Detection of Lexical Stress Errors in Non-Native (L2) English with Data Augmentation and Attention

Daniel Korzekwa[1,2], Roberto Barra-Chicote[3], Szymon Zaporowski[2], Grzegorz Beringer[1],
Jaime Lorenzo-Trueba[3], Alicja Serafinowicz[1], Jasha Droppo[4], Thomas Drugman[3], Bozena Kostek[2]

[1]Amazon, Poland
[2]Gdansk University of Technology, Faculty of ETI, Poland
[3]Amazon, UK
[4]Amazon, USA

`korzekwa@amazon.com`

Abstract

This paper describes two novel complementary techniques that improve the detection of lexical stress errors in non-native (L2) English speech: attention-based feature extraction and data augmentation based on Neural Text-To-Speech (TTS). In a classical approach, audio features are usually extracted from fixed regions of speech such as the syllable nucleus. We propose an attention-based deep learning model that automatically derives optimal syllable-level representation from frame-level and phoneme-level audio features. Training this model is challenging because of the limited amount of incorrect stress patterns. To solve this problem, we propose to augment the training set with incorrectly stressed words generated with Neural TTS. Combining both techniques achieves 94.8% precision and 49.2% recall for the detection of incorrectly stressed words in L2 English speech of Slavic and Baltic speakers.

Index Terms: lexical stress, language learning, data augmentation, text-to-speech, attention, automated speech assessment

1. Introduction

Computer Assisted Pronunciation Training (CAPT) usually focuses on practicing pronunciation of phonemes [1, 2, 3], while there is evidence in non-native (L2) English speakers that practicing lexical stress improves speech intelligibility [4, 5]. Lexical stress is a syllable-level phonological feature. It is a part of the phonological rules that define how words should be spoken in a given language. Stressed syllables are usually longer, louder, and expressed with a higher pitch than their unstressed counterparts [6]. Lexical stress is inter-connected with phonemic representation. For example, placing lexical stress on a different syllable of a word may lead to different phonemic realizations known as 'vowel reduction' [7].

The focal point of our work is the detection of words with incorrect stress patterns. The training data with human speech is usually highly imbalanced, with few training examples of incorrectly stressed words. It makes training machine learning models for this task challenging. We address this problem by augmenting the training set with synthetic speech that is generated with Neural Text-To-Speech (TTS) [8]. Neural TTS allows us generating words with both correct and incorrect stress patterns.

Most of the existing approaches for automated lexical stress assessment are based on carefully designed features that are extracted from fixed regions of speech signal such as the syllable nucleus [9, 10, 11]. We introduce attention mechanism [12] to automatically learn optimal syllable-level representa-

tion. Attention-based approach originates from the intuition of how people detect specific patterns in high dimensional and unstructured data such as visual and speech signals [13]. For example, we might focus our attention on the duration ratio between nuclei of two neighboring syllables, incidentally, an important predictor of lexical stress. The syllable-level representation is derived from frame-level (F0, intensity) and phoneme-level (duration) audio features and the corresponding phonetic representation of a word. We do not indicate precisely the regions of the audio signal that are important for the detection of lexical stress errors. The attention mechanism does it automatically.

To the best of our knowledge, this paper is the first attempt, for the task of lexical stress error detection, to: *i)* augment the training data with Neural TTS, *ii)* use attention mechanisms to automatically extract syllable-level features for lexical stress error detection. Ruan et al. [14] used attention-based architecture of transformers for lexical stress detection. However, their paper concerns recognizing stressed and unstressed phonemes. They do not detect lexical stress errors, which is crucial in CAPT applications.

The paper is structured as follows. In Section 2, we review the related work. Section 3 describes the proposed model. Section 4 reviews human and synthetic speech corpora. In Section 5, we present our experiments, and Section 6 concludes the paper.

2. Related Work

The existing work focuses on the supervised classification of lexical stress using Neural Networks [15, 10], Support Vector Machines [11, 16] and Fisher's linear discriminant [17]. There are two popular variants: a) discriminating syllables between primary stress/no stress [9], and b) classifying between primary stress/secondary stress/no stress [18, 15]. Ramanathi et al. [19] have followed an alternative unsupervised way of classifying lexical stress, which is based on computing the likelihood of an acoustic signal for a number of possible lexical stress representations of a word.

Accuracy is the most commonly used performance metric, and it indicates the ratio of correctly classified stress patterns on a syllable [18] or word level [11]. On the contrary, following Ferrer et al. [9], we analyze precision and recall metrics because we aim to detect lexical stress errors and not just classify them.

Existing approaches for the classification and detection of lexical stress errors are based on carefully designed features. They start with aligning a speech signal with phonetic tran-
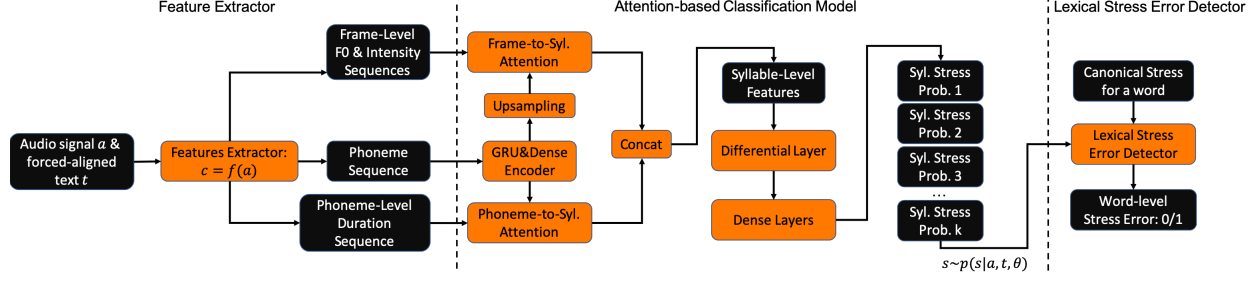
Figure 1: *Attention-based Deep Learning model for the detection of lexical stress errors.*

scription, performed via forced-alignment [10, 11]. Alternatively, Automatic Speech Recognition (ASR) can provide both phonetic transcription and its alignment with a speech signal [18]. Then, prosodic features such as duration, energy and pitch [11] and cepstral features such as MFCC and Mel-Spectrogram [9, 10] are extracted. These features can be extracted on the syllable [10] or syllable nucleus [9, 11] level.

Shahin et al. [10] computed features of neighboring vowels, and Li et al. [18] included the features for two preceding and two following syllables in the model. The features are often preprocessed and normalized to avoid potential confounding variables [9], and to achieve better model generalization by normalizing the duration and pitch on a word level [9, 17]. Li et al. [15] added canonical lexical stress to input features, which improves the accuracy of the model.

In our approach, we use attention mechanisms to derive automatically regions of the audio signal that are important for the detection of lexical stress errors. We also use data augmentation through the generation of artificial data with Neural TTS.

3. Proposed Model

The proposed model consists of three subsystems: Feature Extractor, Attention-based Classification Model, and Lexical Stress Error Detector. It is illustrated in Figure 1.

3.1. Feature Extractor

The Feature Extractor extracts prosodic features and phonemes from speech signal $\mathbf{a}$ and forced-aligned text $\mathbf{t}$. To obtain forced-alignment, we used Montreal toolkit [20] along with an acoustic model pretrained on LibriSpeech ASR corpus [21]. The prosodic features $\mathbf{c} = f(\mathbf{a})$ are formed by: F0, intensity [dB SPL] and phoneme-level durations. The F0 and intensity features are computed at the frame level using Praat library [22] (time step: 10 ms, window size: 40 ms). The F0 contour is linearly interpolated in unvoiced regions. These raw features will be further transformed by the attention-based model to the syllable-level representation.

3.2. Attention-based Classification Model

The Attention-based Classification Model maps frame-level and phoneme-level features to the syllable-level representation. Then, it produces a lexical stress pattern $\mathbf{s}$, modeled as a sequence of Bernoulli random variables $\mathbf{s} = \{s_1, .., s_k\}$ (stressed/unstressed) over K syllables of a multi-syllable word, conditioned on audio $\mathbf{a}$ and text $\mathbf{t}$ representations. Let us define it as a conditional probability distribution $\mathbf{s} \sim p(\mathbf{s}|\mathbf{a}, \mathbf{t}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the parameters of the model.

To extract syllable-level features, we use two dot-product

attentions operating on the frame and phoneme levels. To build better intuition on what these two attention do, in Figure 2 we show the frame-level and phoneme-level attention plots for the word 'garage' pronounced by a Polish speaker and incorrectly stressed on the first syllable in reference to American English. This word has a similar pronunciation but different lexical stress in Polish and American English languages ('G AA1 R AA0 ZH' vs 'G ER0 AA1 ZH'). Both attentions find the most relevant regions of the frame-level and phoneme-level features.

The dot-product attention is presented in Equation 1, and it follows the notation proposed by Vaswani et al. [12]. It is based on three inputs: Query ($\mathbf{Q}$), Keys ($\mathbf{K}$) and Values ($\mathbf{V}$), where d_k is the dimensionality of $\mathbf{K}$.

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^t}{\sqrt{d_k}})\mathbf{V} \qquad (1)$$

The attention inputs are represented as follows. Query refers to the syllable positional embeddings defined by one-hot syllable index encodings. Keys represents a sequence of sub-phonemes. Each sub-phoneme is represented by a set of features: $phoneme\_id$, $syllable\_index$, is_vowel, $left\_or\_right\_sub\_phoneme$. All features are one-hot encoded and processed with a Gated Recurrent Unit (GRU) layer [23] (units:4, dropout: 0.24). In the end, encoded sub-phoneme sequence is passed through linear dense layers. In the case of the frame-level attention, the encoded sub-phoneme sequence is upsampled to the frame level using phoneme durations from forced-alignment. In upsampling, we simply replicate phonemes across aligned frames of audio signal. Similar phoneme-to-frame upsampling has been recently adopted in Text-To-Speech [24]. Finally, Values are the $F0/intensity$ and $duration$ features for frame-level and phoneme-level attentions respectively.

To model relative prominence, we introduce a differential bi-directional layer that computes the ratios of syllable-level acoustic features for each syllable and its two neighbors (Figure 1). The bi-directional layer is implemented as a simple 'division' math operation and it does not contain any trainable parameters. The output of the differential layer is further processed by three dense layers (units: 4, activation: tanh, dropout: 0.24), followed by a linear dense layer (units: 2, dropout: 0.24) that produces a two-dimensional output for each syllable. It is then squeezed by a softmax function to generate lexical stress probabilities.

3.3. Training of the Classification Model

We train the model on a set of N triplets that contains 1) human recorded words and 2) synthetic words generated using Neural
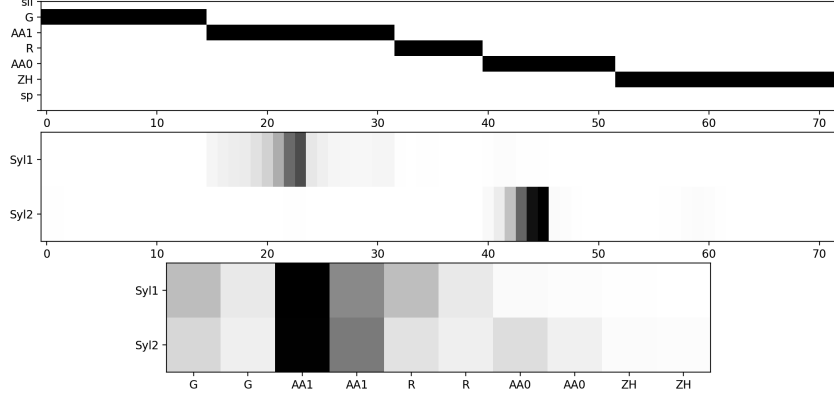
Figure 2: *Top: forced-alignment mapping between phonemes and frames for the word 'garage'. Middle: Frame-to-syllable attention weights matrix. Bottom: (Sub)Phoneme-to-syllable attention weights matrix.*

TTS. A single triplet is represented by $\{\mathbf{s_n}, \mathbf{a_n}, \mathbf{t_n}\}$, where $n = 1..N$ is the index of a training example.

The concept of data augmentation can be explained using a framework of Bayesian Inference. Consider three random variables, lexical stress $\mathbf{s_n}$, audio signal $\mathbf{a_n}$ and text $\mathbf{t_n}$. All variables are observed for the training examples of human speech. However, for the synthetic speech, we only observe the lexical stress and text variables. The audio signal is unobserved (hidden) because we have to generate it.

To train this model, we derive a negative log-likelihood loss over a joint probability distribution of lexical stress $\mathbf{s}$ and audio $\mathbf{a}$ random variables, as depicted in Equation 2. The loss is further approximated with the variational lower bound [25], as presented in Equation 3 (we omit $\boldsymbol{\theta}$ for brevity). For the training examples of synthetic speech, the conditional probability distribution over the audio signal $\mathbf{a_n} \sim p(\mathbf{a_n}|\mathbf{s_n}, \mathbf{t_n})$ is estimated with Neural TTS, and for human recorded words, it is given explicitly.

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_n^N log \int p(\mathbf{s_n}, \mathbf{a_n}|\mathbf{t_n}, \boldsymbol{\theta}) d\mathbf{a_n} \qquad (2)$$

$$log \int p(\mathbf{s_n}, \mathbf{a_n}|\mathbf{t_n}) d\mathbf{a_n} \approx E_{\mathbf{a_n} \sim p(\mathbf{a_n}|\mathbf{t_n}, \mathbf{s_n})}[log p(\mathbf{s_n}|\mathbf{a_n}, \mathbf{t_n})] \qquad (3)$$

The model was implemented in MxNet [26], trained with Stochastic Gradient Descent optimizer (learning rate: 0.1, batch size: 20) and tuned with Bayesian optimization [27]. Training data were split into buckets based on the number of frames in an audio signal, using Gluon-NLP package [28]. A single bucket contains words with the same number of syllables with zero-padded acoustic and sub-phoneme sequences.

3.4. Lexical Stress Error Detector

The Lexical Stress Error Detector reports on lexical stress error if the expected (canonical) and estimated lexical stress for a given syllable do not match and the corresponding probability is higher than a given threshold.

4. Speech Corpus

Our speech corpus consists of human and synthetic speech. The data were split into training and testing sets with disjointed speakers ascribed to each set. Human speech contains L1 and L2 speakers of English. Synthetic data were generated with Neural TTS and are included only in the training set. All audio files were downsampled to a 16 kHz sampling rate. The data are summarized in Table 1, and we provide more details in the following subsections.

Table 1: *Train and test sets details.*

Data set	Speakers (L2)	Words (unique)	Stress Errors
Train set (human)	473 (10)	8223 (1528)	425
Train set (TTS)	1 (0)	3937 (1983)	2005
Test set (human)	176 (21)	2108 (378)	189

4.1. Human Speech

Due to the limited availability of L2 corpora, we recorded our own L2-English corpus of Slavic and Baltic speakers. It also allows us to evaluate the model during interactive English learning sessions with our students. The corpus contains speech from 25 speakers (23 Polish, 1 Ukrainian and 1 Lithuanian): 7 females and 18 males, all between 24 and 40 years old. All speakers read a list of two hundred words. One hundred words were prepared by a professional English teacher, including frequently mispronounced words by Slavic and Baltic students. The second half consists of the most common words that were obtained from Google's Trillion Word Corpus [29] based on n-gram frequency analysis. We excluded abbreviations and one-syllable words.

Additionally, L1 and L2 English speech was collected from publicly available speech data sets, including TIMIT [30], Arctic [31], L2-Arctic [32] and Porzuczek [33].

4.2. Synthetic Speech

Complementary to human recordings, synthetic speech was generated with Neural TTS by Latorre et al. [8]. The Neural TTS consists of two modules. Context-generation module is an attention-based encoder-decoder neural network that generates a mel-spectrogram from a sequence of phonemes. Then, a Neural Vocoder converts it to the speech signal. The Neural Vocoder is a neural network of architecture similar to the work by [34]. The Neural TTS was trained using speech of a professional American voice talent. To generate words with different

lexical stress patterns, we modify lexical stress markers associated with the vowels in the phonemic transcription of a word. For example, with the input of /r iy1 m ay0 n d/ we can place lexical stress on the first syllable of the word 'remind'. 1980 popular English words were synthesized with correct and incorrect stress patterns.

4.3. Lexical Stress Annotations

L1 corpora were segmented into words and annotated automatically using a proprietary Amazon American English Lexicon, taking into account the syntactic context of the word. Neural TTS speech and the speech of L2 speakers were annotated by 5 American English linguists into 'primary' and 'no stress' categories, keeping the words for which a minimum of 4 out of 5 linguists agreed on the stress pattern. Annotators were not able to distinguish between primary and secondary lexical stress. 81.5% of synthesized words matched the intended stress patterns with a minimum of 4 annotators' agreement. It shows that Neural TTS can be used to generate incorrectly stressed speech.

5. Experiments

The proposed model (Att_TTS) from Section 3 is compared to three baseline models that are designed to measure the impact of the Neural TTS data augmentation and the attention mechanism. To compare these models, we plotted their precision-recall curves and gave their corresponding area under a curve (AUC) along with our results, see Figure 3.

The Att_NoTTS model has the same architecture as the Att_TTS, but the synthetic speech is excluded from the 'training set'. The NoAtt_TTS model uses the same training set as the Att_TTS, but it has no attention mechanism. Instead, as a syllable-level representation, it uses mean values of acoustic features for the corresponding syllable nucleus. The NoAtt_NoTTS model has no attention, and it does not use Neural TTS data augmentation.

As a state-of-the-art baseline, we use the work by Ferrer et al. [9]. However, a direct comparison is not possible. In their test corpus, there were 46.4% (191 out of 411) of incorrectly stressed words, far more than 9.4% (189 out of 2109) words in our experiment. The fewer lexical stress errors are made by users, the more challenging it is to detect it. They also used proprietary L2 English of Japanese speakers. Due to the lack of available benchmark and standard speech corpora for the task of lexical stress assessment, we could not make a fairer comparison with the state-of-the-art.

5.1. Experimental Results

First, we compare Att_NoTTS and NoAtt_NoTTS models. Using the attention mechanism for automatic extraction of syllable-level features significantly improves the detection of lexical stress errors. It is illustrated by precision-recall curves and AUC metric in Figure 3. To be comparable with the study by Ferrer et al., we fix recall to around 50% and compare the models using precision as shown in Table 2.

The Att_NoTTS attention-based can be further improved. Augmenting the training set with incorrectly stressed words (Att_TTS) boosts precision from 87.85% to 94.8%, at a recall level of 50%. Data augmentation helps because it increases the number of words with incorrect stress patterns in the training set. It prevents the model from exploiting a strong correlation between phonemes and lexical stress in correctly stressed words. Using data augmentation in the simpler no-
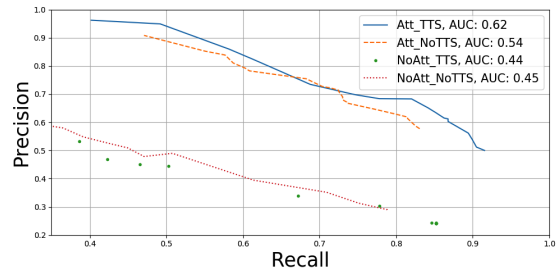

Figure 3: *Precision-recall curves for evaluated systems.*

attention-based model (NoAtt_TTS) does not help. It is because NoAtt_TTS uses only prosodic features for fixed regions of speech, so this model cannot overfit to phonetic input.

Table 2: *Precision and recall [%, 95% Confidence Interval] of detecting lexical stress errors, at around 50% recall. * - Ferrer et al. model has been evaluated on the data with 46.4% of lexical stress errors, compared to 9.4% of errors on our data set. This data point indicates that our proposed model AttTTS should outperform Ferrr et al. model if both were evaluated exactly in the same conditions.*

Model	Precision	Recall
AttTTS	94.8 (89.18-98.03)	49.2 (42.13-56.3)
AttNoTTS	87.85 (80.67-93.02)	49.74 (42.66-56.82)
NoAttTTS	44.39 (37.85-51.09)	50.26 (43.18-57.34)
NoAttNoTTS	48.98 (42.04-55.95)	50.79 (43.70-57.86)
Ferrer et al. [9] *	95.00 (na-na)	48.3 (na-na)

Ferrer et al. [9] reported on a similar performance to our Att_TTS model with a precision of 95% and a recall of 48.3% on L2 English speech of Japanese speakers. However, in their testing data, the proportion of incorrectly stressed words is much larger, which makes it easier to detect lexical stress errors.

6. Conclusion and Future Work

Using an attention-based neural network for the automatic extraction of syllable-level features significantly improves the detection of lexical stress errors in L2 English speech, compared to baseline models. However, this model has a tendency to classify lexical stress based on highly-correlated phonemes. We can counteract this effect by augmenting the training set with incorrectly stressed words generated with Neural TTS. It boosts the performance of the attention-based model by 14.8% in the AUC metric and by 7.9% in precision, while maintaining recall at a level close to 50%. Data Augmentation, however, does not help when applied to a simpler model without an attention mechanism.

We found that the current word-level model is not able to correctly classify lexical stress when two words are linked [35] and stress shift may occur [36]. For example, two neighboring phonemes /er/ in the text 'her arrange' /hh er - er ey n jh/ are pronounced as a single phoneme. Therefore, in future, we plan to move away from the assessment of isolated words and extend the current model to detect lexical stress errors at the sentence level. We plan to replace a single-speaker TTS model to generate synthetic lexical stress errors with a multi-speaker model. We plan to analyze the accuracy of detecting lexical stress errors for speakers with different proficiency levels of English.

7. References

[1] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[2] W.-K. Leung, X. Liu, and H. Meng, "Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8132–8136.

[3] D. Korzekwa, J. Lorenzo-Trueba, S. Zaporowski, S. Calamaro, T. Drugman, and B. Kostek, "Mispronunciation detection in non-native (l2) english with uncertainty modeling," in *Accepted to ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

[4] J. Field, "Intelligibility and the listener: The role of lexical stress," *TESOL quarterly*, vol. 39, no. 3, pp. 399–423, 2005.

[5] A. Lepage and M. G. Busà, "Intelligibility of english l2: The effects of incorrect word stress placement and incorrect vowel reduction in the speech of french and italian learners of english," in *Proc. of the Intl. Symposium on the Acquisition of Second Language Speech*, vol. 5, no. 2014, 2014, pp. 387–400.

[6] Y.-J. Jung, S.-C. Rhee *et al.*, "Acoustic analysis of english lexical stress produced by korean, japanese and taiwanese-chinese speakers," *Phonetics and Speech Sciences*, vol. 10, no. 1, pp. 15–22, 2018.

[7] D. R. v. Bergem, "Acoustic and lexical vowel reduction," in *Phonetics and Phonology of Speaking Styles*, 1991.

[8] J. Latorre, J. Lachowicz, J. Lorenzo-Trueba, T. Merritt, T. Drugman, S. Ronanki, and V. Klimkov, "Effect of data reduction on sequence-to-sequence neural tts," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7075–7079.

[9] L. Ferrer, H. Bratt, C. Richey, H. Franco, V. Abrash, and K. Precoda, "Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems," *Speech Communication*, vol. 69, pp. 31–45, 2015.

[10] M. A. Shahin, J. Epps, and B. Ahmed, "Automatic classification of lexical stress in english and arabic languages using deep learning." in *INTERSPEECH*, 2016, pp. 175–179.

[11] J.-Y. Chen and L. Wang, "Automatic lexical stress detection for chinese learners' of english," in *2010 7th Intl. Symposium on Chinese Spoken Language Processing*. IEEE, 2010, pp. 407–411.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[13] M. I. Posner and S. E. Petersen, "The attention system of the human brain," *Annual review of neuroscience*, vol. 13, no. 1, pp. 25–42, 1990.

[14] Y. Ruan, X. Wang, H. Liu, Z. Ou, Y. Gao, J. Cheng, and Y. Qian, "An end-to-end approach for lexical stress detection based on transformer," *arXiv preprint arXiv:1911.04862*, 2019.

[15] K. Li, S. Mao, X. Li, Z. Wu, and H. Meng, "Automatic lexical stress and pitch accent detection for l2 english speech using multi-distribution deep neural networks," *Speech Communication*, vol. 96, pp. 28–36, 2018.

[16] J. Zhao, H. Yuan, J. Liu, and S. Xia, "Automatic lexical stress detection using acoustic features for computer assisted language learning," *Proc. APSIPA ASC*, pp. 247–251, 2011.

[17] N. Chen and Q. He, "Using nonlinear features in automatic english lexical stress detection," in *2007 Intl. Conference on Computational Intelligence and Security Workshops (CISW 2007)*. IEEE, 2007, pp. 328–332.

[18] K. Li, X. Qian, S. Kang, and H. Meng, "Lexical stress detection for l2 english speech using deep belief networks." in *Interspeech*, 2013, pp. 1811–1815.

[19] M. K. Ramanathi, C. Yarra, and P. K. Ghosh, "Asr inspired syllable stress detection for pronunciation evaluation without using a supervised classifier and syllable level features." in *INTERSPEECH*, 2019, pp. 924–928.

[20] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE Intl. Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[22] P. Boersma, "Praat: doing phonetics by computer," *http://www.praat.org/*, 2006.

[23] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[24] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. Weiss, and Y. Wu, "Parallel tacotron: Non-autoregressive and controllable tts," *arXiv preprint arXiv:2010.11439*, 2020.

[25] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[26] T. e. a. Chen, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," *arXiv preprint arXiv:1512.01274*, 2015.

[27] A. Paleyes, M. Pullin, M. Mahsereci, N. Lawrence, and J. Gonzalez, "Emulation of physical processes with emukit," in *Second Workshop on Machine Learning and the Physical Sciences, NeurIPS*, 2019.

[28] J. Guo, H. He, T. He, L. Lausen, M. Li, H. Lin, X. Shi, C. Wang, J. Xie, S. Zha *et al.*, "Gluoncv and gluonnlp: Deep learning in computer vision and natural language processing." *Journal of Machine Learning Research*, vol. 21, no. 23, pp. 1–7, 2020.

[29] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant *et al.*, "Quantitative analysis of culture using millions of digitized books," *science*, vol. 331, no. 6014, pp. 176–182, 2011.

[30] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *STIN*, vol. 93, p. 27403, 1993.

[31] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[32] G. Zhao, S. Sonsaat, A. O. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-arctic: A non-native english speech corpus," *Perception Sensing Instrumentation Lab*, 2018.

[33] A. Porzuczek and A. Rojczyk, "English word stress in polish learners speech production and metacompetence," *Research in Language*, vol. 15, no. 4, pp. 313–323, 2017.

[34] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *International conference on machine learning*. PMLR, 2018, pp. 3918–3926.

[35] A. E. Hieke, "Linking as a marker of fluent speech," *Language and Speech*, vol. 27, no. 4, pp. 343–354, 1984.

[36] S. Shattuck-Hufnagel, M. Ostendorf, and K. Ross, "Stress shift and early pitch accent placement in lexical items in american english," *Journal of Phonetics*, vol. 22, no. 4, pp. 357–388, 1994.