A Model for Assessor Bias in Automatic Pronunciation Assessment

Jose Antonio Lopez Saenz, Thomas Hain   Speech and Hearing Research, Department of Computer Science, University of Sheffield, UK

Introduction

- 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.
- The bias in PA is a matter of inter-rater reliability attesting the lack of ground truth.
- A model for the assessor bias will benefit PA for the sake of a fair evaluation.

Key Contribution

An interpretable model for the assessor bias in automatic PA consisting of an assessor independent and an assessor sensitive bias term.

Model for the Assessor Bias

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

$$A(\mathbf{O}^{(w)}) = \frac{1}{H}\sum_{\eta\in H}[A_\eta(\mathbf{O}^{(w)}) - b_\eta(\mathbf{O}^{(w)})]$$

Where:
$\mathbf{O}^{(w)}$: Speech segment associated to prompt w.
$A_\eta(\mathbf{O}^{(w)})$: PA scoring function given assessor η.
$A(\mathbf{O}^{(w)})$: Assessor independent PA scoring function.
$b_\eta(\mathbf{O}^{(w)})$: Bias term given assessor η.

Mispronunciation Detection

$$P(\text{Error}|\mathbf{O}^{(w)}) = 1 - P(\mathbf{l}=1|\mathbf{r},\mathbf{O}^{(w)})$$

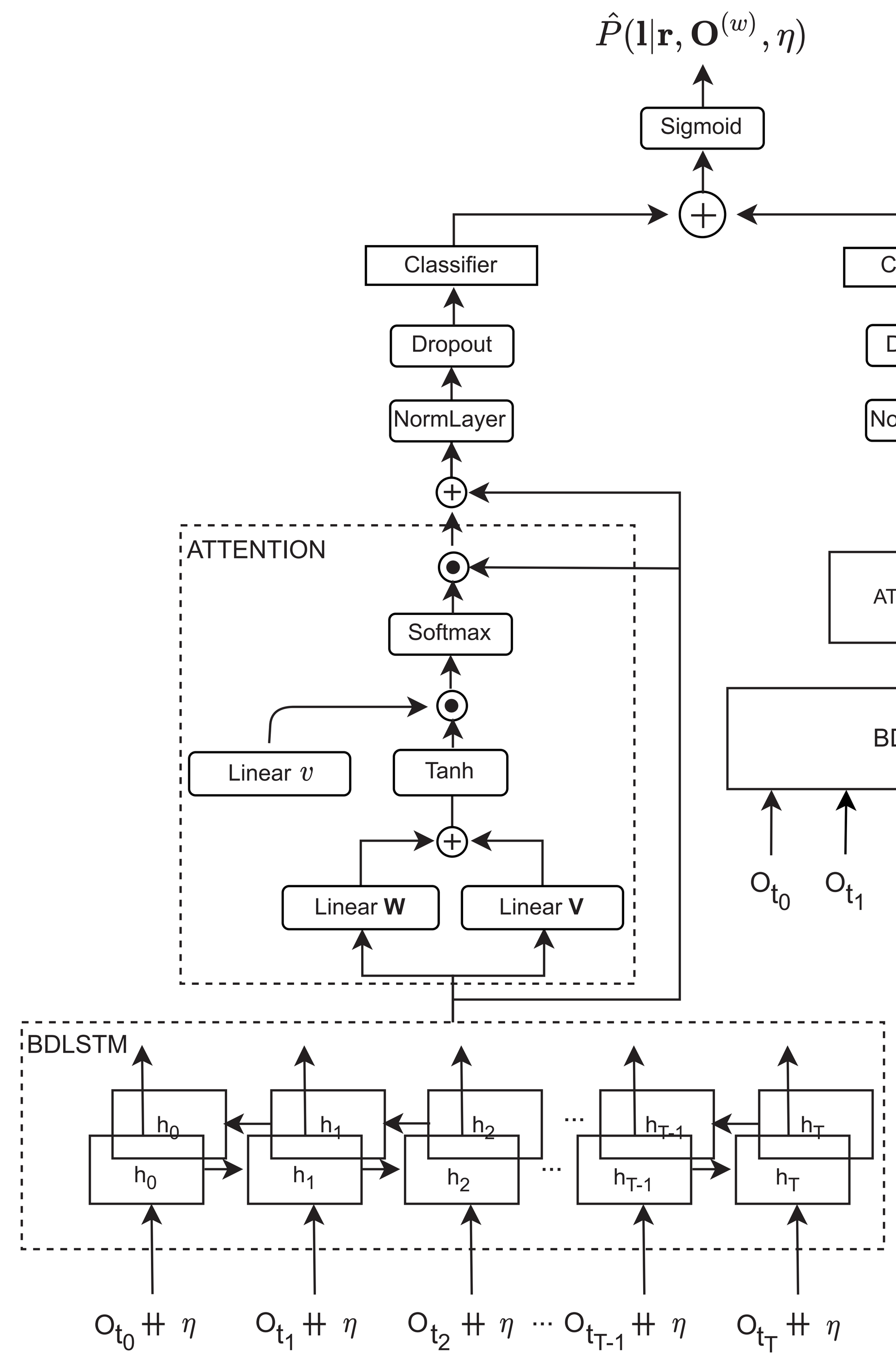
$$P(\mathbf{l}=1|\mathbf{r},\mathbf{O}^{(w)}) = \prod_i P(l_i=1|r_i,\mathbf{O}^{(w)})$$

Where:
$\mathbf{r} = \{r_i; i=0,\dots,R\}$: a phoneme sequence assumed canonical.
$\mathbf{l} = \{l_i; i=0,\dots,R\}$: a binary label where $l_i = 1$ given r_i is marked as correctly pronounced.

Dual Incorrectness Model

The Dual Attention-Based Segmental Incorrectness Model (ASIM) [1] approximates $A_\eta(\mathbf{O}^{(w)})$ as:

$$P(\hat{\mathbf{l}}|\mathbf{r},\mathbf{O}^{(w)},\eta) = P(\hat{\mathbf{l}_\mathbf{A}}|\mathbf{r},\mathbf{O}^{(w)}) + P(\hat{\mathbf{l}_\mathbf{b}}|\mathbf{r},\mathbf{O}^{(w)},\eta)$$



Data

INA set from the ITSLANG corpus of prompted L2 speech of teenage students of English in the Netherlands [2]. The data was annotated by three professional assessors (*a1*, *a2*, *a3*) with agreement percentage (I) and Cohen's kappa (κ):

vs.		I	κ
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

Experimental Setup

Sequence r: Forced-alignment via DNN-HMM trained WSJCAM0 + 46hrs ITSL $\notin$ INA.
Segments $\mathbf{O}^{(w)}$: Sliding window 0.5s size and 0.05s stride.
No Speaker Overlap.

Result - Error Detection

The model was scored on precision (**P**), recall (**R**) and **F1** score on detecting mispronounced segments given η. The model performed better at predicting *a3* while *a2*showed the worst metrics.

	Train			Test		
η	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**

Result - Assessor Sensitivity

The data was scored using a previously unseen η_d or all annotators. The model shows an overall decay in performance when using the wrong η

	Train			Test		
η	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

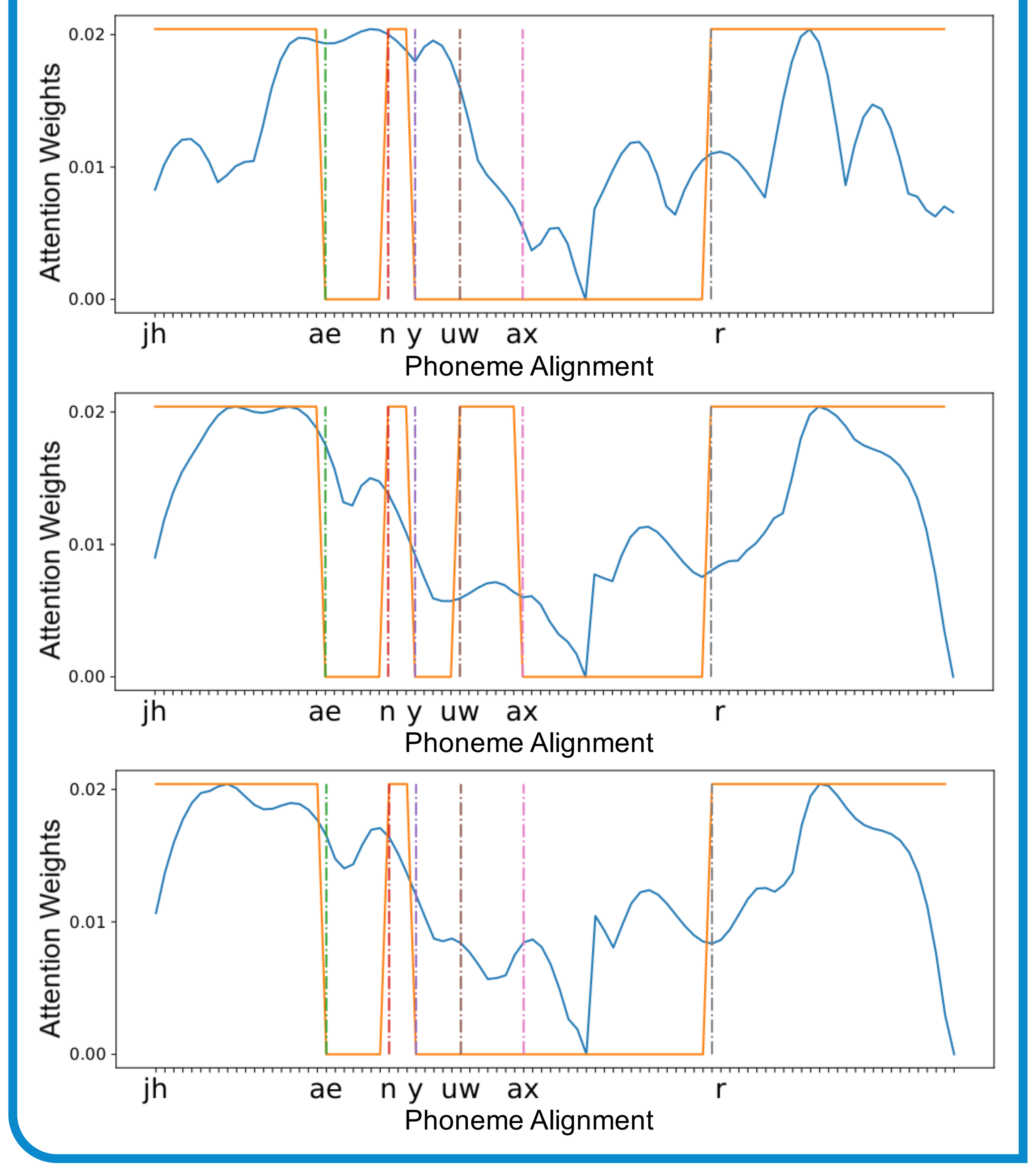
Result - MaxVoting Scoring

Output $\hat{\mathbf{l}_\mathbf{A}}$ was scored against each assessor and a MaxVoting reference (**MAX**), matching *a3* better.

	Train			Test		
η	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

Attention Curve Analysis

The normalised attention curves (blue) for both outputs focused differently on the same acoustic events. The curves for $\hat{\mathbf{l}_\mathbf{A}}$ (top), $\hat{\mathbf{l}_\mathbf{b}}$ for *(a2)* (mid) and *(a3)* (bottom) indicate points of disagreement across assessors on the same observation.



Conclusion

- This work introduced an interpretable model for automatic PA consisting of an assessor independent and a bias term, implemented using a dual ASIM.
- Output $\hat{\mathbf{l}_\mathbf{b}}$ was sensitive to η and would decrease in its performance considerably if the wrong assessor tag was used.
- Output $\hat{\mathbf{l}_\mathbf{A}}$ was more similar to assessor *a3* than to a MAXVoting reference when scored on its own.
- The disagreement between assessors could be observed from the attention curves for $\hat{\mathbf{l}_\mathbf{b}}$.

Selected References

1 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.

2 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.

Acknowledgements

This work was possible with the help of ITSLanguage BV for the data facilitated.