Automatic assessment of pronunciation and its dependent factors by exploring their interdependencies using DNN and LSTM

Aparna Srinivasan, **Chiranjeevi Yarra**, Prasanta Kumar Ghosh

SPIRE LAB
Electrical Engineering,
Indian Institute of Science (IISc), Bangalore, India

Overview

Introduction

- Overall quality of an utterance depends on quality of following factors[1]:
  1. Intelligibility
  2. Phoneme quality
  3. Phoneme mispronunciation
  4. Syllable stress quality
  5. Intonation quality
  6. Correctness of pause location
  7. Mother tongue influence (MTI).
- Exemplary sentence: "Please **begin** rubbing the **blue spot**"

Teacher Learner

[1] Ramanarayanan et al., "Human and automated scoring of fluency, pronunciation and intonation during human–machine spoken dialog interactions", 2017

Introduction

- Based on these factors, features have been proposed for the assessment.
- However, for an utterance, those have been obtained **heuristically** by applying statistics on the sub-segment level features.
- Typically, **utterance level averaging** have been considered.
- Classification based approaches have been used to assess the overall quality and quality of the factors **independently**.

Contributions

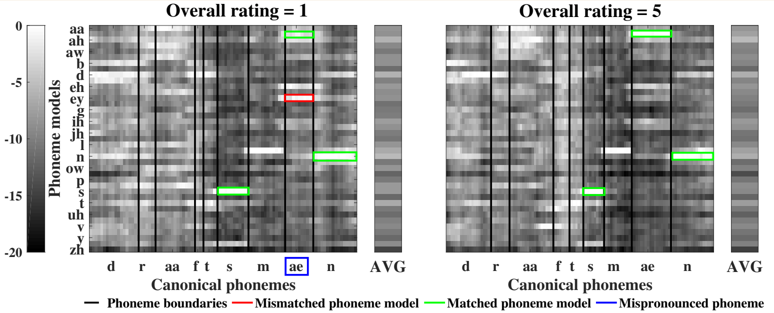
1. Feature computation to overcome the averaging based demerits.
2. Joint modelling to explore interdependencies among overall quality and quality of the factors.

Database

- Read English corpus collected from 16 Indian learners who were in spoken English training.
- Number of utterances: 12375 $\approx$ 800 per subject
- 800 unique utterances were also recorded from the expert.
- Overall quality ratings: Excellent (5: 20.3%), very good (4: 21.0%), good (3: 23.6%), moderate (2: 17.3%) and poor (1: 17.8%).

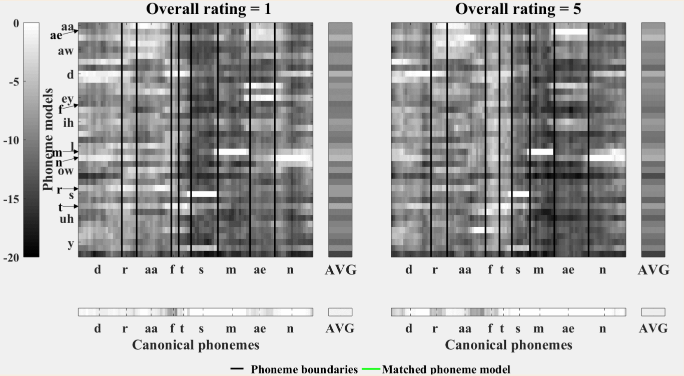
Yes (1)/No (0) questions for factors	1 (%)	0 (%)
Is utterance **intelligible**	88.5	11.5
Is **phoneme quality** good	68.7	31.3
Is **phoneme mispronunciation** exists	49.2	50.8
Is **syllable stress** proper	37.4	62.6
Is **intonation** proper	62.2	37.8
Is **pause locations** are proper	81.2	18.8
Is **MTI** present	57.6	42.4

Proposed feature computation



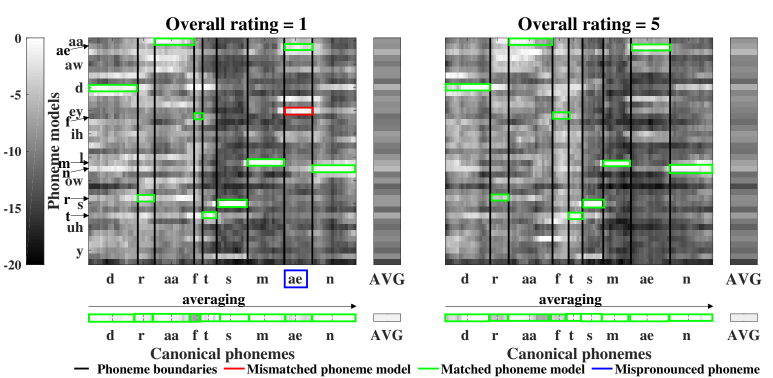
- Computed based on the frame level logarithm of posterior probability values from all phoneme models, referred to as log posteriors.
- Utterance level averaged features could be insufficient for better discrimination between the ratings.

Utterance level features (f_{utt})



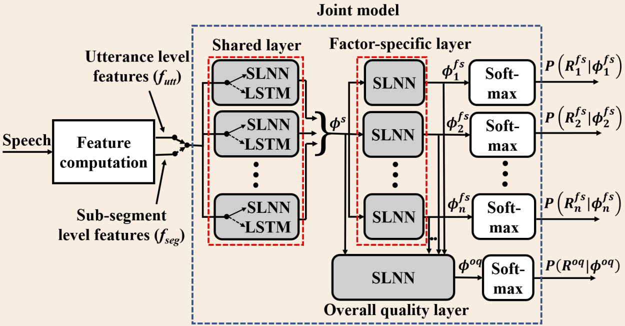
- Log posteriors from the matched phoneme model could be indicative of mispronunciation.
- Construct a one-dimensional vector consisting of the log posteriors from the matched phoneme models.

Sub-segment level features (f_{seg})



- Average performed over fewer frames in the sub-segments could discriminate the ratings better.
- f_{seg} are modelled in a data driven manner using LSTMs to overcome errors due to heuristic based averaging.

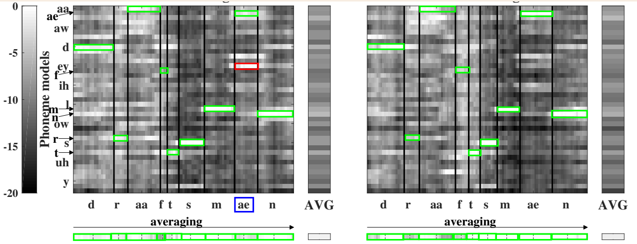
Joint model architecture



- Shared layer is believed to explore the interdependencies by learning common representations in conjunction with factor-specific and over quality layer.
- It uses single layer neural network (SLNN) for f_{utt} and LSTM for f_{seg}.
- The factor-specific and overall quality layer learn representations specific to each factor and overall quality separately.

Experimental setup

- Number of phoneme models: 39
- Baseline features: 78-dimensional paired log posteriors by concatenating the utterance level averaged log posteriors of learner and teacher.
- f_{utt} and f_{seg} dimensions are 80 and $n{\times}80$ respectively, where n is number of words.

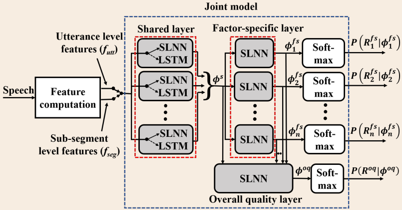


Experimental setup

- Number of phoneme models: 39
- Baseline features[2]: 78-dimensional paired log posteriors by concatenating the utterance level averaged log posteriors of learner and expert.
- f_{utt} and f_{seg} dimensions are 80 and $n \times 80$ respectively, where n is number of words.
- Five-class classification accuracy is used as the objective measure.
- 10-fold cross validation: 8 folds for train, 1 for validation and 1 for test.

[2] Xiao, Soong, and Hu, "Paired Phone-Posteriors Approach to ESL Pronunciation Quality Assessment", 2018

Experimental setup



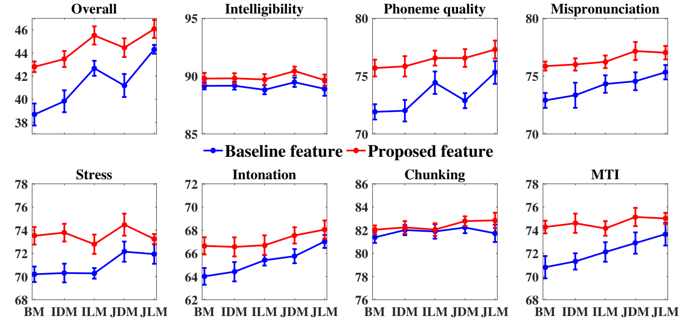
- JDM: joint model when f_{utt} is used.
- JLM: joint model when f_{seg} is used.

	JDM	JLM
Shared	6 layers 32 units	6 layers 128 units
Factor-specific/ Overall quality	32 units	

- Baseline model (BM)[3]: DNN with two hidden layers and 16 units each.
- IDM: DNN with two hidden layers and 32 units each.
- ILM: LSTM with 128 units and a SLNN with 32 units each.

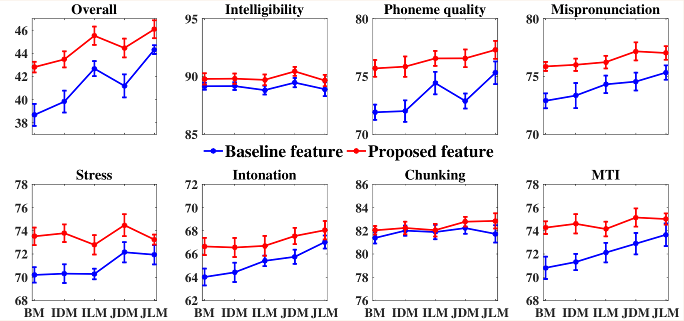
[3]Xiao, Soong, and Hu, "Paired Phone-Posteriors Approach to ESL Pronunciation Quality Assessment", 2018

Classification accuracy on test sets



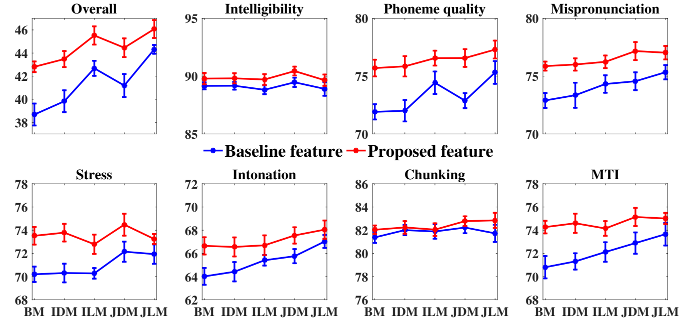
- Accuracies with the proposed features are higher than those with the baseline.
- Relative improvements with JLM and JDM in overall quality with respect to BM are found to be 19.13% and 14.93% respectively.

JDM vs IDM and JLM vs ILM



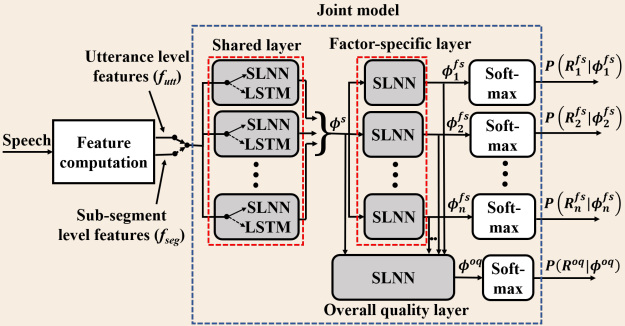
- Accuracies with JDM and JLM are found to be 2.25% and 1.23% (relative) higher in overall quality.
- Similar observations are consistent across all the factors.
- Joint models perform better than the independent models.

ILM vs IDM and JLM vs JDM



- Accuracies with ILM and JLM are found to be 4.69% and 3.64% (relative) higher in overall quality.
- f_{seg} is better than f_{utt}
- Lower performance in the factors intelligibility, stress and MTI could be avoided by considering phonemes or syllables as sub-segments.

Analysis on interdependencies



- Analysed the effect of both representations $\{\phi^s, \phi^{fs}\}$ on the overall quality.
- Compute the difference between the average accuracies with $\{\phi^s, \phi^{fs}\}$ and that with either ϕ^s or ϕ^{fs} separately for JDM and JLM.

Analysis on interdependencies

Table: Difference between the average accuracies obtained with $\{\phi^s,\phi^{fs}\}$ and those obtained with either ϕ^s or ϕ^{fs}. The negative entries are indicated in red.

	JDM		JLM	
	Only ϕ^{fs}	Only ϕ^s	Only ϕ^{fs}	Only ϕ^s
Intelligibility	0.3	0.3	0.09	0.21
Phoneme quality	0.27	0.33	0.03	-0.11
Mispronunciation	0.52	0.44	0.29	0.22
Stress	-0.01	0.32	-0.04	-0.11
Intonation	0.79	1.18	0.13	0.31
Pause locations	0.08	0.17	0.17	-0.01
MTI	-0.1	-0.04	0.19	-0.39
Overall quality	0.81	0.95	0.5	0.78

- The differences are positive in all cases of overall quality.
- The differences are positive in most of the cases for the factors.
- This benefit of joint training could be due to the interdependencies between the factors and overall quality.

Analysis on confusions among the ratings

Table: Confusions among the ratings in overall quality computed from a) BM with baseline feature (BM with baseline), b) JDM with f_{utt} and c) JLM with f_{seg}.

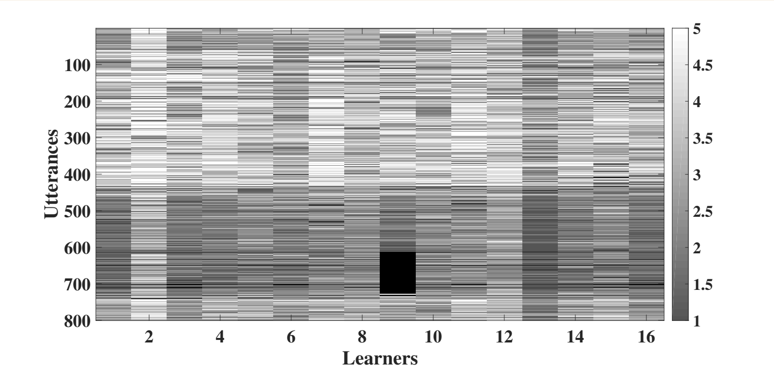
	(a) BM with baseline					(b) JDM with f_{utt}					(c) JLM with f_{seg}				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	38.0	30.6	16.3	4.3	10.8	54.0	28.7	11.7	2.3	3.3	57.9	25.4	9.2	3.2	4.3
2	24.0	39.8	26.2	3.5	6.5	22.7	46.4	23.6	3.7	3.6	20.4	48.3	21.4	5.2	4.7
3	12.6	22.5	38.2	9.4	17.3	9.8	24.8	39.9	11.4	14.1	9.5	22.5	35.9	16.3	15.8
4	8.3	8.5	29.2	14.7	39.3	4.2	8.1	31.6	20.5	35.6	5.2	8.2	23.2	25.7	37.7
5	4.7	2.4	19.0	11.8	62.1	2.1	2.1	17.7	17.4	60.7	2.9	2.9	11.1	19.7	63.4

- Shows the confusions in percentage averaged across 10 folds.
- Row $\longrightarrow$ true ratings; column $\longrightarrow$ predicted ratings.
- Red colored entries indicate where JDM and JLM have values lower in the diagonal and higher in the off-diagonal than the respective values from BM with baseline feature.
- No bias in predicting the ratings with the proposed approach.

Conclusion and Future work

- We predict the ratings for overall quality and its influencing factors by exploring interdependencies among those with joint models.
- In contrast to heuristically computed utterance level averaged features, we consider f_{seg} and model it using LSTMs.
- Experiments on the data collected from Indian learners reveal that the proposed joint approach performs better than the baseline scheme.
- Further investigations are required to identify better sub-segment level features for improving quality of all factors and overall quality.
- Better modeling strategies when the length of sub-segment level features from expert and learner are not identical.

THANK YOU


Utterances
100
200
300
400
500
600
700
800
2
4
6
8
10
12
14
16
Learners
5
4.5
4
3.5
3
2.5
2
1.5
1