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

ABSTRACT

In pronunciation assessment, the assessor's perception is influenced by a particular pronunciation template. This assessor may hold a bias towards certain variations in pronunciation which do not necessarily impact communication, yet they may be penalized during the assessment. This work proposes a model for pronunciation assessment as the combination of an assessor independent (A) and an assessor specific (B) component. The latter could be interpreted as the assessor bias. The resulting assessment function was implemented as a dual model trained to detect mispronounced speech segments. The models incorporate Long-Short Memory and saliency region selection using attention. An experiment was performed using recordings from young Dutch learners of English as second language, which were annotated for mispronunciation by three trained phoneticians (*a1*, *a2*, *a3*). The models combined were able to detect mispronunciations given the assessor identity achieving F1 scores of 0.77, 0.68 and 0.86 for *a1*, *a2*, *a3* respectively on the Train set and 0.66, 0.53 and 0.81 on the Test set. Additionally, the attention weights of the B model were able to illustrate disagreements between assessors related to the bias.

Index Terms— pronunciation assessment, perception bias, L2 learning

1. INTRODUCTION

Speech relies on interlocutors being able to produce and perceive meaningful sounds, known as phonemes, as defined in a pronunciation reference assumed to be canonical. Equally, distinct accents of the same *native language* (L1) show that a constant realisation across speakers is not essential for using a language. As phonemes are conditioned to meaning, a speaker can deviate from the reference without interfering with the intended meaning as long as the linguistic structure remains close and the phonemic variation is kept [1].

In pronunciation assessment of *second language* (L2) speakers, an assessor declares the proficiency of a speaker in communicating using a canonical reference. The assessor's perception plays a key role in determining the identity of the phonemes produced by the speaker. The phonemic variations of L2 speakers are more noticeable than the ones of L1 speakers and likely to cause a bias in the assessor having an effect, positive or negative, on the rating of a speaker [1][2][3][4][5][6]. The existence of a bias in pronunciation assessment has been previously acknowledged, yet for *Automatic Pronunciation Assessment* (APA) it is a matter of inter-rater reliability associated with a lack of ground truth. A model for the assessor bias will benefit pronunciation assessment for the sake of a fair evaluation of a speaker.

This work introduces a model for the assessor bias as part of an assessor dependent scoring function for APA. The function is represented as the contribution of an assessor independent and an assessor specific scoring function, the latter being responsible for the bias. The assessment function was implemented using a combination of sequence encoding and saliency region selection using attention for detecting mispronounced segments. The corresponding models were tested for learning the annotation of three trained assessors on data of young Dutch learners of English as L2.

2. MODEL FOR THE ASSESSOR BIAS

Consider a speech segment $\mathbf{O}$ observed by a pronunciation assessor η. The assessor uses a scoring function $A_\eta(\mathbf{O}^{(w)})$ to estimate the probability of the *correct* pronunciation of a known word or prompt w. A function $D_\eta(\mathbf{O}^{(w)})$ acts as a decision threshold for up to which degree of correctness $\mathbf{O}^{(w)}$ is declared *mispronounced* or not. It is further assumed that a listener independent assessment function $A(\mathbf{O}^{(w)})$ exists, which remains unknown. In the proposed model, the relationship between this function and $A_\eta(\mathbf{O}^{(w)})$ is represented as an additive term $b_\eta(\mathbf{O}^{(w)})$.

$$A_\eta(\mathbf{O}^{(w)}) = A(\mathbf{O}^{(w)}) + b_\eta(\mathbf{O}^{(w)}) \quad (1)$$

Eq. 1 is always true as long as no constraints are defined for $b_\eta(\mathbf{O}^{(w)})$. This additive term can be loosely referred as the assessor bias.

A model that approximates Eq. 1 can be obtained using the annotation of assessor η. The following task is to find a model for the bias-free component $A(\mathbf{O}^{(w)})$ for each η. For this, assuming it is possible to estimate $b_\eta(\mathbf{O}^{(w)})$, the bias-free function can be computed as the average of the corrected

Jose Antonio Lopez Saenz is a doctoral student from Programa de Becas en el Extranjero from CONACYT with the fellowship number 661687 at the University of Sheffield. We also want to thank ITSLanguage B.V. for the data provided.

models across H assessors.

$$\frac{1}{H}\sum_{\eta\in H} A_\eta(\mathbf{O}^{(w)}) = \frac{1}{H}\sum_{\eta\in H}[A(\mathbf{O}^{(w)}) + b_\eta(\mathbf{O}^{(w)})] \quad (2)$$

$$A(\mathbf{O}^{(w)}) = \frac{1}{H}\sum_{\eta\in H}[A_\eta(\mathbf{O}^{(w)}) - b_\eta(\mathbf{O}^{(w)})] \quad (3)$$

3. SEGMENT BASED AUTOMATIC PRONUNCIATION ASSESSMENT

A usual approach for APA is to locate and score phoneme segments using a model based on data assumed correct [7][8][9][10]. This process is strongly prone to alignment [11]. A segment based approach is used instead for detecting the presence of a phoneme rather than its precise location.

For a phoneme sequence $\mathbf{r} = \{r_i; i = 1, \dots, R\}$ considered to be a canonical pronunciation associated with $O^{(w)}$, there is a binary correctness indicator $\mathbf{l} = \{l_i; i = 1\dots, R\}$. In $\mathbf{l}$, $l_i = 1$ if the phoneme r_i is marked as correctly pronounced and $l_i = 0$ otherwise. The real uttered phoneme sequence $\mathbf{s} = \{s_j; j = 1, \dots, S\}$ is not necessarily equal to $\mathbf{r}$. A correctness indicator could be assigned to $\mathbf{s}$, however the association with $\mathbf{l}$ may not be trivial. The probability of detecting a mispronunciation in $\mathbf{O}^{(w)}$ in terms of $\mathbf{r}$ and $\mathbf{l}$ is:

$$P(\text{mispronunciation}|\mathbf{O}^{(w)}) = 1 - P(\mathbf{l} = 1|\mathbf{r}, \mathbf{O}^{(w)}). \quad (4)$$

To keep the model practical, it is assumed $\mathbf{r}$ is known. Mispronunciation independence is also assumed:

$$P(\mathbf{l} = 1|\mathbf{r}, \mathbf{O}^{(w)}) \cong \prod_{i=1}^{R}(l_i = 1|\mathbf{r}, \mathbf{O}^{(w)}) \quad (5)$$

No information about the real sequence $\mathbf{s}$ nor precise timing information is needed as the model focuses mainly on detecting the presence of r_i and whether $l_i = 1$. An explicit model of the relationship between $\mathbf{r}$ and $\mathbf{s}$ may lead to re-weighting with prior common mispronunciations. The insertion of prior knowledge is avoided as it would require to sum over all possible combinations of real and canonical phoneme sequences, making it impractical.

The estimation of the correctness indicator $\mathbf{l}$ depends only on $\mathbf{r}$ and $\mathbf{O}^{(w)}$. This is achieved using a combination of spatial-sequential encoding and self-attention as detailed below.

4. ATTENTION-BASED SEGMENTAL INCORRECTNESS MODEL

An Attention-Based Segmental Incorrectness Model (ASIM) used for estimating Eq. 5 consists of three stages: sequential encoding, self-attention and segment classification. The initial encoding uses *Bidirectional Long Short-Term Memory* (BDLSTM) [12]. The aim is to exploit spatio-sequential relationships in $\mathbf{O}^{(w)}$ and to avoid the need for a precise alignment of non-canonical pronunciations. The use of attention weights on the BDLSTM hidden state outputs $\mathbf{h}_O = \{h_{o_{t_0}}, \dots, h_{o_T}\}$ for a saliency region selection enhances critical features to improve the performance of the upcoming classification stage [13]. The self-attention mechanism [14] computes the energy $e_{c,t}$ as defined in Eq. 6, where v_i, W_i and V_i are weight matrices. The energy is normalised over time to calculate the attention weights $\alpha_{c,t}$ (Eq. 7). The attention vector ψ is obtained from the element-wise multiplication ($\odot$) between α and $\mathbf{h}_O$. A residual connection is implemented by adding $\mathbf{h}_O$ to ψ easing the flow of the gradient [15][16]. A normalisation layer [17] and regularization via dropout ($p = 0.1$) are applied before the final classification stage.

$$e_{c,t} = v_c \odot \tanh(W_c^\top h_{o_t} + V_c^\top h_{o_t}) \quad (6)$$

$$\alpha_{c,t} = \frac{\exp(e_{c,t})}{\sum_{q=0}^{T}(e_{c,q})} \quad (7)$$

A deep feed-forward network associates the encoding with the correctness indicator $\mathbf{l}$. The final layer holds two linear outputs for each r_i, corresponding to the binary label l_i. This configuration allows the detection of incorrect pronunciations which may not be different enough to confuse with a different phoneme class. The network learns $P(l_i|r_i, \mathbf{O}^{(w)})$ as a multi-label classification problem. Since $\mathbf{r}$ is known, the output for any phoneme not in $\mathbf{r}$ is masked during training and the later scoring of a segment. This model was first introduced in [18] as an interpretable model for mispronunciation detection without the need of a precise alignment.

Since the objective of this work is to model the bias term $b_\eta(\mathbf{O}^{(w)})$, two ASIMs A and B are trained jointly to approximate the assessor independent and the bias component of Eq. 1 respectively. Both models observe the same acoustic feature frames, only B receives the assessor identity tag η concatenated ($+\!\!+$) as a constant dimension. The models' outputs are added and passed through a softmax layer to estimate $P(\mathbf{l}|\mathbf{r}, \mathbf{O}^{(w)}, \eta)$. The proportions in which the models contribute to the final posterior is not controlled, yet we consider using attention-based mixture in the future. It is expected that A generalizes over η while the bias B adjusts the output to better predict each assessor. This dual ASIM setup is shown in Figure 1.

5. DATA

The models were trained on recordings of Dutch children and early teenage learners of English as L2. The data was collected in various schools across the Netherlands as part of the ITSLanguage (ITSL) corpus from ITSLanguage B.V. The corpus consists of prompted speech recordings collected in classrooms, hence different environment conditions and noise

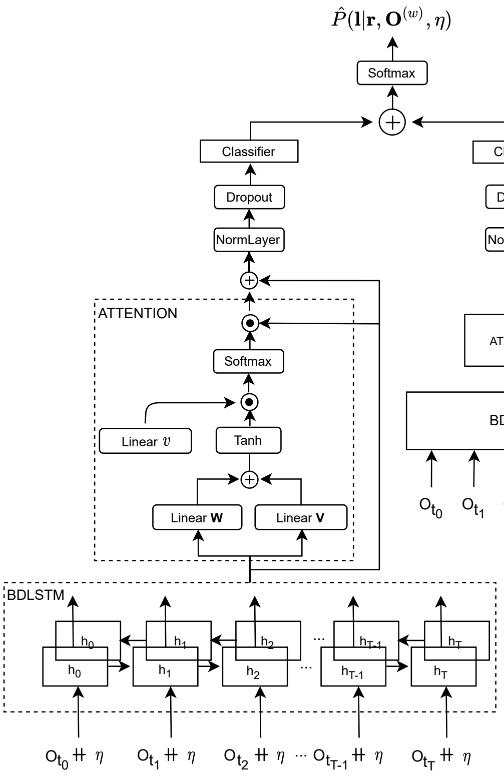
$\hat{P}(\mathbf{1}|\mathbf{r},\mathbf{O}^{(w)},\eta)$
Softmax
Classifier
Classifier
Dropout
Dropout
NormLayer
NormLayer
ATTENTION
ATTENTION
Softmax
Linear v
Tanh
Linear **W**
Linear **V**
BDLSTM
BDLSTM
h_0 h_1 h_2 ... h_{T-1} h_T
o_{t_0} o_{t_1} o_{t_2} ... o_{T-1} o_T
$o_{t_0} + \eta$ $o_{t_1} + \eta$ $o_{t_2} + \eta$ ... $o_{t_{T-1}} + \eta$ $o_{t_T} + \eta$

Fig. 1. Diagram for the dual ASIM setup. The left path observes the assessor tag η concatenated to the input $\mathbf{O}$ of length T.

are present. The students read from a list of 193 short sentences and isolated words and were able to re-record any element themselves until they were gruntled with their pronunciation.

Recordings with low levels of clipping and distortion were selected for the dataset (INA) used in this work. A total of 6 hours of speech from 238 speakers were annotated for mispronunciation at phoneme level by 3 trained phoneticians (*a1*, *a2*, *a3*). The inter-annotation agreement percentage (I) and Cohen's kappa (κ) for INA is shown in Table 1. The INA speakers range in age, L1, their L2 proficiency level and Dutch dialects used. The inter-speaker variability is useful to model the perception bias since the overlapping individual annotations show the disagreement between assessors.

6. EXPERIMENTS

The ASIMs A and B were trained to approximate the components of the proposed assessor scoring function (Eq. 1). The configuration used for each ASIM consisted of 6 BDLSTM layers of size 64, a self-attention module with matrix components of size 128 and a classifier with 6 linear layers of size 1024. The dual ASIM was scored for detecting segments with at least one mispronunciation marked given the assessor η. The performance metrics used were Precision (**P**), Recall (**R**) and **F1** score. The models were trained on data from all three assessors jointly. The data was split into 85% and 15% for Train and Test, respectively. The split was balanced for sex, age, L2 proficiency level without speaker overlap. The first 13 Perceptual Linear Predictor Coefficients with their first and second order differentials were used as feature vectors. The phoneme set $\mathbf{r}$ for each $\mathbf{O}^{(w)}$ was obtained via forced-alignment using a triphone-based DNN-HMM acoustic model trained on WSJCAM0 [19] and 46 hours of ITSL data excluding INA, as outlined in [20]. To help overcome alignment errors, only phoneme segments which alignment boundaries fell within at least two frames inside a segment were assumed part of $\mathbf{r}$. The canonical $\mathbf{r}$ vectors had a mean length of 3.46 phonemes and a standard deviation of 1.54. The models were trained using the Adam optimizer with a binary cross entropy.

Table 1. Inter-annotation agreement percentage (I) and Cohen's kappa (κ) for each assessor pair of the INA set

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

7. RESULTS AND DISCUSSION

The results for the dual ASIM on predicting the annotations for each assessor are shown in Table 2. Results for *a3* showed the best performance with an F1 score of 0.8586 and 0.8054 for Train and Test, respectively; results for *a1* were the second best with F1 scores of 0.7705 and 0.6554 and *a2* scored the lowest with 0.6781 and 0.5277 F1 scores for Train and Test. To test the sensitivity of B to the tag η, the data was scored using a previously unobserved dummy $\eta_d = -2.0$. The original assessor tags used were $\eta_{a1} = -1.22$, $\eta_{a2} = 0$ and $\eta_{a3} = 1.22$.

The results in Table 3 show an overall decay on all the performance metrics for predicting the assessment of *a2* and *a3* compared to the results in Table 2. This confirmed an effect from η in B. The change in the results for *a1* was less noticeable most likely due to the proximity of η_{a1} to η_d as it could not be confused with the other assessors; a similar behaviour was observed for *a3* when the data was scored using $\eta_d = 2.0$. As expected, the tag η acted as the assessor selector for the bias component.

The drop in performance shown in Table 3 had a limit due to the level of inter-assessor agreement (Table 1). Model A is expected to be assessor independent, yet not necessarily similar to the agreement across raters. A consolidated reference MAX was obtained via a majority vote criteria over the asses-

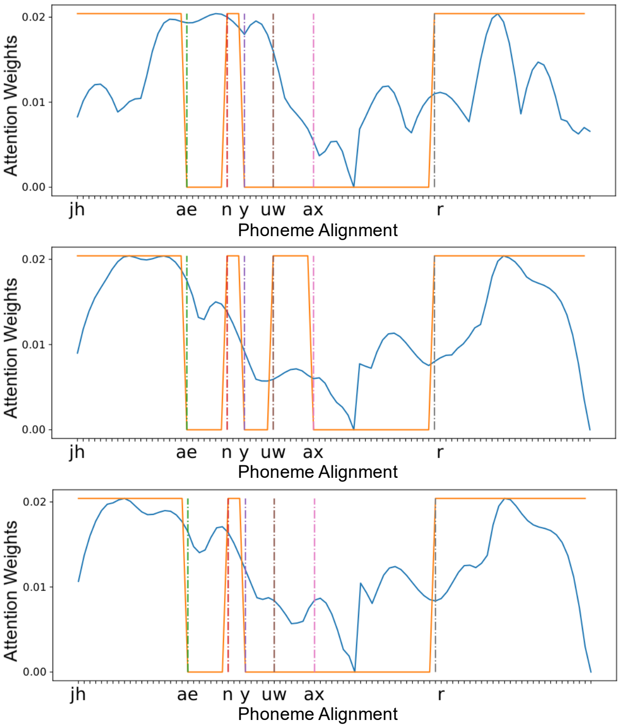

Fig. 2. Attention curves for A (top), B_{a2} (mid) and B_{a3} (bottom). The normalised attention curve is shown in blue and the correctness indicator $\mathbb{l}$ in orange. The MAX reference is used for A while B_{a2} and B_{a3} use their corresponding assessor's $\mathbb{l}$.

Table 2. Precision (P), Recall (R) and F1 score for the combined ASIMs on detecting segments with mispronunciation for each assessor η.

	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

Table 3. Precision (P), Recall (R) and F1 score for the combined ASIMs on detecting segments with mispronunciation for each assessor η while using the dummy tag $\eta_d = -2.0$

	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

Table 4. Precision (P), recall (R) and F1 score for A on detecting segments with mispronunciation for each assessor η and the MAX reference.

	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

sors; this is often assumed to represent the most agreement. The MAX annotation showed κ values of 0.818, 0.555 and 0.571 for *a1*, *a2* and *a3* respectively. The outputs from A were used to predict each assessor and the MAX reference to look for the similarity between A and MAX; the results are shown on Table 4. Assessor *a1* showed the least decrease in the performance on the test set, although the ranking of best to worst assessor prediction remained the same (*a3*, *a1*, *a2*). The MAX reference behaved very similar to *a1*, which was not surprising due the high κ between them. Model A was better at predicting *a3* with the highest F1 scores of 0.7659 and 0.7275 for Train and Test. Regardless of the limited amount of assessors used, A was not similar to criteria intended to represent a lower disagreement in annotation.

Finally, A and B seemed to focus on similar features with some particularities dependant on η. Figure 2 shows the normalized attention weights for A and B for an utterance showing disagreement between *a2* and *a3*. The plots correspond to the attention curves in blue for A (top), B_{a2} (mid) and B_{a3} (bottom). The correctness indicator $\mathbb{l}$ is shown as the orange curve, a high position means the phoneme is marked as *correct* and a low position means *incorrect*. The MAX reference is plotted with A at the top, showing the assessment of *a2* disagrees with the other two assessors. The first thing to notice is the similar tendency between the three plots, yet A and B differ mainly in the longer correct segments. Although the three plots show the same decay and almost vertical rising at segment $/ax/$, only B_{a2} spiked on the disagreement at $/uw/$. The interpretability of the ASIM showed the model is capable of locating disagreements between assessors and useful for illustrating the effects of the assessor bias.

8. CONCLUSIONS

This work introduced a dual model for APA that consists of an assessor independent and an additive bias term. The model was implemented using a pair of ASIMs A and B, the latter performing as the bias. The ASIMs combined showed a good performance for predicting the assessment of three trained phoneticians on the pronunciation of young Dutch learners of English as L2. Model B was sensitive to η and would decrease in its performance considerably if the wrong assessor tag was used. The component A was more similar to assessor *a3* than to the MAX reference when evaluated on its own. Finally, the attention curves of A and B focused differently on the same acoustic events. In the case of B, the disagreement between assessors could be observed from the attention curve making this feature of the ASIM a key tool for the further exploration of the perception bias.

9. REFERENCES

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

[2] Marijt J. Witteman, Andrea Weber, and James M. McQueen, “Tolerance for inconsistency in foreign-accented speech,” *Psychonomic Bulletin & Review*, vol. 21, no. 2, pp. 512–519, apr 2014.

[3] Luke Harding, “What Do Raters Need in a Pronunciation Scale? The User’s View,” in *Second Language Pronunciation Assessment: Interdisciplinary Perspectives*, Talia Isaacs and Pavel Trofimovich, Eds., chapter 2, pp. 12–34. Multilingual Matters / Channel View Publications, 2017.

[4] John Levis, “Assessing speech intelligibility: Experts listen to two students,” *Pronunciation and Intelligibility: Issues in Research and Practice*, p. 56, 2010.

[5] James Milroy, “Language ideologies and the consequences of standardization,” *Journal of Sociolinguistics*, vol. 5, no. 4, pp. 530–555, Nov 2001.

[6] Michael D. Carey, Robert H. Mannell, and Peter K. Dunn, “Does a rater’s familiarity with a candidate’s pronunciation affect the rating in oral proficiency interviews?,” *Language Testing*, vol. 28, no. 2, pp. 201–219, 2011.

[7] Guimin Huang, Jing Ye, Zhenglin Sun, Ya Zhou, Yan Shen, and Ruyu Mo, “English mispronunciation detection based on improved gop methods for chinese students,” in *2017 International Conference on Progress in Informatics and Computing (PIC)*, 2017, pp. 425–429.

[8] Sweekar Sudhakara, Manoj Kumar Ramanathi, Chiranjeevi Yarra, and Prasanta Kumar Ghosh, “An improved goodness of pronunciation (gop) measure for pronunciation evaluation with dnn-hmm system considering hmm transition probabilities.,” in *INTERSPEECH*, 2019, pp. 954–958.

[9] Sweekar Sudhakara, Manoj Kumar Ramanathi, Chiranjeevi Yarra, Anurag Das, and Prasanta Kumar Ghosh, “Noise robust goodness of pronunciation measures using teacher’s utterance.,” in *SLaTE*, 2019, pp. 69–73.

[10] Jiatong Shi, Nan Huo, and Qin Jin, “Context-aware goodness of pronunciation for computer-assisted pronunciation training,” *arXiv preprint arXiv:2008.08647*, 2020.

[11] Shiran Dudy, Steven Bedrick, Meysam Asgari, and Alexander Kain, “Automatic analysis of pronunciations for children with speech sound disorders,” *Computer Speech and Language*, vol. 50, pp. 62–84, 2018.

[12] Hasim Sak, Andrew Senior, and Françoise Beaufays, “Long short-term memory recurrent neural network architectures for large scale acoustic modeling,” 2014.

[13] Rosanna Milner, Md Asif Jalal, Raymond WM Ng, and Thomas Hain, “A cross-corpus study on speech emotion recognition,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 304–311.

[14] 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.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Dec, pp. 770–778, 2016.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*. jun 2017, vol. 2017-Dec, pp. 5999–6009, Neural information processing systems foundation.

[17] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, “Layer normalization,” *arXiv preprint arXiv:1607.06450*, 2016.

[18] 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.

[19] Tony Robinson, Jeroen Fransen, David Pye, Jonathan Foote, and Steve Renals, “Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition,” in *1995 International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1995, vol. 1, pp. 81–84.

[20] 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.