# A MODEL FOR ASSESSOR BIAS IN AUTOMATIC PRONUNCIATION ASSESSMENT

Jose Antonio Lopez Saenz and Thomas Hain
{jalopezsaenz1, t.hain}@sheffield.ac.uk

Speech and Hearing Research Group (SPandH)

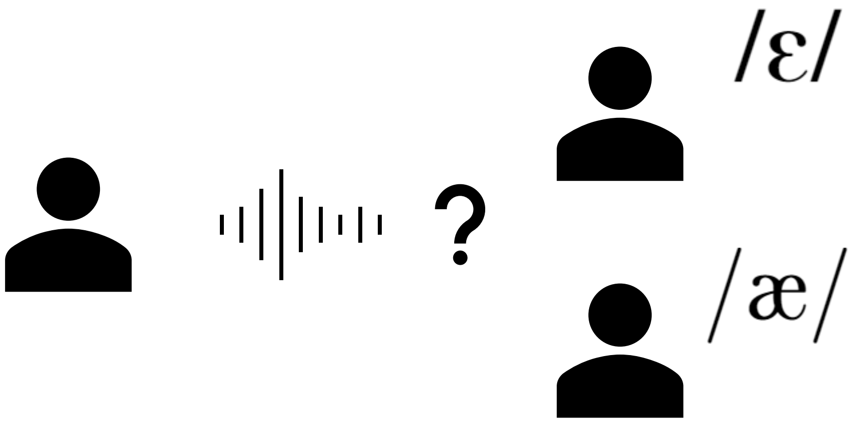
icassp 2022
Singapore, China, Online

# Outline

- Bias in Pronunciation Assessment
- A Model for the Assessor Bias
- Mispronunciation Detection
- Attention-Based Segmental Incorrectness Model
- Experiments:
  - Pronunciation Error Detection
  - Sensitivity to Assessor Tag
  - Similarity Between A and MaxVoting
- Analysis of Normalised Attention Curves
- Conclusion

# Bias in Pronunciation Assessment

- In pronunciation assessment (PA) an assessor declares the proficiency of a speaker using a pronunciation reference.
- The variations in second language (L2) speech are likely to cause a bias in the assessor towards the speaker [1].
- The bias in PA is a matter of inter-rater reliability attesting the lack of ground truth.



# A Model for the Assessor Bias

$$A_{\eta}\left(O^{(w)}\right) = A\left(O^{(w)}\right) + b_{\eta}\left(O^{(w)}\right)$$

$$A\left(O^{(w)}\right) = \sum_{\eta \in H} \left[A_{\eta}\left(O^{(w)}\right) + b_{\eta}\left(O^{(w)}\right)\right]$$

Where:
$O^{(w)}$ : Speech segment related to prompt w.
η : A pronunciation assessor in set $H$.
$A_{\eta}\left(O^{(w)}\right)$ : The pronunciation scoring function used by assessor η.
$A\left(O^{(w)}\right)$ : The assessor independent scoring function.
$b_{\eta}\left(O^{(w)}\right)$ : The η specific bias function.

# Mispronunciation Detection

$$P\left(Error|0^{(w)}\right) = 1 - P(\boldsymbol{l} = 1|r, O^{(w)})$$

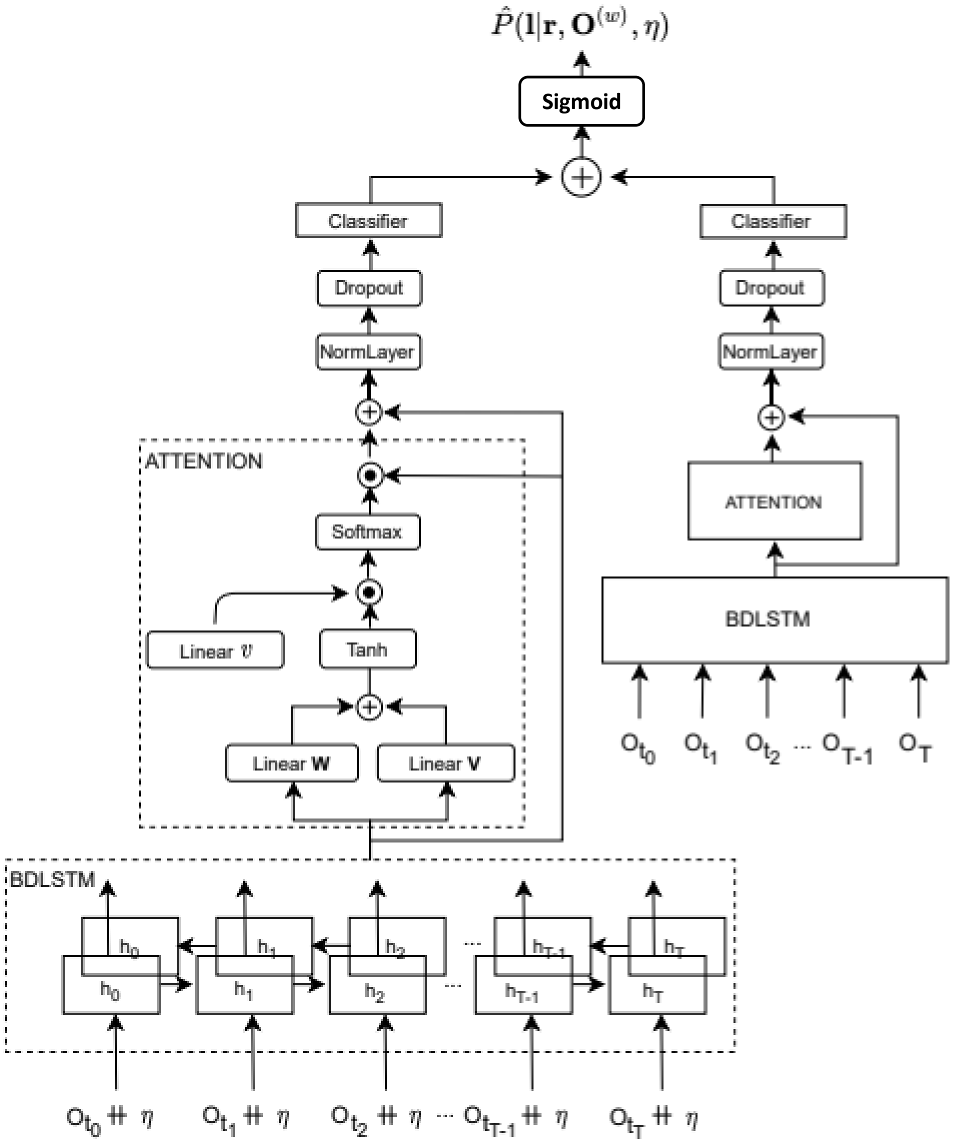
$$P\left(\boldsymbol{l} = 1|r, O^{(w)}\right) \cong \prod_i P\left(l_i = 1|r_i, 0^{(w)}\right)$$

Where:
$\boldsymbol{r} = \{r_i; i = 1, \ldots, R\}$ : a phoneme sequence assumed canonical.
$\boldsymbol{l} = \{l_i; i = 1, \ldots, R\}$ : a binary correctness indicator where $l_i = 1$ if $r_i$ is marked as correct.

# Attention-Based Segmental Incorrectness Model (ASIM)

- The model for the assessor bias is implemented on a dual model for detecting mispronounced segments [2].

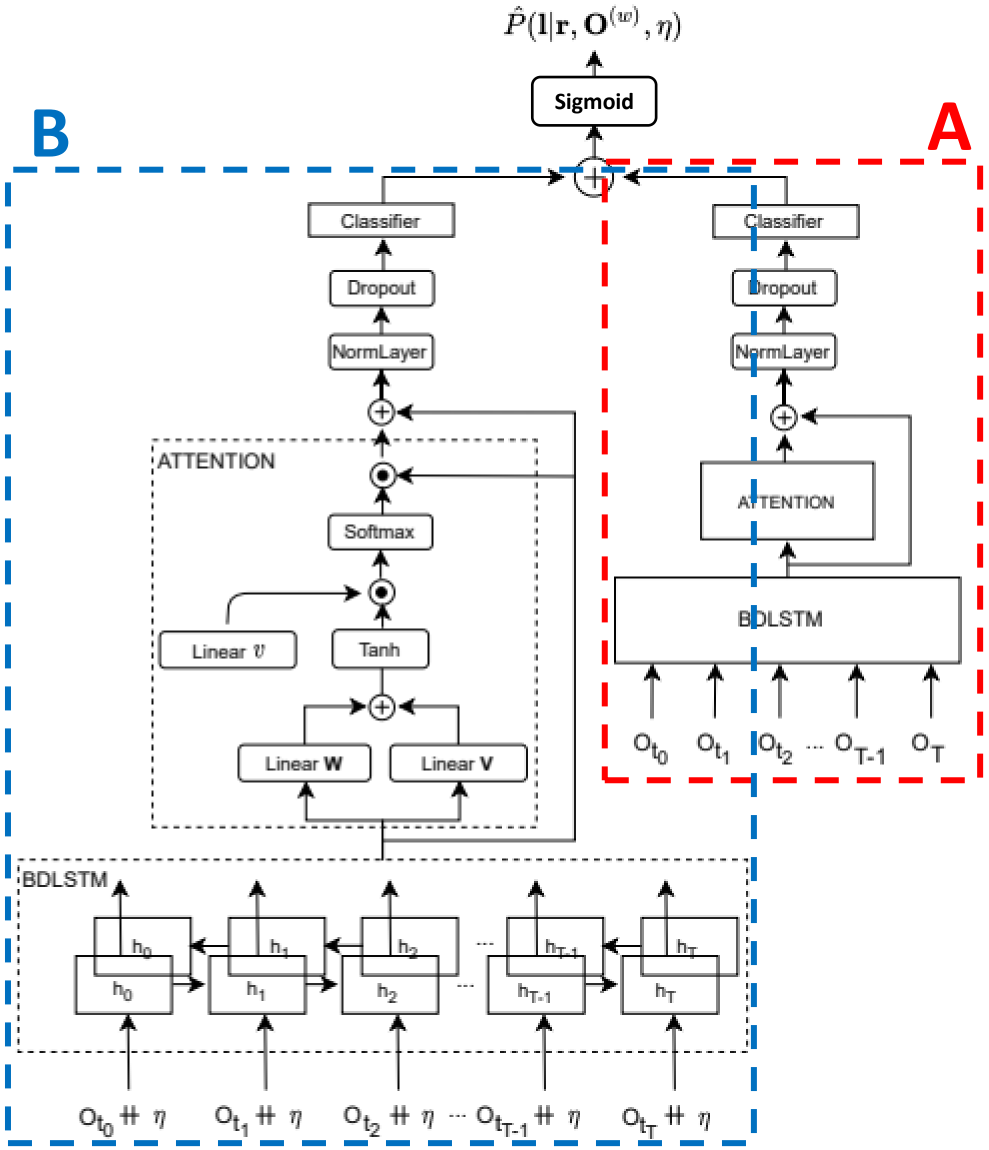
- Each branch estimates $\hat{P}\left(l|r,O^{(w)}\right)$ via:
  - Sequence encoding
    - BDLSTM + Additive self-attention [3]
  - Multilabel Classification
    - Deep feedforward network



# Attention-Based Segmental Incorrectness Model (ASIM) (2)

Each branch estimates $\hat{P}\left(l|r,O^{(w)}\right)$ to obtain:

$$\hat{P}\left(l|r,O^{(w)},\eta\right) = \hat{P}\left(l_A|r,O^{(w)}\right) + \hat{P}\left(l_b|r,O^{(w)},\eta\right)$$



# Experiment: Pronunciation Error Detection

- DATA: INA set from the ITSLANG Corpus of L2 prompted speech from ITSLANG BV [4].
  - 193 words and sentences
  - 230 speakers (early teens)
  - 6 hours annotated for mispronunciation at phoneme level by 3 professionals $a_1$, $a_2$ & $a_3$ (agreement shown below).
  - 85% for train and 15% for test.
  - No speaker overlap.
  - Balanced across sex, age and proficiency levels.

| vs. | | | % | $\kappa$ |
|---|---|---|---|---|
| $a1$ | | $a2$ | 0.871 | 0.349 |
| $a2$ | | $a3$ | 0.770 | 0.254 |
| $a3$ | | $a1$ | 0.808 | 0.446 |
| $a1$ | $a2$ | $a3$ | 0.725 | 0.331 |

# Experiment: Pronunciation Error Detection (2)

- The sequence $\boldsymbol{r} = \{r_i; i = 1, \ldots, R\}$ comes from forced-alignment.
  - DNN-HMM acoustic model trained on WSJCAM0 + 46hrs ITSL ∉ INA

- Segments $O^{(w)}$:
  - Sliding window of length 0.5s with 0.05s stride
  - Segments contained a $\mu = 3.46$ and $\sigma = 1.54$ annotated phonemes.
  - Only phonemes contained within 2 frames in each $O^{(w)}$ where considered for $\boldsymbol{r}$ and $\boldsymbol{l}$.

- The model was scored on precision (P), recall (R) and F1 score on detecting mispronounced segments

# Results: Pronunciation Error Detection

The model performed better at predicting $a_3$ while $a_2$ showed the worst metrics.

| | Train | | | Test | | |
|---|---|---|---|---|---|---|
| $\eta$ | **P** | **R** | **F1** | **P** | **R** | **F1** |
| *a1* | 0.7498 | 0.7923 | 0.7705 | 0.6489 | 0.6620 | 0.6554 |
| *a2* | 0.5861 | 0.8043 | 0.6781 | 0.4635 | 0.6124 | 0.5277 |
| *a3* | 0.8920 | 0.8276 | 0.8586 | 0.8507 | 0.7647 | 0.8054 |

# Experiment: Sensitivity to Assessor Tag

The sensitivity of **B** to tag $\eta$ scoring the data using the same previously unseen dummy $\eta_d$ for all annotators.

# Result : Sensitivity to Assessor Tag

- The model shows an overall decay in performance when using the wrong $\eta_d$ (top) compared to the matching η scores (bottom).

| $\eta_d$ | P | R | F1 | P | R | F1 |
|---|---|---|---|---|---|---|
| *a1* | 0.7449 | 0.7880 | 0.7659 | 0.6433 | 0.6780 | 0.6602 |
| *a2* | 0.4584 | 0.7107 | 0.5573 | 0.3505 | 0.5827 | 0.4377 |
| *a3* | 0.8546 | 0.7735 | 0.8120 | 0.8146 | 0.6981 | 0.7519 |

| | | Train | | | Test | |
|---|---|---|---|---|---|---|
| $\eta$ | P | R | F1 | P | R | F1 |
| *a1* | 0.7498 | 0.7923 | 0.7705 | 0.6489 | 0.6620 | 0.6554 |
| *a2* | 0.5861 | 0.8043 | 0.6781 | 0.4635 | 0.6124 | 0.5277 |
| *a3* | 0.8920 | 0.8276 | 0.8586 | 0.8507 | 0.7647 | 0.8054 |

# Experiment: Similarity Between A and MaxVoting

- The output of **A** was scored against each assessor and a MaxVoting reference (MAX).
    - MaxVoting is often used as inter-annotator agreement.

# Result: Similarity Between A and MaxVoting

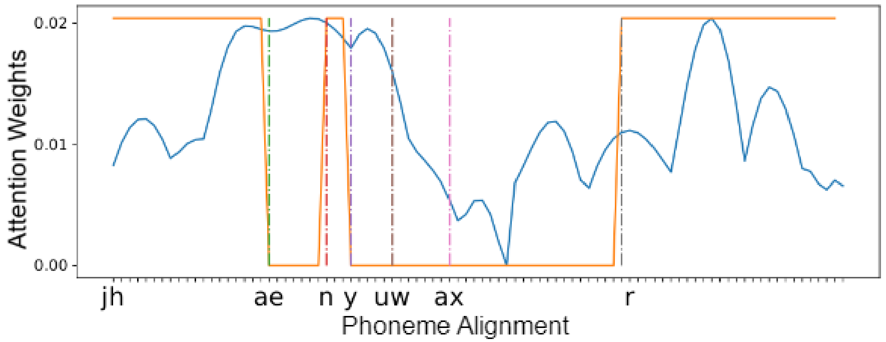
- The **A** output was better at scoring $a_3$ than MAX.

| | Train | | | Test | | |
|---|---|---|---|---|---|---|
| $\eta$ | **P** | **R** | **F1** | **P** | **R** | **F1** |
| *a1* | 0.6345 | 0.6885 | 0.6604 | 0.5480 | 0.6434 | 0.5919 |
| *a2* | 0.4156 | 0.6736 | 0.5141 | 0.3171 | 0.6112 | 0.4176 |
| *a3* | 0.8165 | 0.7211 | 0.7659 | 0.7739 | 0.6864 | 0.7275 |
| *MAX* | 0.6421 | 0.7126 | 0.6755 | 0.5424 | 0.6592 | 0.5951 |

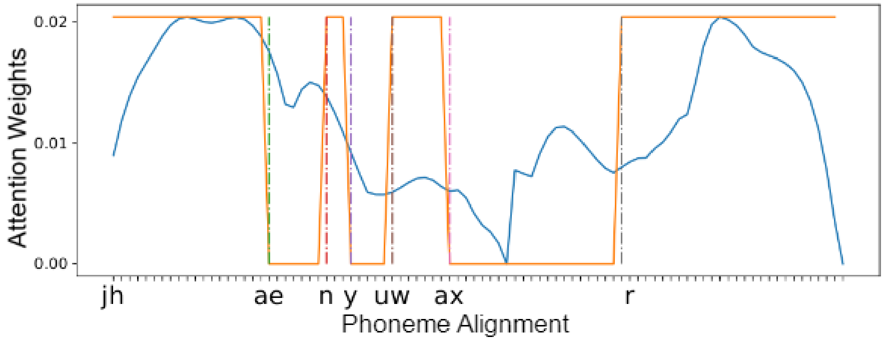
# Analysis of Normalised Attention Curves

- Normalised Attention weights (blue)
- Correctness Label (orange)
  - Correct = High position
  - Incorrect = Low Position

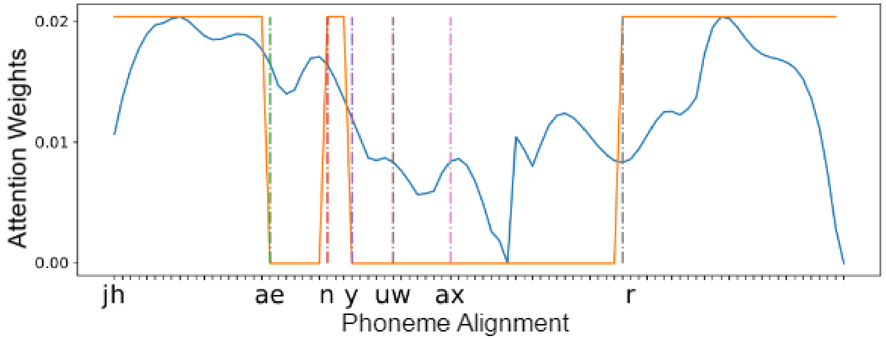
$A_{MAX}$




$B_{a_2}$




$B_{a_3}$



# Conclusion

- This work introduced an interpretable model for automatic PA consisting of an assessor independent and a bias term, implemented using a pair of ASIMs A and B.
- Model B was sensitive to $\eta$ and would decrease in its performance considerably if the wrong assessor tag was used.
- Model A was more similar to assessor $a_3$ than to the MAX reference when evaluated on its own.
- The disagreement between assessors could be observed from the attention curves in B.

# Selected References


1. Stephanie Lindemann, “Variation or ‘error’? perception of pronunciation variation and implications for assessment,” Second Language Pronunciation Assessment, p.193, 2017.

2. Jose Antonio Lopez Saenz, Md Asif Jalal, Rosanna Milner, and Thomas Hain, “Attention based model for segmental pronunciation error detection,” in 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2021.

3. Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio, “Neural machine translation by jointly learning to align and translate,” 3rd International Conference on Learning Representations, ICLR 2015 -Conference Track Proceedings, pp. 1–15, 2015.

4. Mauro Nicolao, Amy V. Beeston, and Thomas Hain, “Automatic assessment of English learner pronunciation using discriminative classifiers,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Apr 2015, pp. 5351–5355, IEEE

Thank you for Listening!
Questions?