Article

Not peer-reviewed version

AI Enhancements for Linguistic E-learning System

Jueting Liu , Sicheng Li , Chang Ren , Yibo Lyu , Tingting Xu , Zehua Wang , Wei Chen *

Posted Date: 20 September 2023

doi: 10.20944/preprints202309.1339.v1

Keywords: linguistic E-learning; phonetic transcription; mel frequency cepstrum coefficient; grapheme-to-phoneme; transformer; speech synthesis

Article

AI Enhancements for Linguistic E-learning System

Jueting Liu [1], Sicheng Li [2], Chang Ren [3], Yibo Lyu [4], Tingting Xu [5], Zehua Wang [6] and Wei Chen [7,*]

[1] China University of Mining and Technology; 6476@cumt.edu.cn
[2] Auburn University; szl0072@auburn.edu
[3] Auburn University; czr0072@auburn.edu
[4] Auburn University; laolvyenq@gmail.com
[5] China University of Mining and Technology; tingting_xu@cumt.edu.cn
[6] China University of Mining and Technology; zwang@ece.ubc.ca
[7] China University of Mining and Technology; chenwdavior@163.com
* Correspondence: chenwdavior@163.com

Abstract: The E-learning system has achieved great development after the pandemic. In this work, we proposed three artificial intelligence-based enhancements to our linguistic interactive E-learning system from different aspects. Compared with the original phonetic transcription exam system, our enhancements include an MFCC+CNN-based disordered speech classification module, a Transformer-based Grapheme-to-Phoneme converter, and a Tacotron2-based IPA-to-Speech speech synthesis system. This work not only provides a better experience for the users of this system but also explores the utilization of artificial intelligence technologies in the E-learning field and linguistic field.

Keywords: linguistic E-learning; phonetic transcription; mel frequency cepstrum coefficient; grapheme-to-phoneme; transformer; speech synthesis

1. Introduction

Phonetic Transcription, a process that represents speech sounds by special symbols, plays an important role in the linguistic education field. Generally, the International Phonetic Alphabet(IPA) characters are utilized in the process of phonetic transcription [1,2]. The IPA is an alphabet system generated from Latin script that aims to indicate the pronunciation of words, for example, the phonetic format of /Phonetic/ is /fəˈnɛtik/.

In the previous work, we developed an interactive E-learning system focused on the **phonetic transcription** and **pronunciation** for language learners. This system named **APTgt** is an online exam system that provides phonetic transcription exams for IPA language students and automated grading tools for teachers [3]. To improve the intelligence and extensibility of the original system, in this work, we propose three enhancements for the system based on machine learning and deep learning technology. Figure 1 illustrates the function of our original system and the proposed enhancements. The primary system includes two parts, in the teacher's part a teacher can create a question by attaching an audio file of word/phrase pronunciation and also uploading its corresponding phonetic format as the answer. On the other hand, a student will listen to the questions and type the answers on an IPA keyboard. The system will then automatically calculate the similarity between the student's answers with the pre-stored correct answer by the Edit distance algorithm and generate the grade [3,4]. The enhancements we propose include three parts:

- A MFCC+CNN based disordered speech classification module.
- A Transformer-based grapheme-to-phoneme(G2P) converter module.
- A Tacotron2-based IPA-to-Speech speech synthesis module.

The disordered speech classification module aims to provide a function to distinguish disordered speech and non-disordered speech. As an intelligent linguistic E-learning system, this module will help students understand the difference between correct pronunciation and disordered speech. During

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2023 doi:10.20944/preprints202309.1339.v1

this work, we employ the Mel Frequency Cepstrum Coefficient (MFCC) as the feature to represent the speech sound and the convolutional neural network (CNN) as the classification model [5].

The second enhancement module to our E-learning system is a transformer-based grapheme-to-phoneme(G2P) converter. In the phonetic transcription exam, teachers need to upload the pronunciation of English words/phrases/sentences, the pronunciation of speech sounds can be represented as phonemes. A G2P conversion is a process that converts the written words(grapheme) to their pronunciations(phonemes) [6]. With this G2P converter, teachers can easily extract the words in their phonetic format. We utilize neural machine translation ideas to design this G2P converter. After comparing several models, the transformer model, an encoder-decoder model with self-attention mechanisms can provide superb performance with a low word error rate(WER) and phoneme error rate(PER) [7].

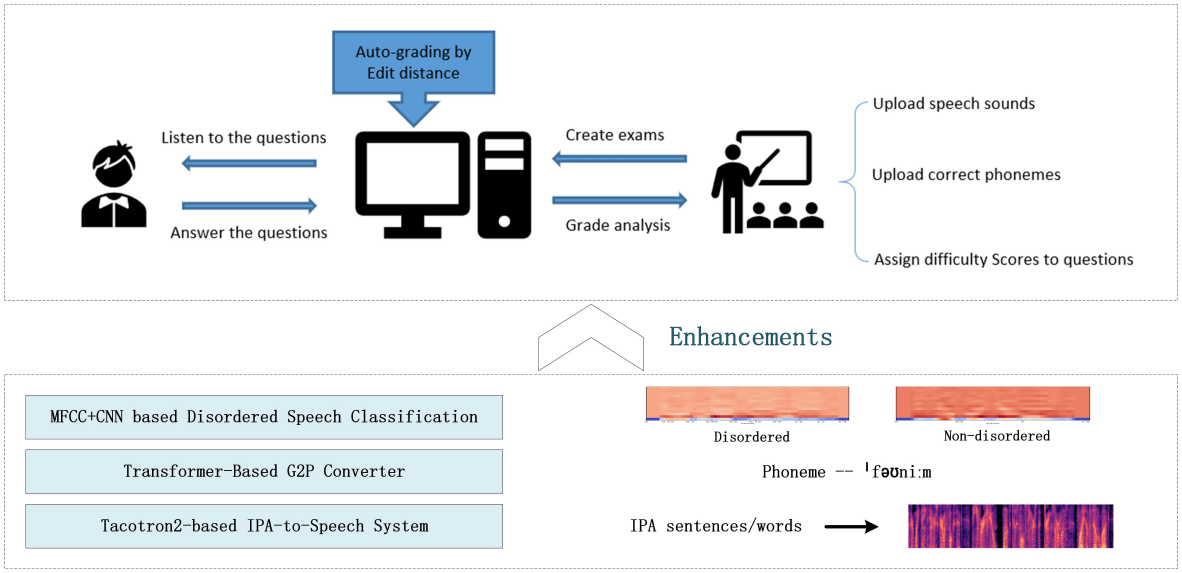

Figure 1. The enhancements to the linguistic E-learning system.

The Tacotron2-based IPA-to-Speech speech synthesis(text-to-speech) system is the last enhancement to our system. In order to help students better understand the pronunciation of IPA symbols, this module is introduced for directly generating high-quality speech audio files from IPA symbols. Furthermore, this module will help teachers easily acquire audio files as a part of the question in the exam system. It takes two steps to build the IPA-to-Speech system:

- Build a Grapheme-to-Phoneme system to convert all the English text to IPA format.
- Build the TTS system with the processed data.

2. Related work

2.1. Linguistic E-Learning

During the peak of the COVID-19 pandemic, according to data from UNESCO, over 1 billion children were affected and out of the classroom globally. The shift to E-learning or online learning is significantly increasing in the Internet access area. For example, Zhejiang University deployed over 5,000 courses online in two weeks to the platform 'DingTalk ZJU' [8] and the Imperial College London started offering courses on Coursera from 2020.

E-learning is an approach that delivers knowledge or skills remotely and interactively by electrical devices such as smartphones, tablets and laptops. Compared with traditional classroom learning, E-learning can offer students flexible topics or subjects, the interactions with teachers or professors by email or platforms without the restrictions by physical distance. It cannot entirely replace traditional

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2023 doi:10.20944/preprints202309.1339.v1

classroom learning but provides an augmented learning environment. Coursera, and Udemy are all successful in E-learning(Online Learning) platforms in the market that provide high-quality courses with quizzes and interactive exams [9].

Linguistics is a scientific subject of human language. E-learning can play an important role in linguistic education since the advantages of E-learning ideally benefit linguistics pedagogy. Automated Phonetic Transcription - the grading tool(APTgt) is a well-designed interactive web-based E-learning system that focuses on phonetic transcription for students(learners) and teachers(faculty). The phonetic transcription is a process that represents the speech sounds by special characters or symbols [3].

2.2. Speech Disorders Classification

A speech disorder is a condition in which a person has problems creating or forming the speech sounds needed to communicate with others. It is a sub-problem of speech classification. To solve the speech classification problems, both feature extraction function and classification algorithm are required. There are two major features in speech classification/recognition subjects: the linear prediction coding (LPC) and the Mel Frequency Cepstrum Coefficient (MFCC) [10]. The classification algorithms include dynamic time warping (DTW) [11], Hidden Markov Models (HMM) [12] and deep learning-based classification.

2.3. Grapheme-to-Phoneme Conversion

In linguistics, a grapheme is the smallest unit of a written language, while a phoneme is the smallest unit of speech sound. A grapheme-to-phoneme(G2P) conversion is a process that converts a spelled-out word to its phonetic format(a sequence of IPA symbols). [13] G2P plays an essential role in the natural language processing(NLP) field including Text-to-Speech(TTS) systems and Automated Speech Recognition(ASR) systems. Generally, the International Phonetic Alphabets(IPA) characters are employed to represent the phoneme.

G2P conversion has always been a popular top in the NLP field. We have investigated different approaches for G2P conversion. The phoneme error rate(PER) and word error rate(WER) can be utilized to evaluate the performance of the G2P conversion system. In 2005, the Hidden Markov Model was employed for G2P conversion by Paul Taylor with 9.02% PER and 42.69% WER [14]. In 2008, Maximilian Bisani introduced Joint-sequence models for G2P conversion, Joint-sequence models are theoretically stringent probabilistic framework that is applicable to this problem. On different English data sets, the joint-sequence models provide better performance compared with the Hidden Markov Models, for example, the PER on CMUdict is 5.88% and the WER on CMUdict is 24.53% [15]. With the development of neural network technology, deep learning models play important roles in NLP field. The G2P conversion, as a text-to-text task, got investigated and trained by different deep-learning models. In 2015, the Seq2Seq model was employed for G2P conversion by Kaisheng Yao from Microsoft Research, in this work, the PER on CMUDict is 5.45% while the WER is 23.55% [16]. In the same year, the Long Short-Term Memory recurrent neural networks were utilized for the same task with 9.1% PER and 21.3% by Kanishka Rao [17]. In 2020, with the start-of-the-art model, the Transformer, the PER increased to 5.23% and the WER is 22.1% [18].

2.4. Speech Synthesis System

Speech synthesis, also known as Text-to-Speech(TTS) is a process that generates human speech sounds from text. Speech synthesis has been a hot topic since the later 20th century. The early computer-based speech synthesis approaches includes **Articulatory Synthesis**, **Formant Synthesis**, **Concatenative Synthesis** and **Statistical Parametric Synthesis** [19].

With the development of neural network technology, deep learning-based end-to-end speech synthesis models are proposed and become major methods in TTS research. A modern TTS system usually consists of three basic components: a text analysis module, an acoustic model, and a vocoder. As shown in Figure 2, the text analysis module converts a text sequence into linguistic features, the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2023 doi:10.20944/preprints202309.1339.v1

acoustic models generate acoustic features from linguistic features, and then the vocoders synthesize waveform from acoustic features. **Tacotron 1, Tacotron 2, Deep Voice, Fast Speech** are all end-to-end TTS examples [20–23].

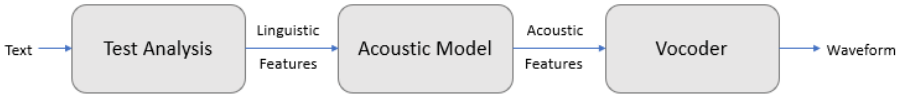

Figure 2. The structure of end-to-end TTS system.

3. The MFCC+CNN based disordered speech classification

The speech classification can be divided into two sub-questions: feature extraction and classification. In this work, we choose MFCC in image format to represent the feature of human speech and the CNN model to achieve classification function.

3.1. Feature Extraction

In sound processing, the Mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. The following Figure 3 illustrates the steps to generate MFCCs from audio [24].

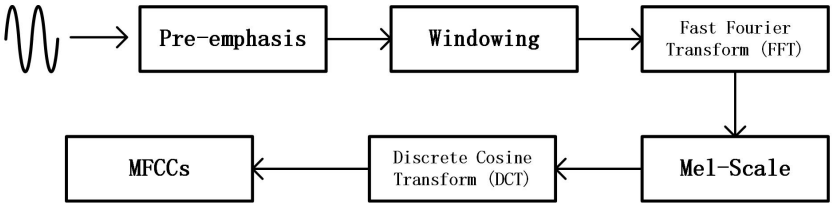

Figure 3. MFCCs extraction process.

- Pre-emphasis the audio signal to increase to energy of the signal at a higher frequency.
- Break the sound signal into the overlapping window.
- Take the Fourier transform to transfer the signal from the time domain to the frequency domain.
- Compute the Mel spectrum by passing the Fourier-transformed signal through the Mel-filter bank. The transformation from the Hertz scale to the Mel scale is:

$$Mel(f) = 2595log(1 + \frac{f}{700})$$

- Take the discrete cosine transform of the mel log signals and the result of this conversion is MFCCs.

3.2. Data selection

The Speech Exemplar and Evaluation Database (SEED) dataset was utilized to train our classification model. The SEED contains about 16,000 recorded speech samples, grouped by age (child vs. adult) and speech health status(with or without speech disorder). The child's speech disorders were determined by parent reports and standardized assessments. Speakers in the SEED are between the ages of 2 to 85, A significant aspect of SEED is that it provides **samples with speech disorders** and **samples without speech disorders** [25].

3.3. Implementation and Evaluation

About 1,000 samples from SEED were selected, 80% of them were used for training and the rest were utilized for validation. We used the Python Librosa library to process the MFCC values into their

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2023 doi:10.20944/preprints202309.1339.v1

image format. The following Figure 4 shows two MFCC images with the same content but recorded by different recorders (With speech disorder and without speech disorder).

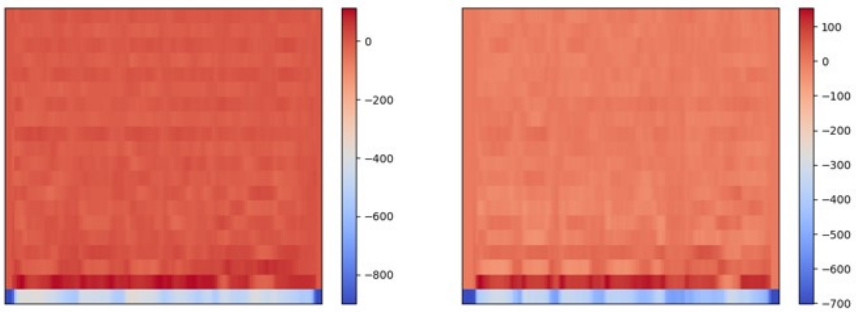

Figure 4. The MFCC images of Disordered Speech and Non-disordered Speech.

Thus, this disordered speech classification problem can be transformed into an image classification problem. We then employed the CNN model to build the classification module. Convolutional Neural Network (ConvNet/CNN) is a deep learning algorithm widely used for image classification and computer vision tasks. The following Table 1 shows the structure of the CNN model used in our classification module.

Table 1. CNN model utilized in classification module.

Layer	Output shape	Param Number
conv2d	(None, 148, 148, 32)	896
max_pooling2d	(None, 74, 74, 32)	0
conv2d_1	(None, 72, 72, 64)	18496
max_pooling2d_1	(None, 36, 36, 64)	0
conv2d_2	(None, 34, 34, 128)	73856
max_pooling2d_2	(None, 17, 17, 128)	0
conv2d_3	(None, 15, 15, 128)	147584
max_pooling2d_3	(None, 7, 7, 128)	0
flatten	(None, 6272)	0
dense	(None, 512)	3211776
dense_1	(None, 1)	513

It takes 150 epochs for training, the following Figure 5 shows the loss and accuracy of our model. The average classification accuracy is about 83%, this disorder speech classification module can neglect the contents and the recorders of the speech which means it is quite efficient and extensive.

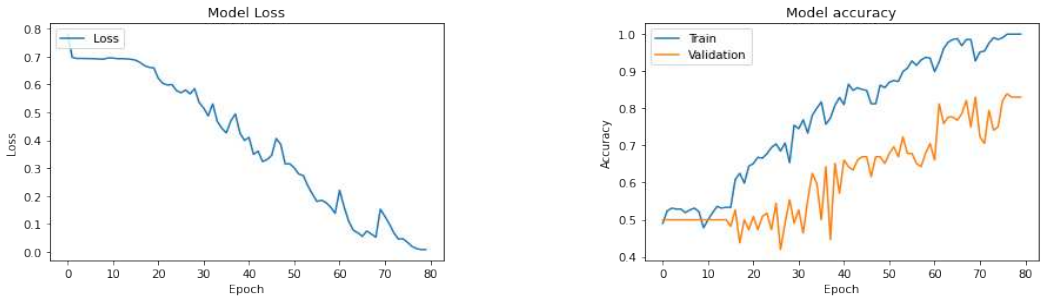

Figure 5. The Loss and Accuracy for Speech Classification Module.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2023 doi:10.20944/preprints202309.1339.v1

4. The Transformer-based Multilingual G2P converter

As we discussed above, the core function of our E-learning system is the interactive phonetic transcription exams. The question in this exam consists of audio of a word/phrase and its corresponding pronunciation(presented in IPA format). The teacher needs to pre-input the correct answer to the system to activate the auto-grading module, so, generating the IPA characters from written language can be a challenge for teachers, it also becomes the inspiration for building the G2P converter. The grapheme-to-phoneme(G2P) conversion is a process of generating words in their IPA format from written format. The G2P converter can be regarded as a variant of a machine translator [5].

4.1. Data Selection

There are two different kinds of characters to represent the pronunciation of words/phrases: the CMUDict characters and the IPA characters. The Carnegie Mellon University Pronouncing Dictionary is an open-source machine-readable pronunciation dictionary for North American English that contains over 134,000 words and their pronunciations, on the other hand, the IPA symbols are more widely used in multi-languages. The following Table 2 illustrates some samples of CUMdict symbols and IPA symbols:

Table 2. CMUDict and IPA symbols.

Written Format	CMUDict	IPA symbols
eat	IY T	it
confirm	K AH N F ER M	kənˈfɜrm
minute	M IH N AH T	ˈminət
quick	K W IH K	kwik
maker	M EY K ER	ˈmeɪker
relate	R IH L EY T	rɪˈleɪt

In our system, we select IPA symbols as the representation of pronunciation. Furthermore, to build a multilingual G2P system, we also investigate the French-IPA converter and Spanish-IPA converter. Table 3 are all the datasets we employed during this work. The first two English-IPA datasets are used to investigate how the size of the data will influence the G2P systems' performance, and the French-IPA and Spanish-IPA datasets are utilized to inspect the feasibility of the multi-linguistic.

Table 3. Datasets for Training.

Dataset	Number of pairs of words	For validation
English-IPA	125,912	20%
French-IPA	122,986	20%
Spanish-IPA	99,315	20%

4.2. The Transformer-based G2P converter

The Transformer model is an encoder-decoder model with attention mechanism. Without using any recurrent layers, the self-attention mechanism allows the model to process the input text as a whole rather than word by word/character by character. This structure makes the Transformer model avoid long dependency issues. The encoder in Transformer is composed of two major elements: the self-attention mechanism (multi-head attention) and feed-forward layer, the decoder includes two multi-head attention layers and one feed-forward layer. The encoder maps input sequences/words into attention-based representations while the decoder then takes the continuous representations and generates the output. The following Figure 6 shows the structure of the attention mechanism in the Transformer.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2023 doi:10.20944/preprints202309.1339.v1

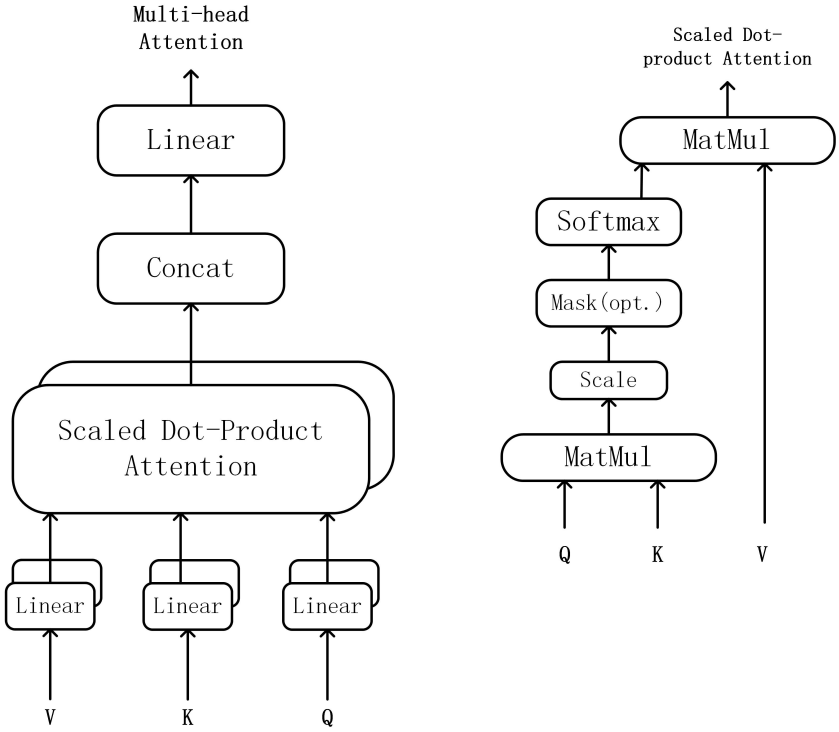

Figure 6. Scaled dot-product attention and multi-head attention in Transformer model.

The Scaled Dot-Product Attention mechanism means the dot products are scaled down by $\sqrt{d_k}$.**Query Q** represents a vector word, **keys K** are all other words in the sequence, and **value V** illustrates the vector of the word. The attention function can be represented as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Multi-head attention is a module that runs through an attention mechanism multiple times in parallel, concatenates the results and produces the result. Each head of the multi-head attention extracts the specific representation, which allows the whole model to receive information from different subspaces. For multi-head attention:

$$multihead(Q, K, V) = concat(head_1, head_2, ..., head_n)W_0$$

$$wherehead_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

W_i^Q, W_i^K, W_i^V are the respective weight matrices calculated from Q, K, and V [26].

4.3. Implementation and Evaluation

Multiple pieces of training have been implemented with the Nvdia Tesla P100 graphic card. We employed a six layers Transformer model and Adam optimizer in Keras with a learning rate of 0.0001. The phoneme error rate(PER) and word error rate(WER) are utilized for evaluating the performance of our G2P converter.The PER is the distance between two phonetic words calculated by the edit distance divided by the total number of phonemes while the WER is a standard parameter for measuring the accuracy in the ASR system. The formulation of WER is:

$$WER = \frac{S + D + I}{N}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2023 doi:10.20944/preprints202309.1339.v1

Where S is the number of substitutions, D is the number of deletions, I is the number of insertions and N refers to the total number of words [7].

The Figures 7–9 display the performance of our multilingual G2P converter. For the English G2P converter, it took 220 epochs of training, the PER is about 2.6% and the WER is 10.7%. For the French G2P converter and Spanish G2P converter, 190 epochs of training have been taken, the PER and WER for the French-IPA converter are 2.1% and 12.3% while the PER and WER for the Spanish-IPA converter are 1.7% and 12.7%. Compared with the other models discussed in **section2**, our models have out performance in converting accuracy. Table 4 illustrates the comparison between different G2P converters.

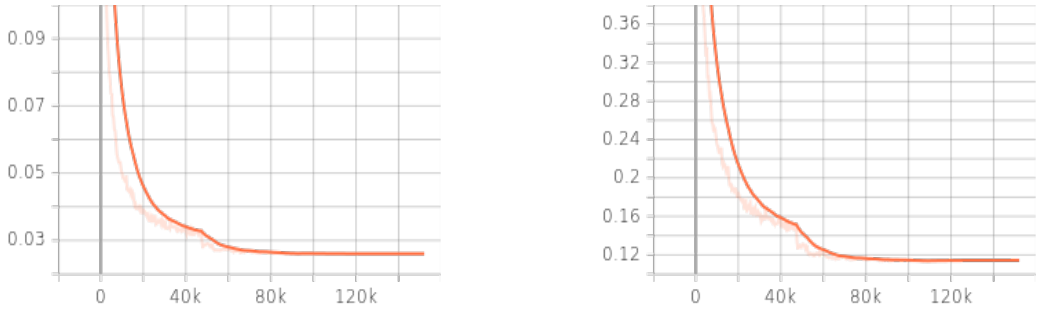

Figure 7. The PER and WER for English-IPA converter.

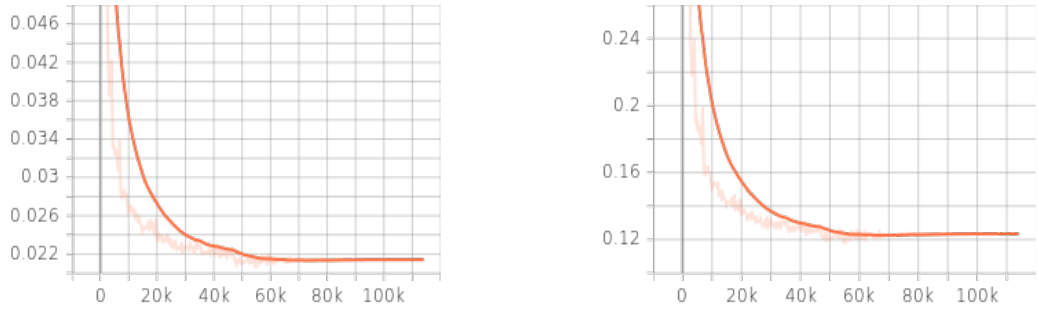

Figure 8. The PER and WER for French-IPA converter.

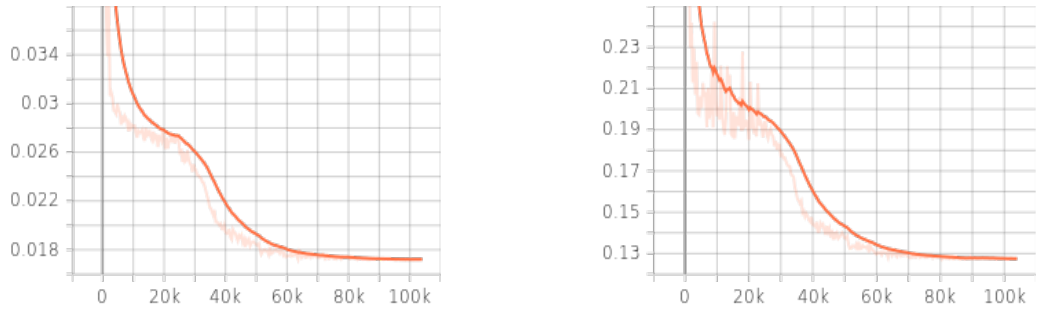

Figure 9. The PER and WER for Spanish-IPA converter.

Table 4. The different models utilized in English G2P conversion.

Models	PER	WER
Hidden Markov Model	9.02%	42.69%
Joint-sequence	5.88%	24.53%
Seq2Seq	5.45%	23.55%
LSTM	9.1%	21.3%
Transformer(ours)	**2.6%**	**10.7%**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2023 doi:10.20944/preprints202309.1339.v1

The following Table 5 shows the results of our multilingual G2P converter.

Table 5. The results of the multilingual G2P converter.

Language	Written Format	Correct Phonemes	Generated Phonemes
English	displeasure	dɪsplˈɛʒə	dɪsplˈʒə
	buoyant	bˈɔɪənt	bˈɔɪənt
	immortal	ɪmˈɔːtəl	ɪmˈɔːtəl
Spanish	ababillarais	aβaβiʎaɾis	aβaβiʎaɾis
	cacofónicos	kakoˈfonikos	kakoˈfonikos
	cadañega	kaðaeɣa	kaðaɲeɣa
French	câlineriez	kalinəʁje	kalinəʁje
	damasquiner	damaskine	damaskine
	effrangé	efʁɑ̃ʒe	efʁɑ̃ʒe

5. The Tacotron2-based IPA-to-Speech System

As we mentioned before, the questions in the phonetic transcription exams consist of speech audios and IPA symbols. From the teachers' views, searching and acquiring appropriate speech audios with high quality along with their texts is not simple work. Even more, the texts of the audio are required to convert to IPA formats. From this perspective, we proposed to design a Text-To-Speech(TTS) system that can directly generate speech sounds from words/phrases/sentences in IPA formats [19].

Figure 10 shows the main process to build our IPA-to-Speech system. The English sentences in LJSpeech will be first converted to their IPA format in batches by the G2P converter. The format of the data in the LJSpeech dataset will be transformed to <*IPASentence, Speechsamples*>. The Mel spectrograms will be predicted and calculated by the Tacotron 2 and we employed WaveGlow as the Vocoder. The Vocoder can generate high-quality speech sounds from the Mel spectrograms.

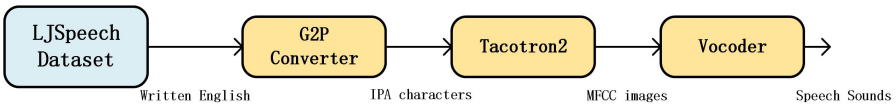

Figure 10. The main process of the IPA-to-Speech system.

5.1. Data preprocess

Speech synthesis, also known as text-to-speech(TTS), is a comprehensive technology that involves linguistics, digital signal processing and acoustics. The main task for TTS is to convert text into speech sounds. We employ the LJSpeech dataset to build our IPA-to-Speech system. The LJSpeech data is a public domain speech dataset consisting of 13,100 speech audio of a single speaker. Clips vary in length from 1 to 10 seconds and have a total length of approximately 24 hours. The data of LJSpeech is constructed by pairs of <English Sentences, Speech Samples>. To build the IPA-to-Speech system, all the written English sentences in LJSpeech dataset should be converted to their IPA format, this data preprocess step requires our English G2P converter discussed in **section 4**. The Table 6 gives several converted samples from the LJSpeech dataset [27].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2023 doi:10.20944/preprints202309.1339.v1

Table 6. The original text in LJSpeech dataset and the converted text.

Original text in LJSpeech	Converted text
The overwhelming majority of people in this country know how to sift the wheat from the chaff in what they hear and what they read.	ðə oʊvɜwɛlmɪŋ mədʒɔrəti ʌv piːpəl iŋ ðɪs kʌntri noʊ haʊ tuː sɪft ðə wiːt frʌm ðə tʃaef ɪn wʌt ðeɪ hiːr ənd wʌt ðeɪ rɛd.
All the committee could do in this respect was to throw the responsibility on others.	ɔl ðə kəmɪti kʊd duː ɪn ðɪs rɪspɛkt wɑːz tuː θroʊ ðə riːspɑːnsəbɪləti ɑːn ʌðɜz.
since these agencies are already obliged constantly to evaluate the activities of such groups	sɪns ðiːz eɪdʒəsiːz ɑːr ɔlrɛdi əblaɪdʒd kɑːnstəntli tuː ɪvaeljuːeɪt ðə aektɪvɪtiːz ʌv sʌtʃ gruːps.

5.2. Tacotron2-based IPA-to-Speech system

Tacotron is an end-to-end text-to-speech synthesis system that synthesizes speech from characters introduced by the Google team. The input of the Tacotron model is characters and the output of the model is corresponding raw spectrograms. The defect of the Tacotron model is the vocoder part, the sound generated by the Griffin-Lim algorithm can't keep in high quality. Thus in this work, we employ the Tacotron2 model to build our IPA-to-Speech system. Compared with the original Tacotron model, the Tacotron 2 uses simpler building blocks, using vanilla LSTM and convolutional layers in the encoder and decoder instead of CBHG stacks and GRU recurrent layers. It consists of two components:

- A recurrent sequence-to-sequence feature prediction network with attention which predicts a sequence of Mel spectrogram frames from an input character sequence
- A modified version of WaveNet which generates time-domain waveform samples conditioned on the predicted Mel spectrogram frames

The following Figure 11 illustrates the structure of the Tacotron2 model [21].

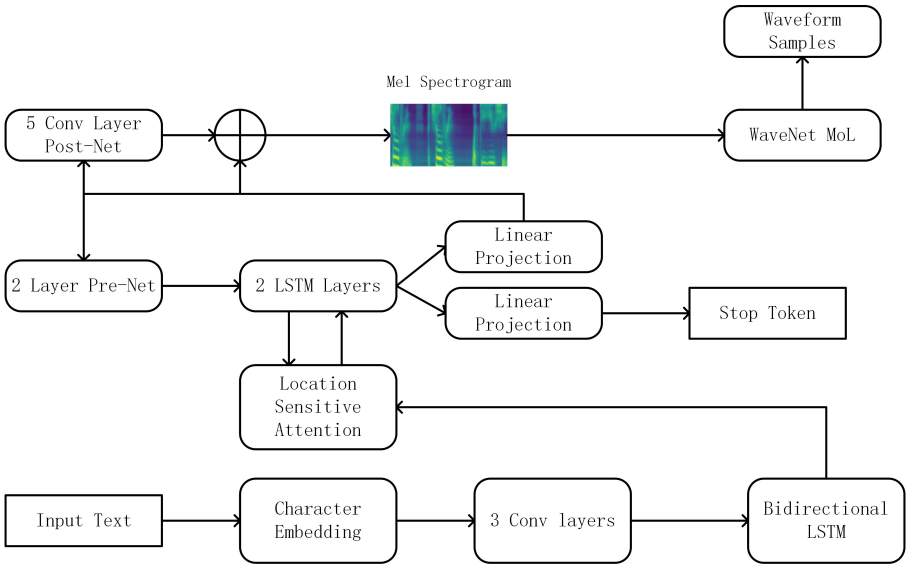

Figure 11. The structure of the Tacotron2 TTS model.

5.3. Implementation and Evaluation

The Nvidia Tesla P100 GPU was employed to train the Tacotron2 model. We selected the batch size of 48 on a single GPU with a 0.0001 learning rate. All the audios in the LJSpeech were used for training which contains about 24.6 hours of audio recorded by a female. Every text in the dataset

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2023 doi:10.20944/preprints202309.1339.v1

needs to be spelled out, for example, the number ”10” should be represented as ”ten”. It takes about 20 hours to finish 180,000 steps which means about 200 epochs. The following Figures 12 and 13 show the training loss, validation loss, target mel, and predicted mel.

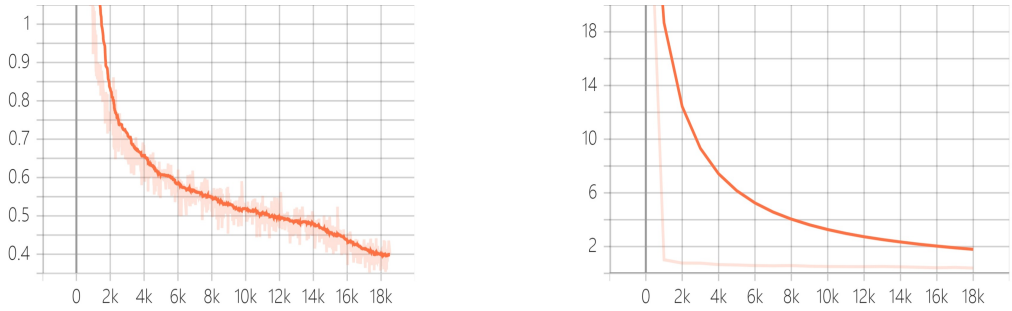

Figure 12. The training loss and validation loss of our Tacotron2 model.

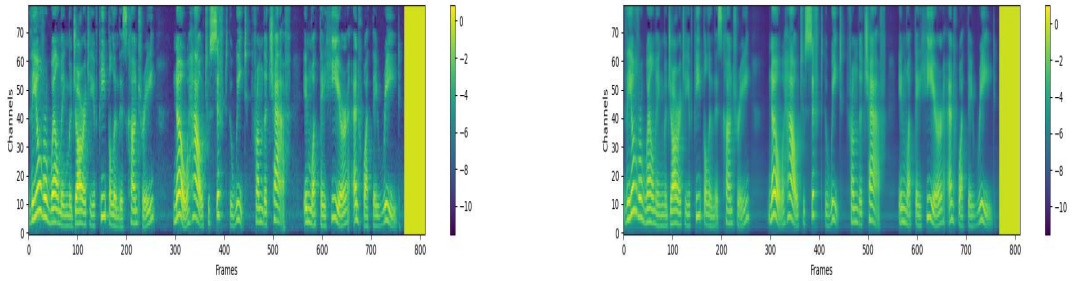

Figure 13. The target mel and predicted mel.

Mean Opinion Score(MOS) is utilized to evaluate the performance of our IPA-to-Speech system, the MOS is a widely-used metric to measure the quality of speech generated by the TTS system. Generally, the MOS score is a rating from 1 to 5 which refers to the perceived quality of audio from worst to best. After rating by human subjects, the MOS is calculated as the arithmetic mean:

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N}$$

where R_n is the single rate score and N is the total number of participants.

We employed 20 students from the linguistic department to help us evaluate our system by MOS, the mean opinion score of our IPA-to-Speech system is about 4.05 [28].

6. Conclusion and Future work

In this work, we proposed three artificial intelligence enhancements for our linguistic E-learning system. The disordered speech classification module utilizes the MFCC to represent the features of the speech and the CNN model to build the classification function which achieves the classification of disordered speech and non-disordered speech; the Grapheme-to-Phoneme module uses the Transformer model that provides high-accuracy G2P conversion; the IPA-to-Speech module employs the Tacotron2 model and generate high-quality speech sound from IPA characters.

All of these enhancements aim to improve the functionality of the system, with this work, our system will not only provide text-based phonetic transcription exams but also has the ability to help users understand speech disorders and the relationship between IPA characters and their pronunciations. In the future, we will first continue improving the performance of the IPA-to-Speech module and employ more students for testing; we will then figure out other potential novelties of our system.

Author Contributions: Conceptualization, Jueting Liu; methodology, Yibo Lyu; software, Sicheng Li and Chang Ren.; validation, Tingting Xu; formal analysis, Tingting Xu.; investigation, Jueting Liu.; writing—original draft preparation, Jueting Liu; writing—review and editing, Zehua Wang and Wei Chen; supervision, Wei Chen; All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the Fundamental Research Funds for the Central Universities, the grant number is 2023QN1079.

References

1. Brown, Adam. International phonetic alphabet. *The encyclopedia of applied linguistics* **2012**
2. Howard, Sara J and Heselwood, Barry C. Learning and teaching phonetic transcription for clinical purposes. *Clinical Linguistics & Phonetics* **2002**, *16*, 371-401.
3. Seals, Cheryl D., et al. "Applied webservices platform supported through modified edit distance algorithm: automated phonetic transcription grading tool (APTgt)." Learning and Collaboration Technologies. Designing, Developing and Deploying Learning Experiences: 7th International Conference, LCT 2020, Held as Part of the 22nd HCI International Conference, HCII 2020.
4. Liu, Jueting, et al. "Optimization to automated phonetic transcription grading tool (APTgt)–automatic exam generator." International Conference on Human-Computer Interaction. Cham: Springer International Publishing, 2021.
5. Liu, Jueting, et al. "Transformer-Based Multilingual G2P Converter for E-Learning System." International Conference on Human-Computer Interaction. Cham: Springer International Publishing, 2022.
6. Liu, Jueting, et al. "Speech Disorders Classification by CNN in Phonetic E-Learning System." International Conference on Human-Computer Interaction. Cham: Springer International Publishing, 2022.
7. Schwarz, Petr, Pavel Matějka, and Jan Černocký. "Towards lower error rates in phoneme recognition." International Conference on Text, Speech and Dialogue. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.
8. Wu, Xiaoping and Fitzgerald, Richard. Reaching for the stars: DingTalk and the Multi-platform creativity of a 'one-star'campaign on Chinese social media. *Discourse, Context & Media* **2021**, *44*, 100540.
9. Downes, Stephen. E-learning 2.0. *ELearn* **2005**, *10*, 1.
10. Madan, Akansha and Gupta, Divya. Speech feature extraction and classification: A comparative review. *International Journal of computer applications* **2014**, *90*.
11. Mohan, Bhadragiri Jagan. "Speech recognition using MFCC and DTW." 2014 international conference on advances in electrical engineering (ICAEE). IEEE, 2014.
12. Lin, Yi-Lin, and Gang Wei. "Speech emotion recognition based on HMM and SVM." 2005 international conference on machine learning and cybernetics. Vol. 8. IEEE, 2005.
13. Hunnicutt, S. Grapheme-to-phoneme rules: A review. *Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden, QPSR* **1980**, *2-3*, 38–60.
14. Taylor, Paul. "Hidden Markov models for grapheme to phoneme conversion." Ninth European Conference on Speech Communication and Technology. 2005.
15. Bisani, Maximilian and Ney, Hermann. Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication* **2008**, *50*, 434–451.
16. Yao, Kaisheng and Zweig, Geoffrey. Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. *arXiv preprint arXiv:1506.00196* **2015**.
17. Rao, Kanishka, et al. "Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks." 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015.
18. Yolchuyeva, Sevinj and Németh, Géza and Gyires-Tóth, Bálint. Transformer based grapheme-to-phoneme conversion. *arXiv preprint arXiv:2004.06338* **2020**.
19. Tan, Xu and Qin, Tao and Soong, Frank and Liu, Tie-Yan. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561* **2021**.
20. Wang, Yuxuan and Skerry-Ryan, RJ and Stanton, Daisy and Wu, Yonghui and Weiss, Ron J and Jaitly, Navdeep and Yang, Zongheng and Xiao, Ying and Chen, Zhifeng and Bengio, Samy and others. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135* **2017**.

21. Shen, Jonathan, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018.
22. Arık, Sercan Ö., et al. "Deep voice: Real-time neural text-to-speech." International conference on machine learning. PMLR, 2017.
23. Ren, Yi and Ruan, Yangjun and Tan, Xu and Qin, Tao and Zhao, Sheng and Zhao, Zhou and Liu, Tie-Yan. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems* **2019**.
24. Gupta, Shikha and Jaafar, Jafreezal and Ahmad, WF Wan and Bansal, Arpit. Feature extraction using MFCC. *Signal & Image Processing: An International Journal* **2013**, *4*, 101–108.
25. Speights Atkins, Marisha and Bailey, Dallin J and Boyce, Suzanne E. Speech exemplar and evaluation database (SEED) for clinical training in articulatory phonetics and speech science. *Clinical Linguistics & Phonetics* **2020**, *34*, 878–886.
26. Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Łukasz and Polosukhin, Illia. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
27. The LJ Speech Dataset. Available online: URL (https://keithito.com/LJ-Speech-Dataset/).
28. Streijl, Robert C and Winkler, Stefan and Hands, David S. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems* **2016**, *22*, 213–227.