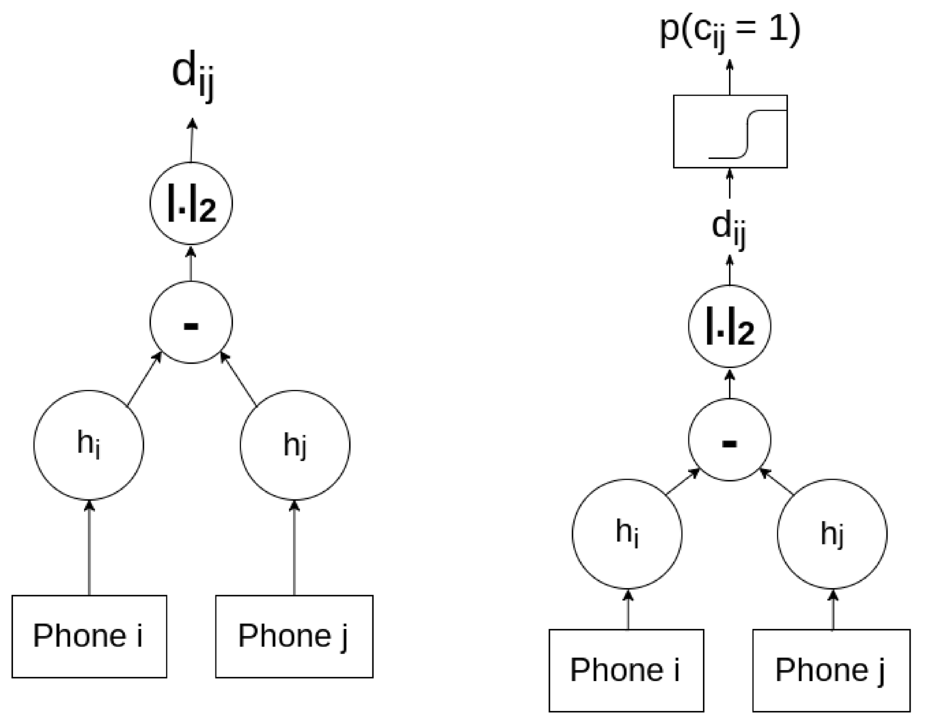
- Three different ways of getting instance vector $\boldsymbol{h}^{(i)}$ from $\boldsymbol{h}_t^{(f,i)}$ and $\boldsymbol{h}_t^{(b,i)}$:



Standard LSTM (left): Projects vector from last frame of each pass. Problematic as boundary frames not representative of phone.
Centre frame method (mid): Uses middle frame of each pass
Attention (right): Attention mechanism determines salience of each frame.

4. Siamese network distance metric

- Project instances i and j (of phones ϕ and ψ) from same speaker to vectors $\boldsymbol{h}^{(i)}$ and $\boldsymbol{h}^{(j)}$, then obtain distance $d_{i,j} = ||\boldsymbol{h}^{(i)} - \boldsymbol{h}^{(j)}||_2$



K-L training (left): Distance $d_{i,j}$ directly predicts model-based K-L distance $D(\phi, \psi)$ for that speaker
Binary training (right): Distance $d_{i,j}$ passed through sigmoid to predict:

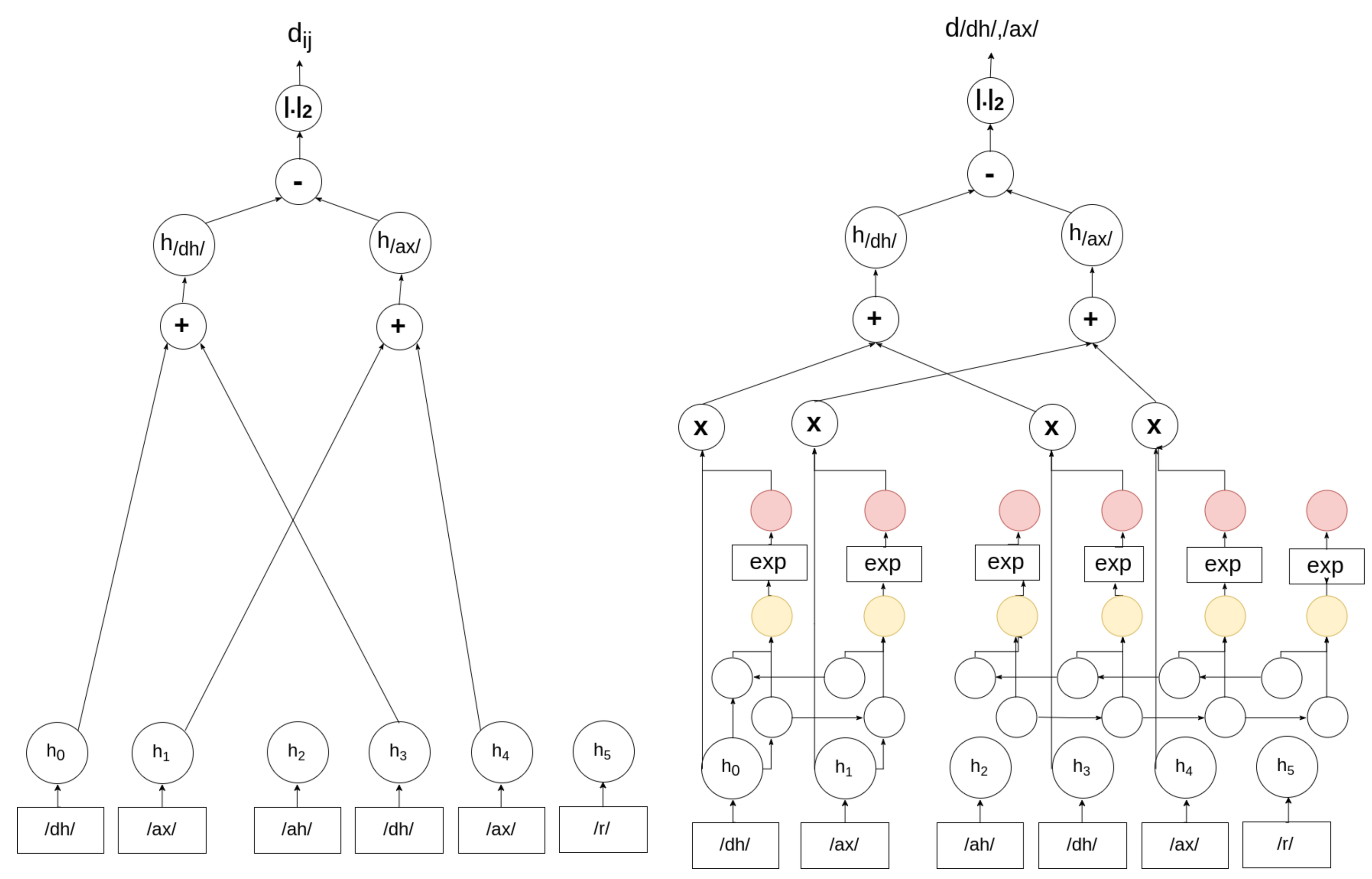
$$c_{ij} = \begin{cases} 1, & \phi = \psi \\ 0 & \phi \neq \psi \end{cases} \tag{1}$$

(i.e. whether the two instances are of the same phone)

- For both, train with random sample of instance pairs from each speaker

5. Predicting grade

- Bi-LSTM (trained as above) projects each instance to vector $\boldsymbol{h}^{(i)}$



Averaging (left): Obtain mean of vectors of all instances of each phones.
Attention (right): Use attention mechanism to obtain weighted sum of vectors

- In both cases, set of 1081 Euclidean distances between all pairs of phones projected to predict grade.
- Attention method allows feature extractors to be fine-tuned for task.
- Attention weights interpretable as importance to grade of phone instances:

$a/10^{-3}$	0.50	0.01	1.2	5.6	0.93	1.8
	/dh/	/ax/	/ah/	/dh/	/ax/	/r/

6. Experimental Results

Projection	Criterion	Combination	PCC
Standard	Binary	Average	0.698
Centre	Binary	Average	0.742
Attention	Binary	Average	0.762
Attention	K-L	Average	0.775
Attention	K-L	Attention	0.790
	Baseline		0.785

- Attention LSTM performance > centre frame LSTM > standard LSTM
- Initialising Siamese distance using model K-L divergences improves performance over binary classification training
- With attention mechanism and end-to-end training, deep method outperforms baseline