Interspeech 2022
18-22 September 2022, Incheon, Korea

Automatic Prosody Evaluation of L2 English Read Speech in Reference to Accent Dictionary with Transformer Encoder

Yu Suzuki[1], *Tsuneo Kato*[1], *Akihiro Tamura*[1]

[1]Graduate School of Science and Engineering, Doshisha University.
tsukato@mail.doshisha.ac.jp

Abstract

Automatic prosody evaluation models for second language (L2) read speech are classified into two categories: reference-based and reference-free. Reference-based models refer to native speakers' speech of the uttered text while reference-free models do not. Conventional reference-free models do not even take the uttered text into account. We propose an automatic prosody evaluation model that takes the uttered text into account by estimating native speakers' prosodic patterns using a Transformer encoder. The Transformer encoder used in FastSpeech 2 estimates a sequence of native speakers' prosodic features in a phoneme-segment level, and a subsequent neural network module evaluates an L2 learner's utterance by comparing the sequence of prosodic features with the estimated sequence of native speakers' utterances. We evaluated the model by Spearman's correlation between the objective and subjective scores on L2 English sentence speech read by Japanese university students. The experimental results indicated that our model achieved a higher subjective-objective score correlation than that with a reference-free model and even higher than an inter-rater score correlation.

Index Terms: automatic prosody evaluation, second language (L2) speech, reference-free, Transformer encoder

1. Introduction

Prosody conveys a speaker's emotions, intentions, and attitudes to listeners. For L2 learners, it is important to learn correct prosodic patterns, i.e., accents, intonation, and rhythm, in a neutral emotion because it greatly affects intelligibility of their words. Computer-assisted pronunciation training systems require accurate evaluation and diagnosis of prosody as well as of pronunciation.

Automatic evaluation of overall fluency in L2 spontaneous speech has been actively studied [1, 2, 3, 4, 5, 6, 7]. These studies are based on multiple regression of various speech features, such as word count, speech rate, frequencies of pause, and phoneme posteriogram without reference, and show high subjective-objective score correlation. They usually do not take into account correctness of pronunciation or prosody in reference to an accent dictionary.

On the other hand, more basic educational programs demand automatic prosody evaluation on read speech of isolated words or short sentences. The most basic reference-based approach compares L2 speech with native speakers' speech of the same text [8, 9, 10]. This approach requires collection of the reference speech produced by native speakers in advance. It is not difficult to collect native isolated word utterances, but it is to collect short sentence utterances in advance.

A stress detection approach [11, 12] does not require reference speech produced by native speakers. Comparison of a detected stress pattern with a canonical pronunciation with lexical stress information determines if the stress pattern is correct. However, stress detection becomes difficult for sentence utterances. Another solution was proposed to refer to text-to-speech (TTS) synthetic speech. A study showed that there was no significant difference between using native speakers' speech and synthetic speech as a reference in a binary classification task of L1 and L2 speech [13].

We propose an automatic prosody evaluation model of L2 English read speech in which a Transformer encoder [14] estimates a sequence of native speakers' prosodic features on the basis of a phoneme sequence of the text and a subsequent module grades an L2 learner's speech on the basis of the difference of the measured prosodic features of L2 speech and estimated ones of native speakers' speech. We call this model Transformer-encoder-based L2 English prosody evaluation model (TEBE). To estimate native speakers' prosodic features, we implemented a modified Transformer encoder used in the high-quality TTS engine: FastSpeech 2 [15].

2. Transformer-encoder-based L2 English Prosody Evaluation Model

2.1. Model structure

Figure 1 illustrates the overall architecture of the TEBE model. The model consists of two modules: a native-speaker prosody estimation module (hereafter, native-speaker module) and an L2-learner prosody evaluation module (hereafter, L2-learner module). The native-speaker module takes a phoneme sequence with lexical stress information i.e., primary/secondary/no stress, as an input and outputs an estimated sequence of native speakers' prosodic features in a phoneme segment level. The L2-learner module receives a sequence of L2 learner's prosodic features to assess, the estimated sequence of native speakers' prosodic features for reference, and the same phoneme sequence with lexical stress information as inputs, and outputs an estimated score of subjective evaluation rated by native speakers. Figure 2 shows the structure of the native-speaker module. This module is based on the Transformer encoder of FastSpeech 2 [15]. Figure 3 shows the structure of the L2-learner module. This module estimates the subjective score on the basis of the differences in the phoneme-segment-level prosodic features between the estimated native speakers' and measured L2 learners'. The following subsections describe in detail.

2.1.1. Native-speaker module

As Figure 2 shows, the native-speaker module consists of a phoneme embedding layer, N-stacked feed-forward Transformer (FFT) blocks [16], and M juxtapositional prosodic feature predictors that estimate prosodic features on phoneme duration, fundamental frequency (F0), and intensity.

Let $\boldsymbol{S} = \{s_1, ..., s_l\}, (l < L)$ be a canonical phoneme se-

10.21437/Interspeech.2022-10344

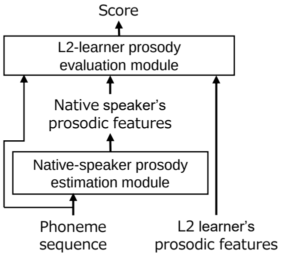
Score
L2-learner prosody evaluation module
Native speaker's prosodic features
Native-speaker prosody estimation module
Phoneme sequence
L2 learner's prosodic features

Figure 1: *Overall structure of TEBE model.*

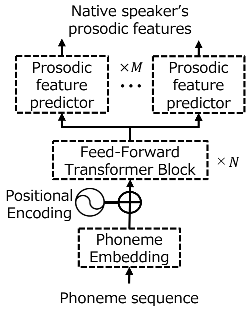
Native speaker's prosodic features
Prosodic feature predictor
×M
Prosodic feature predictor
Feed-Forward Transformer Block
× N
Positional Encoding
Phoneme Embedding
Phoneme sequence

Figure 2: *Native-speaker prosody-estimation module.*

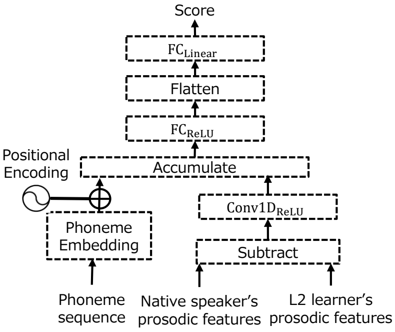

Figure 3: *L2-learner prosody-evaluation module.*

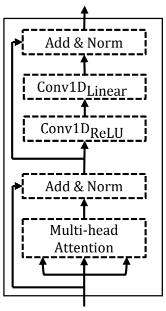

Figure 4: *Feed-forward Transformer block.*

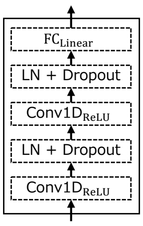

Figure 5: *Prosodic-feature predictor.*

quence with accent information of an L2 learner's read speech. After zero-padding the null phonemes, the phoneme embedding layer converts every phoneme into a d_E-dimensional embedding vector. Then, a positional encoding adds positional embeddings and forms an input matrix to the first FFT block. This process is expressed as follows:

$$\boldsymbol{X}_0 = \mathrm{PhonemeEmbed}(\boldsymbol{S}) + \mathrm{PositionEnc} \tag{1}$$

where $\boldsymbol{X}_0$, an $L \times d_E$ matrix, is an input into the first FFT block.

An FFT block is a modified Transformer encoder block, where two position-wise feed-forward neural networks are substituted with two 1D convolutional layers, one with a rectified linear unit (ReLU) and one with an identity activation function, to capture temporally characteristic patterns across phoneme segments. Figure 4 illustrates the structure of an FFT block. An FFT block collects contextual information with its multi-head self-attention and outputs a matrix of the same size as follows:

$$\boldsymbol{X}_n = \mathrm{FFTblock}^n(\boldsymbol{X}_{n-1}) \qquad 1 \le n \le N \tag{2}$$

where $\boldsymbol{X}_n$ is an output of the n-th FFT block.

Each of M juxtapositional prosodic feature predictors on top of the native-speaker module takes $\boldsymbol{X}_N$ as an input and outputs an estimated sequence of a native speaker's prosodic feature in a phoneme segment level. Figure 5 illustrates the structure of this block. The block consists of two 1D convolutional layers with an ReLU activation function each followed by a linear layer with a dropout and one fully-connected (FC) layer with an identity activation function. This is a modified duration/pitch/energy predictor in FastSpeech 2 [15] without the length regulator.

$$\hat{\boldsymbol{z}}_m^{NS} = \mathrm{ProsodicFeatPredict}^{\mathrm{m}}(\boldsymbol{X}_N) \tag{3}$$

where $\hat{\boldsymbol{z}}_m^{NS}$ denotes an estimated sequence of the m-th prosodic feature with standardization. The concatenation of M sequences of $\hat{\boldsymbol{z}}_m^{NS}$ forms an $L \times M$ matrix $\hat{\boldsymbol{Z}}^{NS} = \{\hat{\boldsymbol{z}}_1^{NS}, ..., \hat{\boldsymbol{z}}_M^{NS}\}$.

The native-speaker module is trained separately from the L2-learner module by minimizing the following mean square error L_f^{NS}.

$$L_f^{NS} = \frac{1}{l}\sum_{k=1}^{l}\sum_{m=1}^{M}\left(\hat{z}_{m,k}^{NS} - z_{m,k}^{NS}\right)^2 \tag{4}$$

where $z_{m,k}^{NS}$ and $\hat{z}_{m,k}^{NS}$ respectively denote the m-th element of the reference and estimated prosodic feature vectors for the k-th phoneme segment of a native speaker's utterance in a training corpus.

2.1.2. L2-learner module

As Figure 3 shows, the L2-learner module consists of a phoneme embedding layer, subtraction layer, 1D convolutional layer with an ReLU activation function, accumulation layer, and two FC layers, one with an ReLU and one with an identity activation function, with a flatten layer between.

The phoneme embedding layer and the subsequent positional encoding layer convert a phoneme sequence $\boldsymbol{S}$ into a matrix $\boldsymbol{X}_0$ with the weights and positional embedding common to those in the native-speaker module. The subtraction layer computes a difference matrix $\boldsymbol{D}$ of $\hat{\boldsymbol{Z}}^{NS}$ and $\boldsymbol{Z}^{L2}$, that is, a sequence of measured prosodic features for an L2 learner's utterance. The 1D convolutional layer extracts d_E-dimensional features by applying d_E filters in the temporal direction to each feature in $\boldsymbol{D}$ and outputs an $L \times M \times d_E$ tensor $\boldsymbol{D}_E$.

$$\boldsymbol{D} = \hat{\boldsymbol{Z}}^{NS} - \boldsymbol{Z}^{L2} \tag{5}$$

$$\boldsymbol{D}_E = \mathrm{Conv1D}(\boldsymbol{D}) \tag{6}$$

The accumulation layer accumulates $\boldsymbol{D}_E$ in the feature direction, adds up $\boldsymbol{X}_0$, and outputs an $L \times d_E$ matrix. The following FC layer with an ReLU activation function receives the output matrix of the accumulation layer and outputs an $L \times d_D$ matrix. The flatten layer transforms the output of the FC layer into a $(d_D L)$-dimensional vector and the final FC layer with an identity activation function outputs an estimated subjective score $\hat{y}$. The process is expressed as follows:

$$\hat{y} = \mathrm{FC}_{\mathrm{Linear}}(\mathrm{Flatten}(\mathrm{FC}_{\mathrm{ReLU}}(\mathrm{Accumulate}(\boldsymbol{D}_{\boldsymbol{E}}, \boldsymbol{X}_0)))) \tag{7}$$

The L2-learner module is trained separately after the training of the native-speaker module by minimizing the squared error L_s^y between the estimation $\hat{y}$ and reference y.

2.2. Prosodic features

The prosodic features that the native-speaker module outputs and the L2-learner module accepts for both reference and test are designed manually on phoneme duration, fundamental frequency (F0), and intensity. All prosodic features are standardized into Z-scores.

2.2.1. Phoneme duration

Phoneme durations are calculated on the basis of phoneme segments obtained using Montreal Forced Aligner [17]. Let dur be a phoneme duration i.e., a time difference between the beginning and ending points of a phoneme segment in seconds. As a prosodic feature, standardized phoneme duration d in a logarithmic scale is obtained with the following equation based on dur.

$$d_l = \log\left(dur + 1\right) \tag{8}$$

$$d = \frac{d_l - \overline{d_l}}{std(d_l)} \tag{9}$$

where $\overline{d_l}$ and $std(d_l)$ denote the mean and standard deviation of d_l. The phoneme segments provide base points in time for obtaining F0 and intensity features, as described as follows.

2.2.2. Fundamental frequency (F0)

An F0 contour is obtained using the F0 estimator Harvest [18] in the speech analysis and synthesis system WORLD[19]. The F0 values extracted at 10-ms intervals are converted to mel-frequency values $f_{0_m}(t)$ then normalized by subtracting the mean of the utterance.

$$f_{0_m}(t) = 1127.0 \times \log\left(\frac{f_0(t)}{700} + 1\right) \tag{10}$$

$$f_{0_n}(t) = f_{0_m}(t) - \overline{f_{0_m}} \tag{11}$$

where $\overline{f_{0_m}}$ and $f_{0_n}(t)$ denote the mean F0 in mel-frequency of an utterance and normalized F0. After standardization with all utterances, three features f_m, f_s, and f_e are extracted as the mean and at the beginning and ending points of each phoneme segment, respectively. In this process, frames in which F0 is not detected are excluded, and f_s, f_e, and f_m are set to 0 if no f_0 is detected in all frames of a segment.

2.2.3. Intensity

A power sequence $i_{dB}(t)$ in decibel is computed by a shift window with length of 25 ms and shift of 10 ms and normalized by subtracting the mean of the utterance. The obtained values are converted to decibel values i_{dB}, and for speaker normalization, the mean per utterance $\overline{i_{dB}}$ is subtracted to obtain the normalized intensity i_n.

$$i_n(t) = i_{dB}(t) - \overline{i_{dB}} \tag{12}$$

where $\overline{i_{dB}}$ and $i_n(t)$ are the mean intensity of an utterance and normalized intensity, respectively. After standardization with all utterances, three features i_m, i_s, and i_e are extracted as the mean and at the beginning and ending points of each phoneme segment, respectively, as with F0.

Table 1: *Statistics of speech corpora.*

	LibriTTS (L1)	ERJ (L2)
#Utterances	154,344	3,654
#Speakers	1,230	80
#Words	2,699,724	25,434
Vocabulary size	41,763	264

3. Experiments

3.1. Speech Corpora

We used two corpora of English read speech, one produced by native speakers of English (LibriTTS [20]) for training and validating the native-speaker module and one produced by Japanese university students (English Read by Japanese, ERJ [21]) for training the L2-learner module and testing the overall model.

We chose the LibriTTS corpus because the native speakers' speech was split at sentence breaks. We used only four clean subsets: dev-clean, test-clean, train-clean-100, and train-clean-360 for training and validating the native-speaker module. Utterances in which Harvest detected no F0 contour were excluded. The clean subsets included 154k utterances produced by 1,230 speakers. The average length of an utterance was 17.5 in words and 65.7 in phonemes. The speech sampled at 24 kHz was downsampled to 16 kHz.

The ERJ corpus included read speech of English words and sentences produced by 160 Japanese university students. We used a subset comprised of 3,654 utterances of 120 short sentences that were designed for assessing sentence stress and rhythm. All utterances were produced in a neutral emotion. The average length of an utterance was 7.0 in words and 26.0 in phonemes. The speech was sampled at 16 kHz.

We conducted a subjective evaluation of the ERJ subset involving native speakers of American English. Every utterance was rated on a five-point discrete scale (5: excellent, 1: very poor) with a focus on prosody i.e., accents and rhythm, by two raters. A total of ten raters listened to the utterances with headphones and scored them on a personal computer. The mean of two scores for each utterance was used as a reference signal for training the L2-learner module. The mean of the inter-rater subjective score correlation was 0.505, which is a target of the subjective-objective score correlation R. The statistics of the two corpora are listed in Table 1.

In the experiments, the clean subsets of LibriTTS corpus were shuffled and split into 8:1:1 for training, development, and test sets, respectively. We used the 3,654 utteances of the ERJ subset in stratified nested 10-fold cross validation, i.e., nested 10-fold cross validation was conducted on the basis of the utterances stratified according to the mean subjective score. The development set was used to determine when to stop the model training.

3.2. Overview

We trained the TEBE model in two steps: training the native-speaker module using the clean subsets of the LibriTTS corpus first, then training the L2-learner module using the ERJ subset. In the first step, we measured the accuracy of native speakers' prosodic feature estimation by using the coefficient of determination R^2. In the second step, we measured subjective-objective score correlation R for the test set of L2 learners' speech.

We evaluated the TEBE model with various combinations

Table 2: *Coefficients of determination R^2 for each prosodic feature estimated from native-speaker module.*

	d	f_s	f_e	f_m	i_s	i_e	i_m
R^2	0.818	0.184	0.166	0.202	0.771	0.760	0.751

of prosodic features to examine which prosodic features are effective. The combinations are listed as follows:

1) **TEBE Dur.** A single feature on phoneme duration d.

2) **TEBE Dur. and F0** Four features on phoneme duration and F0, i.e., d, f_m, f_s, and f_e.

3) **TEBE Dur. and Int.** Four features on phoneme duration and intensity, i.e., d, i_m, i_s, and i_e.

4) **TEBE Dur., F0, and Int.** Seven features on phoneme duration, F0, and intensity, i.e. d, f_m, f_s, f_e, i_m, i_s, and i_e.

We compared R with a regression model without referencing phoneme and accent information.

3.3. Setup

We used ARPABET with accent information, with which the CMU pronunciation dictionary [22] defines canonical phoneme sequences, for phoneme sequence inputs. ARPABET is composed of 39 phonemes with 3 accent symbols 1/2/0 that gives additional primary/secondary/no stress information for vowels. Pauses between words were included in the input and output of the model as well as phonemes because they affect the subjective score.

We set the parameters of the native-speaker module as follows. The maximum length of phoneme segments L was 357, dimensions of phoneme embedding d_E was 64, number of FFT blocks N was 4, heads in the multi-head self-attention was 4, and kernel size and number of channels of the 1D convolutional layers were 3 and 256, respectively. We adopted Adam [23] optimizer for model training with an initial learning rate $\alpha = 0.0001$, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-7}$. We trained the module up to 60,000 steps with a batch size of 512.

We set the parameters of the L2-learner module as follows. The kernel size and number of channels of the 1D convolutional layer were 3 and 64, respectively, and dimensions of phoneme embedding d_D of $\mathrm{FC}_{\mathrm{ReLU}}$ was 64. We also adopted Adam optimizer with the same parameter settings as the native-speaker module. We trained the module up to 3,000 steps with a batch size of 16 .

3.4. Regression model without reference to phoneme sequence

As a baseline model without reference to a phoneme sequence, we trained and evaluated a simple multiple regression model based on various features on prosody and fluency in reference to previous studies [1, 5]. The model was built on ElasticNet [24] with five features used in an early study of estimating fluency [1]. Specifically, the standardized five features are as follows: average chunk length in words, mean deviation of chunks in words, duration of silence per word, mean of silence duration, and articulation rate. The elastic net was trained to estimate the five-point-scale subjective scores. Note that this model was not originally designed to evaluate the prosodic aspect of read speech but to evaluate fluency in spontaneous speech.

Table 3: *Subjective-objective score correlation R for 3,656 L2 English sentence utterances read by Japanese learners.*

Model	R
Regression w/o ref. phoneme sequence	0.376
TEBE Dur.	0.515
TEBE Dur. and F0	0.479
TEBE Dur. and Int.	0.570
TEBE Dur., F0, and Int.	0.540

3.5. Results

Before evaluating the performance of the TEBE model, we evaluated the accuracy of native-speaker module by using R^2. Table 2 lists the R^2 of the native speakers' prosody estimation for each prosodic feature. The results indicate that d had the highest R^2 among all features and the prosodic features on intensity had generally high values, whereas the prosodic features on F0 had very low values.

Table 3 lists the R for the regression model without reference to a phoneme sequence and TEBE model with four combinations of prosodic features. The R of the TEBE model was higher than that of the multiple regression model (0.376) and even higher than the inter-rater score correlation (0.505). Among the combinations of prosodic features in the TEBE models, the phoneme duration gave a high baseline (0.505) and the three intensity features increased R, whereas the three F0 features decreased R. The maximal R was 0.570 for TEBE Dur. and Int. We consider that the decrease in R w.r.t the F0 features is associated with the low R^2 in the native-speaker module and a single F0 feature sequence estimated from the module is insufficient to express the diversity of correct F0 contours.

4. Conclusions

We proposed a Transformer-encoder-based automatic prosody evaluation model of L2 English read speech, called the TEBE model. The model first estimates native speakers' prosodic features on the basis of a phoneme sequence with a Transformer encoder and evaluates L2 learner's read speech on the basis of the comparison of the estimated native speakers' and measured L2 learner's prosodic features. The TEBE model achieved a subjective-objective score correlation of 0.570, which surpassed that from a regression model without reference to phoneme sequences (0.376) and even an inter-rater score correlation (0.505). The experimental results also indicated that the subjective-objective score correlation decreased with prosodic features on F0. They also suggest that a single estimation of a F0 feature sequence is insufficient to evaluate L2 learners prosody with diversity of correct F0 contours. Automatic evaluation on the basis of multiple estimated sequences of F0 features will be for future work.

5. Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 20K00789.

6. References

[1] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, “Automatic scoring of non-native spontaneous speech in tests of spoken english,” *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[2] M. Black, D. Bone, Z. Skordilis, R. Gupta, W. Xia, P. Papadopoulos, S. Chakravarthula, B. Xiao, M. Van Segbroeck, J. Kim, P. Georgiou, and S. Narayanan, “Automated evaluation of non-native english pronunciation quality: Combining knowledge- and data-driven features at multiple time scales,” in *Proc. Interspeech 2015*, 2015, pp. 493–497.

[3] R. C. van Dalen, K. M. Knill, and M. J. F. Gales, “Automatically grading learners’ english using a gaussian process,” in *ISCA International Workshop on Speech and Language Technology in Education, SLaTE 2015, Leipzig, Germany, September 4-5, 2015*, 2015, pp. 7–12.

[4] Y. Wang, M. J. F. Gales, K. M. Knill, K. Kyriakopoulos, A. Malinin, R. C. van Dalen, and M. Rashid, “Towards automatic assessment of spontaneous spoken english,” *Speech Communication*, vol. 104, pp. 47–56, 2018.

[5] Y. Shen, A. Yasukagawa, D. Saito, N. Minematsu, and K. Saito, “Optimized prediction of fluency of L2 english based on interpretable network using quantity of phonation and quality of pronunciation,” in *IEEE Spoken Language Technology Workshop, SLT 2021*, 2021, pp. 698–704.

[6] L. Fontan, M. Le Coz, and S. Detey, “Automatically measuring l2 speech fluency without the need of asr: a proof-of-concept study with japanese leaners of french,” in *Proc. Interspeech 2018*, 2018, pp. 2544–2548.

[7] M. Kondo, L. Fontan, M. Le Coz, T. Konishi, and S. Detey, “Phonetic fluency of japanese learners of english: Automatic vs native and non-native assessment,” in *Proc. ISCA International Conference on Speech Prosody 2020*, 2020, pp. 784–788.

[8] J. Arias, N. Yoma, and H. Vivanco, “Automatic intonation assessment for computer aided language learning,” *Speech Communication*, vol. 52, pp. 254–267, 2010.

[9] J. Cheng, “Automatic assessment of prosody in high-stakes english tests,” in *Proc. Interspeech 2011*, 2011, pp. 1589–1592.

[10] Q. T. Truong, T. Kato, and S. Yamamoto, “Automatic assessment of l2 english word prosody using weighted distances of f0 and intensity contours,” in *Proc. Interspeech 2018*, 2018, pp. 2186–2190.

[11] 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. 32–45, 2015.

[12] 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.

[13] Y. Xiao and F. Soong, “Proficiency assessment of esl learner’s sentence prosody with tts synthesized voice as reference,” in *Proc. Interspeech 2017*, 2017, pp. 1755–1759.

[14] 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*, vol. 30, 2017, p. 6000–6010.

[15] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” in *9th International Conference on Learning Representations, ICLR 2021*, 2021. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA

[16] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, “Fastspeech: Fast, robust and controllable text to speech,” in *Advances in Neural Information Processing Systems*, 2019, pp. 3165–3174.

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

[18] M. Morise, “Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals,” in *Proc. Interspeech 2017*, 2017, pp. 2321–2325.

[19] M. Morise, F. Yokomori, and K. Ozawa, “World: A vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE Transactions on Information and Systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[20] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech,” in *Proc. Interspeech 2019*, 2019, pp. 1526–1530.

[21] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, “English speech database read by japanese learners for call system development,” in *Proc. Language Resources and Evaluation 2002*, 2002, pp. 896–903.

[22] “The carnegie mellon pronouncing dictionary,” http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[23] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[24] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society, Series B*, vol. 67, pp. 301–320, 2005.