A hierarchical architecture for automatic pronunciation assessment of spontaneous non-native English speech based on phone distances

UNIVERSITY OF CAMBRIDGE

Konstantinos Kyriakopoulos, Kate K.M. Knill, Mark J.F. Gales *{kk492,kate.knill,mjfg}@eng.cam.ac.uk*

ALTA Institute / Department of Engineering, University of Cambridge

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 oral exams

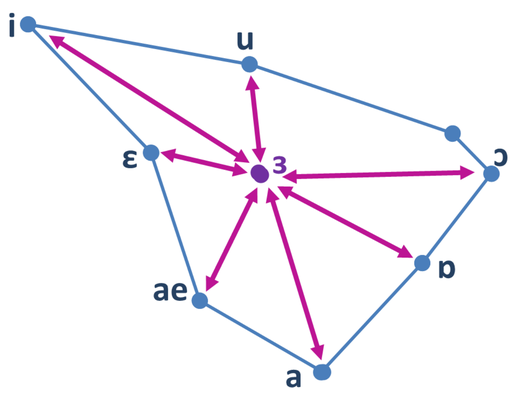
- Features should be predictive of grade and interpretable
- DNN used as grader (see process outline ➔)
- Extraction and grading initially separate, then combined

Constraints:

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

Feature Extraction

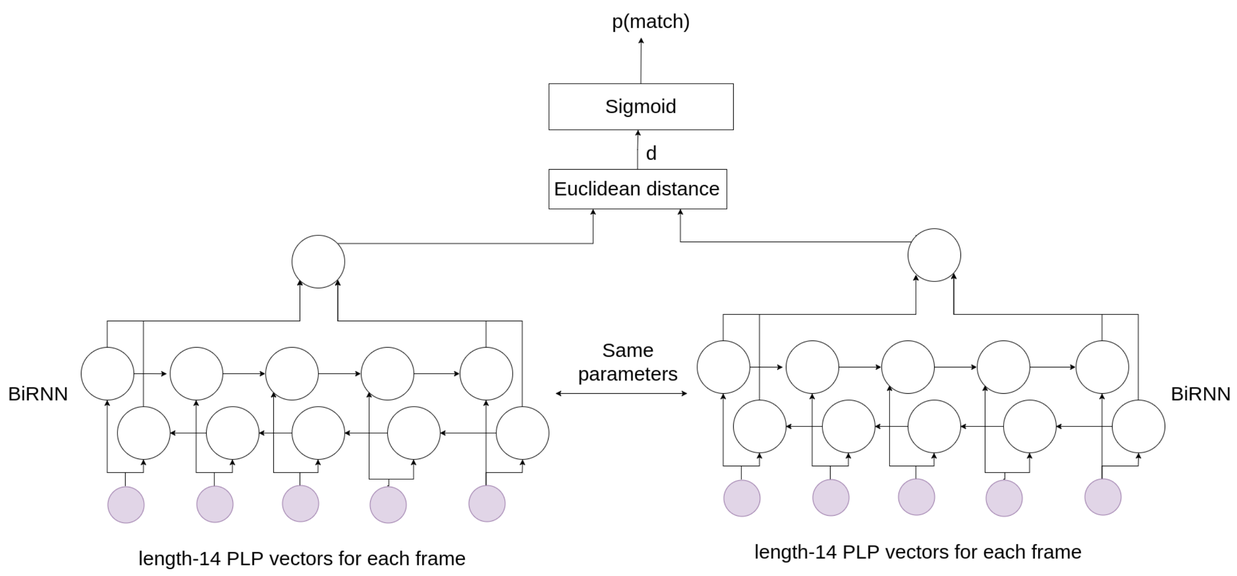
- Each phone characterised relative to others
- Phone-to-phone distances acts as features



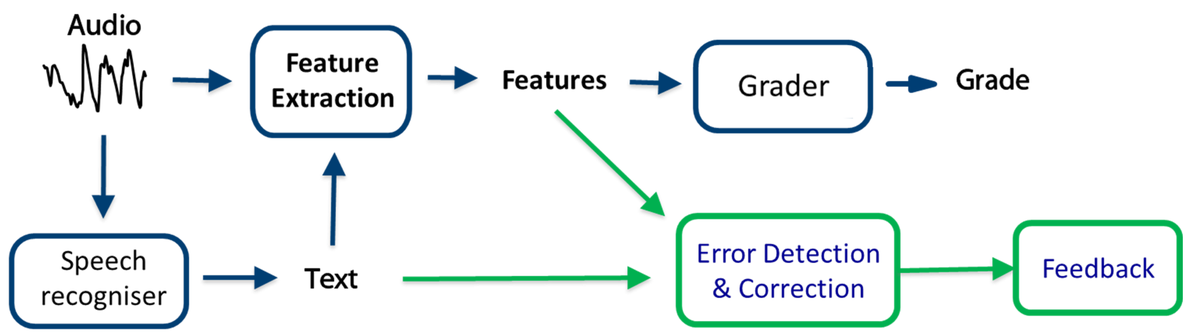
Baseline: Train Gaussian model for each phone and calculate K-L divergences between each pair

Siamese:

- Bi-LSTM projects frame sequence to phone instance feature
- Euclidean distance (d) between each pair of instances
- Two ways to train:
  - Classifier of instance pairs as same vs. different phone
  - Predictor of baseline phone distance features
- Average instance pairs for each phone pair to get features



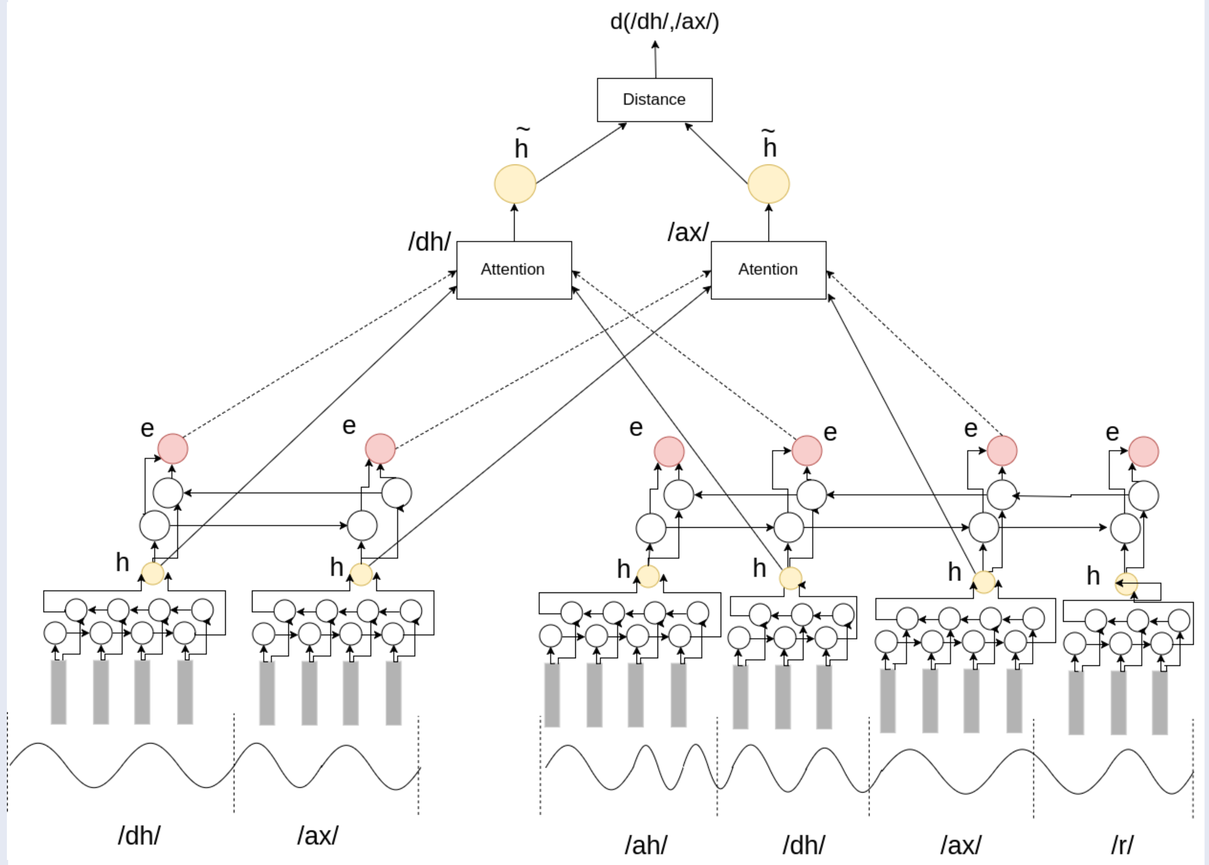
Process Outline



Hierarchical System

Combines feature extraction and grading

- Bi-LSTM projects frame sequence to phone instance feature (as in Siamese ↙)
- Weight-and-sum instance features to overall phone feature
  - Weights from instance features and phone identities (i.e. attention mechanism over instances)
- Siamese networks get distance between each phone pair
- Distances projected to score via DNN
- Full system backpropagated across



Attention weights identify salient phone instances:

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

Assessment Performance

- Evaluated as Pearson correlation between predicted and actual scores in held out evaluation set
- L1 of speakers is Gujarati

	Siamese extractor => DNN grader	Hierarchical System
PCC	0.644	0.680

- Hierarchical system outperforms separate systems as expected
- Need method to evaluate feedback