applied sciences

MDPI

Article

AI Enhancements for Linguistic E-Learning System

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

[1] Department of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221000, China; 6476@cumt.edu.cn (J.L.); tingting_xu@cumt.edu.cn (T.X.)
[2] Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36830, USA; czr0072@auburn.edu (C.R.)
[3] Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36830, USA; laolvyenq@gmail.com
* Correspondence: chenwdavior@163.com

Abstract: E-learning systems have been considerably developed after the COVID-19 pandemic. In our previous work, we developed a linguistic interactive E-learning system for phonetic transcription learning. In this paper, we propose three artificial-intelligence-based enhancements to this system from different aspects. Compared with the original system, the first enhancement is a disordered speech classification module; this module is driven by the MFCC-CNN model, which aims to distinguish disordered speech and nondisordered speech. The accuracy of the classification is about 83%. The second enhancement is a grapheme-to-phoneme converter. This converter is based on the transformer model and designed for teachers to better generate IPA words from the regular written text. Compared with other G2P models, our transformer-based G2P model provides outstanding PER and WER performance. The last part of this paper focuses on a Tacotron2-based IPA-to-speech synthesis system, this deep learning-based TTS system can help teacher generate high-quality speech sounds from IPA characters which significantly improve the functionality of our original system. All of these three enhancements are related to the phonetic transcription process. and 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

Citation: Liu, J.; Li, S.; Ren, C.; Lyu, Y.; Xu, T.; Wang, Z.; Chen, W. AI Enhancements for Linguistic E-Learning Systems. *Appl. Sci.* **2023**, *13*, 10758. https://doi.org/10.3390/app131910758

Academic Editors: Yu Liang, Wenjun Wu and Ying Li

Received: 30 August 2023
Revised: 23 September 2023
Accepted: 25 September 2023
Published: 27 September 2023

1. Introduction

Phonetic transcription, a process that represents speech sounds using special symbols, plays an important role in the linguistic education field. Generally, 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 a previous work, we developed an interactive E-learning system focused on 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 paper, 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 uploading its corresponding phonetic format as the answer; the student listens to the questions and types the answers on an IPA keyboard. The system then automatically calculates the similarity between the student's answers and the prestored correct answer using the edit distance algorithm and generates the grade [3,4].

Our study mainly focused on the teacher part and aimed at improving the experience of teachers. The prototype of the linguistic system was designed for the Communication Disorder Department. Teachers always require large amounts of time to distinguish disordered speech from nondisordered speech. Moreover, thonetic transcription is a process that generates words in phonetic format (IPA characters) from their written format. For teachers, adding questions in the system also requires them to add the corresponding correct answer in phonetic format. A grapheme-to-phoneme converter assists teachers in finishing this step and significantly improves the user experience. In the last part, the phonetic transcription exams in our system require the speech sound of English words or brief phrases; using the recordings of students or staff may cause privacy issues. A high-quality speech synthesis system perfectly solves this problem. All of the above challenges inspired us to build the following enhancements for our E-learning system.

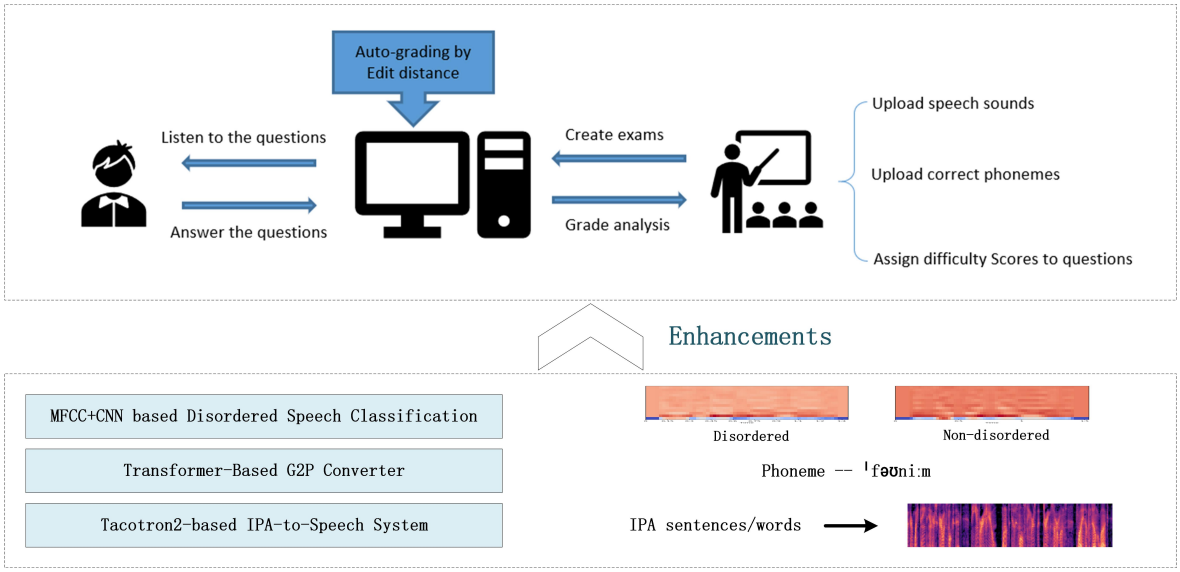

Figure 1. The enhancements to the linguistic E-learning system.

The three enhancements that we propose include the following:

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

The disordered speech classification module aims to to distinguish disordered speech from nondisordered speech. As an intelligent linguistic E-learning system, this module helps students understand the difference between correct pronunciation and disordered speech. During this study, we employed the Mel frequency cepstrum coefficient (MFCC) as the feature to represent the speech sound and a convolutional neural network (CNN) as the classification model [5].

The second enhancement module for 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, and sentences, where the pronunciation of speech sounds can be represented as phonemes. G2P conversion is a process that converts written words (grapheme) to their pronunciations (phonemes) [6]. With this G2P converter, teachers can easily extract the words in their phonetic format. We used 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 low word error rate (WER) and phoneme error rate (PER) [7].

The Tacotron2-based IPA-to-speech synthesis (text-to-speech) system is the last enhancement for our system. In order to help students better understand the pronunciation

of IPA symbols, this module was introduced for directly generating high-quality speech audio files from IPA symbols. Furthermore, this module helps teachers easily acquire audio files as a part of the question in the exam system. It took 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 Literature

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 also significantly increasing with increases in Internet access. For example, Zhejiang University deployed over 5000 courses online in two weeks on the platform DingTalk ZJU [8], and the Imperial College London started offering courses on Coursera starting in 2020.

E-learning is an approach that delivers knowledge or skills remotely and interactively using electrical devices such as smartphones, tablets, and laptops. Compared with traditional classroom learning, E-learning can offer students flexible topics or subjects; students can interact with teachers or professors through email or platforms without the restrictions imposed by physical distance. It cannot entirely replace traditional classroom learning but provides an augmented learning environment. Coursera and Udemy are both successful in E-learning (online learning) platformsonin 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 linguistics education since the advantages of E-learning 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). Phonetic transcription is a process that represents the speech sounds using special characters or symbols [3].

2.2. Speech Disorders Classification

A speech disorder is a condition in which a person experiences problems creating or forming the speech sounds needed to communicate with others. It is a subproblem of speech classification. To solve speech classification problems, both a feature extraction function and a classification algorithm are required. There are two major features in speech classification/recognition subjects: the linear prediction coding (LPC) and the Mel grequency cepstrum coefficient (MFCC) [10]. Classification algorithms include dynamic time warping (DTW) [11], hidden Markov models (HMMs) [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. 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 in text-to-speech (TTS) systems and automated speech recognition (ASR) systems. Generally, the International Phonetic Alphabet (IPA) characters are employed to represent phonemes.

G2P conversion has long been a popular top in the NLP field. Researchers 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 a G2P conversion system. In 2005, the hidden Markov model was employed for G2P conversion by Paul Taylor with a 9.02% PER and 42.69% WER [14]. In 2008, Maximilian Bisani introduced joint-sequence models for G2P conversion. A joint-sequence model is a theoretically stringent probabilistic framework that is applicable to this problem. On different English data

sets, joint-sequence models provide better performance than hidden Markov models; for example, the PER on CMUdict was 5.88% and the WER on CMUdict was 24.53% [15]. With the development of neural network technology, deep learning models are playing important roles in the NLP field. G2P conversion, as a text-to-text task, was studied by training different deep learning models. In 2015, the Seq2Seq model was employed for G2P conversion by Kaisheng Yao from Microsoft Research; in this study, the PER on CMUDict was 5.45%, while the WER was 23.55% [16]. In the same year, long short-term memory recurrent neural networks were utilized for the same task, achieving a 9.1% PER and 21.3%, by Kanishka Rao [17]. In 2020, with a start-of-the-art model, the Transformer, the PER was increased to 5.23% and the WER to 22.1% [18].

2.4. Speech Synthesis Systems

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 part of the 20th century. The early computer-based speech synthesis approaches include 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 have been proposed and have become the main methods used 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 acoustic models generate acoustic features from linguistic features; then, the vocoders synthesize waveforms from acoustic features. Tacotron 1, Tacotron 2, Deep Voice, and Fast Speech are all end-to-end TTS examples [20–23].

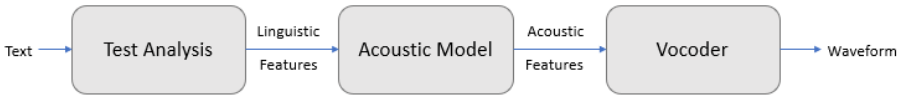

Figure 2. The structure of end-to-end TTS systems.

3. MFCC+CNN-Based Disordered Speech Classification

The first enhancement we developed is a speech classification module. This module has the ability to classify disordered speech and nondisordered speech. The speech classification problem can be divided into two subproblems: feature extraction and classification. In this study, we chose MFCC in image format to represent the features of human speech and the CNN model to perform the 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. Figure 3 illustrates the steps to generate MFCCs from audio [24].

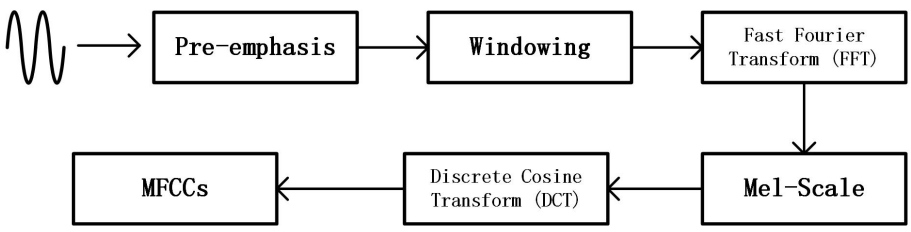

Figure 3. MFCC extraction process.

- Pre-emphasize the audio signal to increase to energy of the signal at a higher frequency.
- Break the sound signal into an 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 children's speech disorders were determined through parent reports and standardized assessments. Speakers in the SEED are between the ages of 2 and 85 years. A significant aspect of SEED is that it provides samples with and without speech disorders [25].

3.3. Implementation and Evaluation

About 1000 samples from the SEED were selected: 80% were used for training and the rest were used for validation. We used the Python Librosa library to process the MFCC values into their image format. Figure 4 shows two MFCC images with the same content but recorded by different recorders (with and without speech disorder).

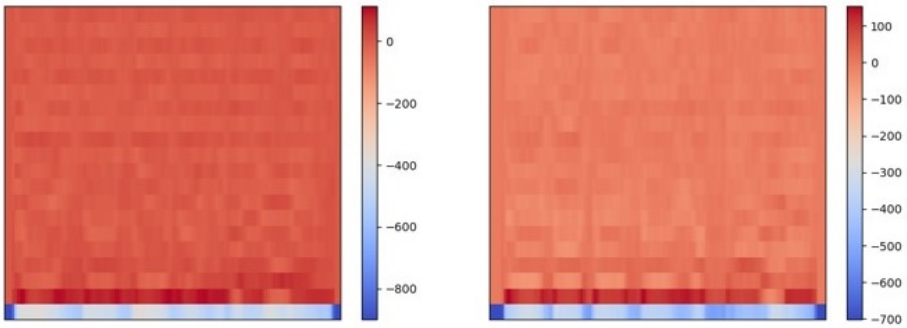

Figure 4. MFCC images of disordered and nondisordered speech.

Thus, this disordered speech classification problem could be transformed into an image classification problem. We then employed a CNN model to build a classification module. Convolutional neural network (ConvNet/CNN) is a deep learning algorithm widely used for image classification and computer vision tasks. Table 1 shows the structure of the CNN model used in our classification module, which took 150 epochs for training. Figure 5 shows the loss and accuracy of our model. The average classification accuracy was about 83%; this disordered speech classification module can neglect the content and the recorder of the speech, which means it is quite efficient and extensive.

Table 1. CNN model utilized in classification module.

Layer	Output Shape	Parameter Number
conv2d	(None, 148, 148, 32)	896
max_pooling2d	(None, 74, 74, 32)	0
conv2d_1	(None, 72, 72, 64)	18,496
max_pooling2d_1	(None, 36, 36, 64)	0
conv2d_2	(None, 34, 34, 128)	73,856
max_pooling2d_2	(None, 17, 17, 128)	0
conv2d_3	(None, 15, 15, 128)	147,584
max_pooling2d_3	(None, 7, 7, 128)	0
flatten	(None, 6272)	0
dense	(None, 512)	3,211,776
dense_1	(None, 1)	513

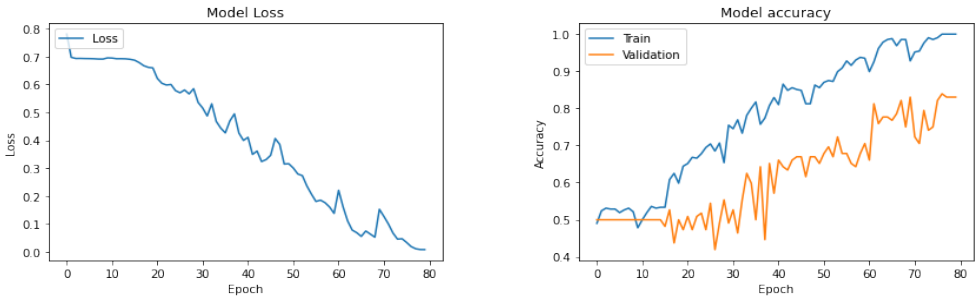

Figure 5. Loss and accuracy of speech classification module.

3.4. Discussion

In our first enhancement, we designed an MFCC+CNN-based disorder speech classification module. It can be regarded as a variant of speech classification but focuses on disordered and nondisordered speech. We developed a new solution in this classification field, where MFCCs ideally represent the features of human speech, and the image of MFCCs is classified by the CNN model. This classification module can help teachers remove disordered speech files from exams and improve the quality of the questions using our E-learning system.

4. Transformer-Based Multilingual G2P Converter

As we discussed above, the core function of our E-learning system is the provision of interactive phonetic transcription exams. The questions in this exam consist of the audio of words/phrases and their corresponding pronunciation (presented in IPA format). The teacher needs to preinput the correct answer to the system to activate the auto-grading module; so, generating IPA characters from written language can be a challenge for teachers. This was our inspiration for building the G2P converter. Grapheme-to-phoneme (G2P) conversion is a process of generating words in their IPA format from written format. G2P converters can be regarded as variants of machine translators [5].

4.1. Data Selection

There are two different kinds of characters used to represent the pronunciation of words and phrases: CMUDict characters and 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, IPA symbols are more widely used in multiple languages. Table 2 illustrates some samples of CUMdict 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 chose IPA symbols for the representation of pronunciation. Furthermore, to build a multilingual G2P system, we also investigated the French–IPA and Spanish–IPA converters. Table 3 lists all the datasets we employed in this study. The first two English–IPA datasets were used to investigate how the size of the data influences the G2P systems' performance, and the French–IPA and Spanish–IPA datasets were utilized to inspect the feasibility of the multilinguistics.

Table 3. Datasets used for training.

Dataset	Number of Word Pairs	For Validation
English–IPA	125,912	20%
French–IPA	122,986	20%
Spanish–IPA	99,315	20%

4.2. Transformer-Based G2P Converter

The Transformer model is an encoder–decoder model with an 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 (multihead attention) and a feed-forward layer. The decoder includes two multihead attention layers and one feed-forward layer. The encoder maps input sequences/words into attention-based representations; the decoder then takes the continuous representations and generates the output. Figure 6 shows the structure of the attention mechanism in the Transformer.

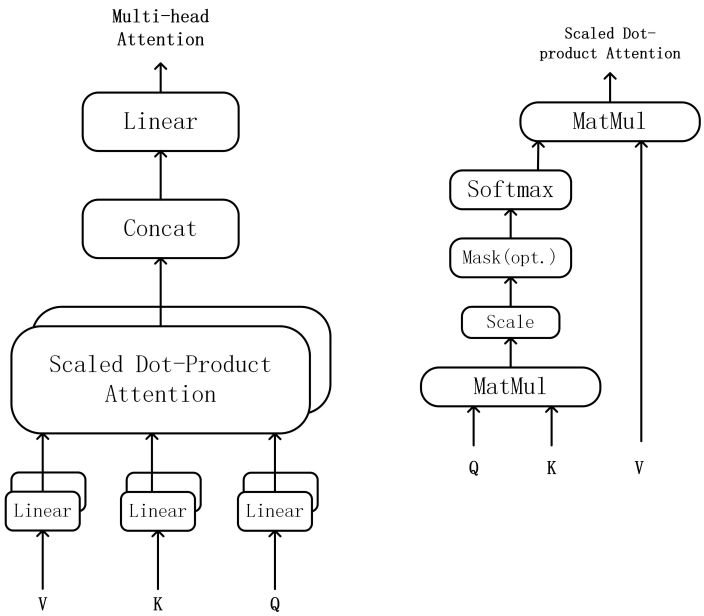

Figure 6. Scaled dot-product attention and multihead attention in Transformer model.

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:

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

Multihead 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 multihead attention extracts the specific representation, which allows the whole model to receive information from different subspaces. For multihead attention:

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

$$where head_i = Attention(QW_i^Q, KW_i^K, VW_i^V);$$

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

4.3. Implementation and Evaluation

Multiple pieces of training were implemented with the Nvidia Tesla P100 graphic card. We employed a six-layer Transformer model and the Adam optimizer in Keras with a learning rate of 0.0001. The phoneme error rate (PER) and word error rate (WER) were 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 used for measuring the accuracy in an ASR system. The formulation of WER is:

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

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].

Figures 7–9 display the performance of our multilingual G2P converter. For the English G2P converter, it took 220 epochs of training, the PER was about 2.6%, and the WER was 10.7%. For the French and Spanish G2P converters, 190 epochs of training were needed. The PER and WER for the French–IPA converter were 2.1% and 12.3%, while the PER and WER for the Spanish–IPA converter were 1.7% and 12.7%. Compared with the other models discussed in Section 2, our models outperformed the others in converting accuracy. Table 4 provides the comparison among different G2P converters.

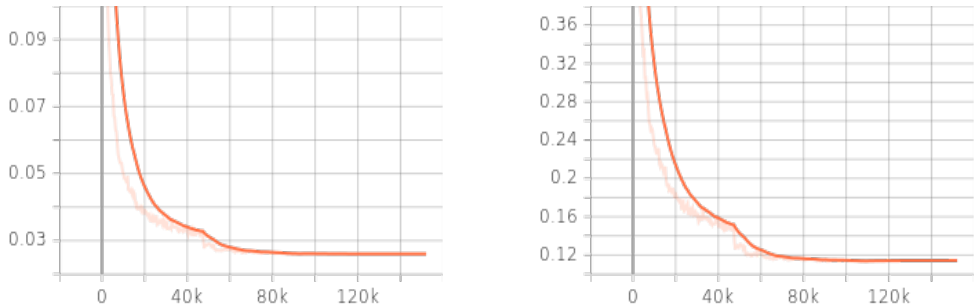

Figure 7. The PER and WER for English–IPA converter.

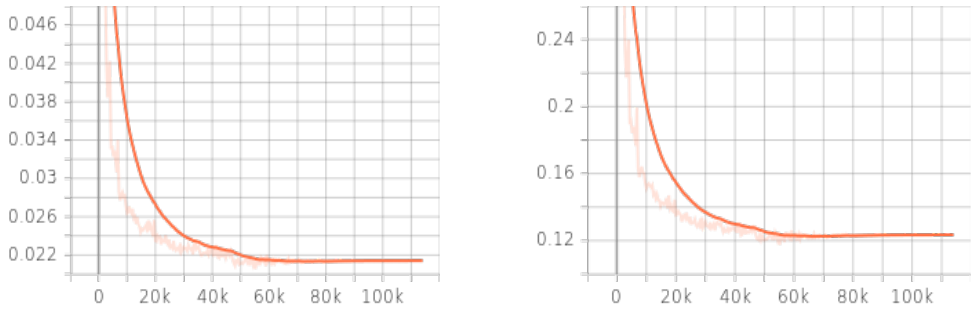

Figure 8. The PER and WER for French–IPA converter.

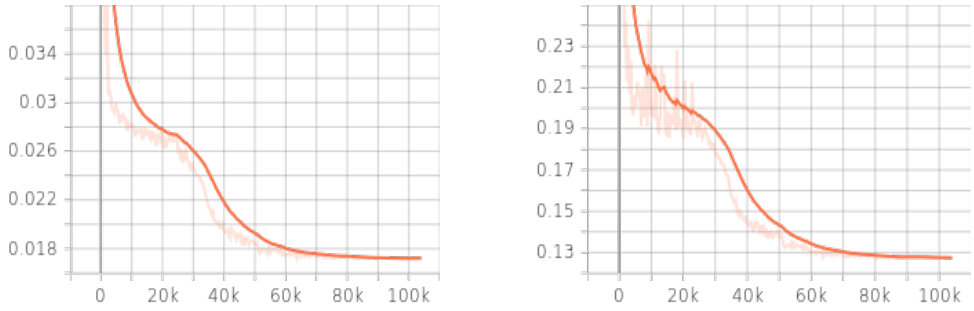

Figure 9. The PER and WER for Spanish–IPA converter.

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%**

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	dɪspl'ɛʒə	dɪspl'ʒə
	buoyant	b'ɔɪənt	b'ɔɪənt
	immortal	ɪm'ɔːtəl	ɪm'ɔːtəl
Spanish	ababillarais	aβaβiʎaɾis	aβaβiʎaɾis
	cacofónicos	kako'fonikos	kako'fonikos
	cadañega	kaðaeɣa	kaðaɲeɣa
French	câlineriez	kalinəʁje	kalinəʁje
	damasquiner	damaskine	damaskine
	effrangé	efʁɑ̃ʒe	efʁɑ̃ʒe

4.4. Discussion

In this part, we describe our second enhancement, a Transformer-based G2P converter. The results showed that the performance of our model is superior to that of other statistical methods and deep learning approaches. Grapheme-to-phoneme conversion is a process that is closely related to phonetic transcription. The multilingual G2P converter is a tremendous contribution to teachers teaching phonetic transcription. With this converter, teachers can easily generate IPA characters from one regular word without browsing a

dictionary; moreover, our G2P converter has the ability to be extended to other languages. This feature gives our system the potential to be used as a multilingual E-learning system.

5. Tacotron2-Based IPA-to-Speech System

As we mentioned above, the questions in phonetic transcription exams consist of speech audio and IPA symbols. From the teacher's view, searching and acquiring appropriate speech audio with high quality along with their text are challenging. Additionally, the text of the audio must be converted to IPA format. From this perspective, we designed a text-to-speech (TTS) system that can directly generate speech sounds from words/phrases/sentences in IPA format [19].

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

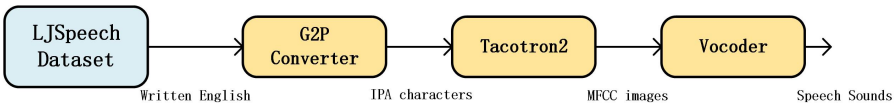

Figure 10. The main process of the IPA-to-speech system.

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 of TTS conversion is to convert text into speech sounds. We employed the LJSpeech dataset to build our IPA-to-speech system. LJSpeech is a public domain speech dataset consisting of 13,100 speech audio files of a single speaker. Clips vary in length from 1 to 10 s and have a total length of approximately 24 h. The LJSpeech data are constructed in pairs of <English Sentences, Speech Samples>. To build our IPA-to-speech system, all the written English sentences in LJSpeech dataset were converted to their IPA format; this data preprocessing step required our English G2P converter discussed in Section 4. Table 6 gives several converted samples from the LJSpeech dataset [27].

Table 6. The original text in the 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 ɪŋ 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 ɵroʊ ðə riːspɑːnsəbɪləti aːn 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 the characters introduced by the Google team. The input of the Tacotron model is characters, and the output of the model is the corresponding raw spectrograms. The limitation of the Tacotron model is the vocoder part: the sound generated by the Griffin–Lim algorithm is not of high quality. Thus, in this study, we employed the Tacotron2 model to build our IPA-to-Speech system. Compared with the original Tacotron model, 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.

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

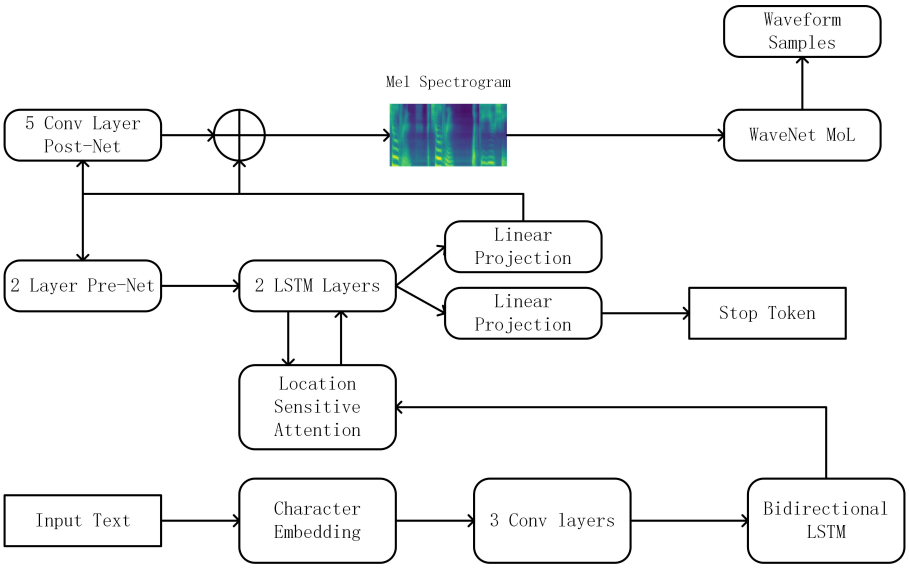

Figure 11. The structure of the Tacotron2 TTS model.

5.3. Implementation and Evaluation

A Nvidia Tesla P100 GPU was employed to train the Tacotron2 model. We selected a batch size of 48 on a single GPU with a 0.0001 learning rate. All the audio in LJSpeech was used for training, which contains about 24.6 h of audio recorded by a woman. Every text in the dataset needs to be spelled out; for example, the number ”10” should be represented as ”ten”. This required about 20 h to finish 180,000 steps, which was about 200 epochs. Figures 12 and 13 show the training loss, validation loss, target Mel, and predicted Mel.

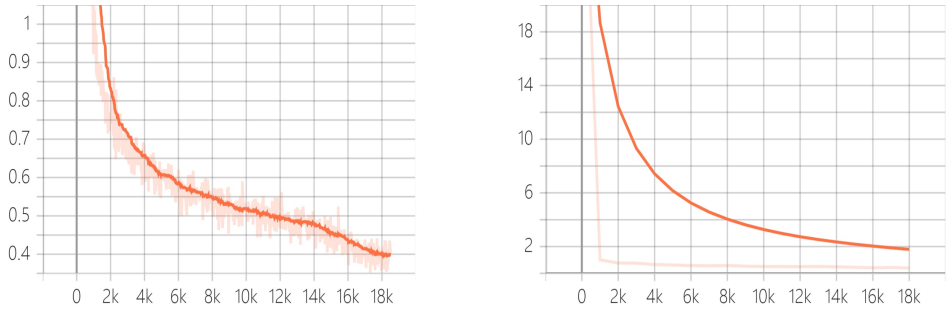

Figure 12. The training and validation losses of our Tacotron2 model.

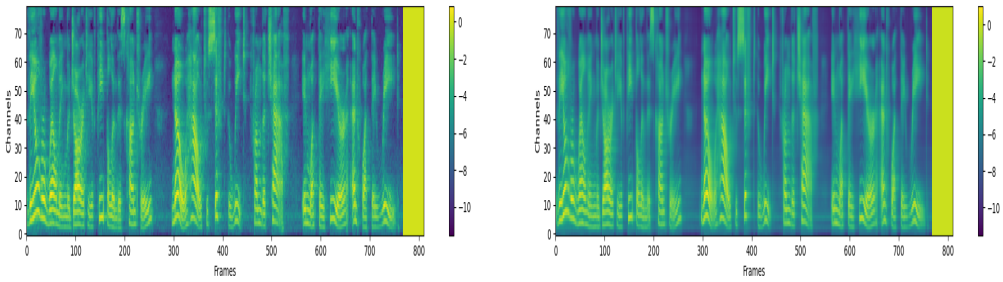

Figure 13. The target−Mel and predicted−Mel.

The mean opinion score (MOS) was utilized to evaluate the performance of our IPA-to-Speech system. The MOS is a metric that is widely used to measure the quality of speech generated by a TTS system. Generally, the MOS score is a rating from one to five, 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 a linguistic department to help us evaluate our system using the MOS. The mean opinion score of our IPA-to-speech system is about 4.05 [28].

5.4. Discussion

In this part, we developed the last enhancement, which is a Tacotron2-based IPA-to-speech system. This system aims to directly generate speech from IPA characters. It has two advantages for those teaching phonetic transcription: the first is that with this speech synthesis system, teachers do not need to record their or other volunteers' speech, which helps to avoid the recorder accents problem and potential privacy problems; second, from the pedagogical aspect, the direct process of conversion from IPA characters to speech can also help students to better understand the pronunciation of single IPA characters and combined words.

6. Conclusions and Future Work

In this paper, we proposed three artificial intelligence enhancements for our linguistic E-learning system. Firt, the disordered speech classification module utilizes MFCCs to represent the features of speech, and a CNN model is used to build the classification function, which achieves the classification of disordered and nondisordered speech. Second, the grapheme-to-phoneme module uses a Transformer model, which provides high-accuracy G2P conversion. Finally, the IPA-to-speech module employs the Tacotron2 model and generates high-quality speech sounds from IPA characters.

All of these enhancements improve the functionality of the system compared with that of the traditional methods. The deep-learning-based speech classification module can better extract the features of disordered and nondisordered speech, and the CNN model provides better classification performance. Transformer-based G2P has a notably higher conversion accuracy than the statistical and other DL approaches. Moreover, IPA-to-speech provides a new idea to directly generate speech sounds from IPA characters. We found that with the development of artificial intelligence, more and more deep-learning-based technologies can be utilized in the education field, especially in linguistics, since linguistics is related to the natural language process in the field of AI. Our original E-learning system is a simple system focusing on phonetic transcription exams: the questions are speech sounds, while the answers are IPA characters. After the implementation of our three AI enhancements, our system is now a comprehensive linguistic E-learning system. It not only distinguishes disordered speech but also generates IPA characters from regular English words and

creates high-quality speech sounds from IPA characters. The last two enhancements will greatly improve the experience for teachers. Also, these deep learning-based models can be pretrained and embedded into the system, so new teachers or educators do not need to complete any extra work to familiarize themselves with the system.

There are still other tasks that should be performed after this study. The first is evaluation: as we could not find a baseline against which to evaluate the performance of our IPA-to-speech system, we hired 20 volunteers to evaluate the system using the MOS. Future studies should include performing a more objective evaluation of the system. Also, the linguistic system may suffer from potential privacy problems. Since some of the speech sounds were recorded by students or teachers, we will try to replace all the speech sounds that may be related to these privacy problems with the generated speech sounds.

Author Contributions: Conceptualization, J.L.; methodology, Y.L.; software, S.L. and C.R.; validation, T.X.; formal analysis, T.X.; investigation, J.L.; writing—original draft preparation, J.L.; writing—review and editing, Z.W. and W.C.; supervision, W.C. All authors have read and agreed to the published version of the manuscript.

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

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We leveraged two open datasets for evaluation, named The LJSpeech Dataset and SEED. The LJSpeech can be downloaded at https://keithito.com/LJ-Speech-Dataset/ (accessed on 1 January 2017) and the SEED is not available due to privacy.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Brown, A. International phonetic alphabet. In *The Encyclopedia of Applied Linguistics*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
2. Howard, S.J.; Heselwood, B.C. Learning and teaching phonetic transcription for clinical purposes. *Clin. Linguist. Phon.* **2002**, *16*, 371–401.
3. Seals, C.D.; Li, S.; Speights Atkins, M.; Bailey, D.; Liu, J.; Cao, Y.; Bassy, R. Applied webservices platform supported through modified edit distance algorithm: Automated phonetic transcription grading tool (APTgt). In *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, Copenhagen, Denmark, 19–24 July 2020*; Springer: Cham, Switzerland, 2020.
4. Liu, J.; Speights, M.; Bailey, D.; Li, S.; Luan, Y.; Mishra, I.; Cao, Y.; Seals, C. Optimization to automated phonetic transcription grading tool (APTgt)—Automatic exam generator. In Proceedings of the International Conference on Human-Computer Interaction, Virtual, 24–29 July 2021; Springer: Cham, Switzerland, 2021.
5. Liu, J.; Ren, C.; Luan, Y.; Li, S.; Xie, T.; Seals, C.; Speights Atkins, M. Transformer-Based Multilingual G2P Converter for E-Learning System. In Proceedings of the International Conference on Human-Computer Interaction, Virtual, 26 June–1 July 2022; Springer: Cham, Switzerland, 2022.
6. Liu, J.; Ren, C.; Luan, Y.; Li, S.; Xie, T.; Seals, C.; Speights Atkins, M. Speech Disorders Classification by CNN in Phonetic E-Learning System. In Proceedings of the International Conference on Human-Computer Interaction, Virtual, 26 June–1 July 2022; Springer: Cham, Switzerland, 2022.
7. Schwarz, P.; Matějka, P.; Černocký, J. Towards lower error rates in phoneme recognition. In Proceedings of the International Conference on Text, Speech and Dialogue, Brno, Czech Republic, 8–11 September 2004; Springer: Berlin/Heidelberg, Germany, 2004.
8. Wu, X.; Fitzgerald, R. 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, S. E-learning 2.0. *ELearn* **2005**, *10*, 1.
10. Madan, A.; Gupta, D. Speech feature extraction and classification: A comparative review. *Int. J. Comput. Appl.* **2014**, *90*, 20–25.
11. Mohan, B.J. Speech recognition using MFCC and DTW. In Proceedings of the 2014 International Conference on Advances in Electrical Engineering (ICAEE), Vellore, India, 9–11 January 2014.
12. Lin, Y.-L.; Wei, G. Speech emotion recognition based on HMM and SVM. In Proceedings of the 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, 18–21 August 2005; Volume 8.
13. Hunnicutt, S. *Grapheme-to-Phoneme Rules: A review*; Speech Transmission Laboratory, Royal Institute of Technology: Stockholm, Sweden, 1980; Volume QPSR 2–3, pp. 38–60.

14. Taylor, P. Hidden Markov models for grapheme to phoneme conversion. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005.
15. Bisani, M.; Ney, H. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Commun.* **2008**, *50*, 434–451.
16. Yao, K.; Zweig, G. Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. *arXiv* **2015**, arXiv:1506.00196.
17. Rao, K.; Peng, F.; Sak, H.; Beaufays, F. Grapheme-to-Phoneme Conversion Using Long Short-Term Memory Recurrent Neural Networks. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015.
18. Yolchuyeva, S.; Németh, G.; Gyires-Tóth, B. Transformer based grapheme-to-phoneme conversion. *arXiv* **2020**, arXiv:2004.06338.
19. Tan, X.; Qin, T.; Soong, F.; Liu, T.-Y. A survey on neural speech synthesis. *arXiv* **2021**, arXiv:2106.15561 .
20. Wang, Y.; Skerry-Ryan, R.J.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards end-to-end speech synthesis. *arXiv* **2017**, arXiv:1703.10135.
21. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018.
22. Arık, S.Ö.; Chrzanowski, M.; Coates, A.; Diamos, G.; Gibiansky, A.; Kang, Y.; Li, X.; Miller, J.; Ng, A.; Raiman, J.; et al. Deep voice: Real-time neural text-to-speech. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, NSW, Australia, 6–11 August 2017.
23. Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.-Y. Fastspeech: Fast, robust and controllable text to speech. In Proceedings of the 33rd Conference on Neural Information Processing Systems, Vancouver, QC, Canada, 8–14 December 2019.
24. Gupta, S.; Jaafar, J.; Ahmad, W.W.; Bansal, A. Feature extraction using MFCC. *Signal Image Process. Int. J.* **2013**, *4*, 101–108.
25. Speights Atkins, M.; Bailey, D.J.; Boyce, S.E. Speech exemplar and evaluation database (SEED) for clinical training in articulatory phonetics and speech science. *Clin. Linguist. Phon.* **2020**, *34*, 878–886.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
27. The LJ Speech Dataset. Available online: https://keithito.com/LJ-Speech-Dataset/ (accessed on 1 January 2017).
28. Streijl, R.C.; Winkler, S.; Hands, D.S. Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives. *Multimed. Syst.* **2016**, *22*, 213–227.