# Generation of target speech with speaker individuality based on accent conversion for English pronunciation learning

Rei Hamakawa* and Michiharu Niimi†
* Graduate School of Computer Science and Systems Engineering
Kyushu Institute of Technology, Iizuka, Japan
E-mail: hamakawa.rei112@mail.kyutech.jp
† Faculty of Computer Science and Systems Engineering
Kyushu Institute of Technology, Iizuka, Japan
E-mail: niimi@ai.kyutech.ac.jp

***Abstract*—This paper first provides an overview of the English pronunciation learning support tool. The tool aims to use "accent-modified speech that retains the learner's voice quality" as the "target speech." We propose a new conversion model based on conventional methods for this accent conversion. Specifically, we improve the conventional LSTM-based DNN model for accent conversion by adopting a transformer-based model. Our experiments investigated the model's ability to handle the unique katakana pronunciation characteristic of Japanese speakers. The results confirmed the effectiveness of the proposed conversion method, although challenges remain, such as the scarcity of Japanese speech data and the need to improve the accuracy of speaker identity retention.**

## I. Introduction

According to a report by EF Education First[1], Japan ranks 87th out of 113 non-native English-speaking countries and regions, and 15th out of 23 in Asia, in other words, Japan is classified as a "low proficiency" level. Therefore, the reform of English education is an urgent issue. The new curriculum guidelines were introduced in 2020 for elementary schools, 2021 for junior high schools, and 2022 for high schools. These guidelines aim to develop communication skills through a balanced learning approach to listening, reading, speaking, and writing in English.

However, it is extremely challenging to provide appropriate instruction to each student due to the insufficient English proficiency of English teachers and the shortage of Assistant Language Teachers (ALTs). From several reports with respect to English teacher quality, it is exceedingly difficult to provide appropriate instruction in speaking to each student. Pronunciation instruction is particularly important for Japanese students, as there are significant differences between Japanese and English pronunciation. In pronunciation instruction, simply listening to the target audio makes it difficult for beginner learners of English to recognize the differences between their own pronunciation and the target pronunciation. If we can provide ideal audio that makes the differences between the learner's pronunciation and the target pronunciation easily recognizable, and visually represent these differences using phonetic symbols or other means, it can lead to more effective learning. Furthermore, developing this as a PC-based system (English pronunciation learning support tool) will enable individual learning.

The concept of a "golden speaker," which refers to a speaker with a voice similar to second-language (L2) learners, has been applied in computer-assisted pronunciation training (CAPT) for L2 learners[2][3]. Because the "golden speaker" has the same voice quality as them, it serves as an ideal target for pronunciation training compared to native speakers' voices. Nagano et al.'s [4] study explored the potential of using modified learner speeches for pronunciation training. In this study, Japanese learners were divided into two groups: one trained to imitate the speech of a standard English speaker, and the other trained to imitate their own speeches pre-modified to match the prosody of the standard English speaker. The results showed that the latter group's post-training speech was rated as closer to native-like than that of the first group. Hence, we propose using accent-converted speech that retains the learner's voice quality as the target speech in pronunciation instruction. We expect that this approach allows learners to hear the correct pronunciation in their own voice quality, which makes it easier for them to recognize the differences between their own pronunciation and the target pronunciation.

Based on this concept, we're trying to develop an English pronunciation learning support tool that generates golden speaker's speech with the L2 learner's voice quality and the correct pronunciation as target speech, shows pronunciation differences visually, and provides feedback. The golden speaker's speech with the learner's voice quality and the correct pronunciation is generated using accent conversion technology, a type of speech processing technology. The input to the accent converter is the L2 learners' spoken English, and this technology adjusts the learners' accent to the native accent while preserving the speaker's voice quality. Traditional accent converters use an LSTM-based DNN model, but we improve it by adopting a transformer-based model. While DNNs require a large amount of training data, Japanese data have not

traditionally been included as processing targets. This paper collects a small amount of Japanese English pronunciation data and conducts experiments.

First, in section 2, we describe the proposed tool. Then, in section 3, we discuss the generation of model speech with the learner's voice quality. Section 4 presents the experimental results of accent conversion using Japanese data, and finally, section 5 concludes the paper.

## II. Proposed English Speech Learning Support Tool

In this section, we outline the concept of an English speech learning support tool designed to address the issues identified in the pronunciation learning process.

### A. The Process of Learning Pronunciation

To acquire native English pronunciation, it is necessary to work on narrowing the gap between one's pronunciation and that of native English speakers. This requires first learning the fundamental knowledge of pronunciation, such as phonetic symbols and phonological changes, and then engaging in actual pronunciation training. In other words, we assume that the cycle: recording own pronunciation, comparing it with native audio, and then receiving the feedback is needed, and this cycle represents what we consider the ideal pronunciation training.

### B. Current Issues

Typically, when learning pronunciation, learners first listen to the recorded pronunciation that comes with their study materials, imitate it, and practice. They progress from words to phrases and then to sentences. However, there are phonological changes such as linking, reduction, and flapping, leading to pronunciations that often differ from the phonetic symbols provided. Additionally, the recorded pronunciations in learning materials are usually voiced by a single male or female native speaker for each word, sentence, or passage. Therefore, if the voice quality of the English learner differs significantly from that of the recorded speaker, it becomes difficult to recognize the differences in pronunciation. Moreover, learners cannot receive feedback on how their pronunciation differs from the ideal pronunciation, which makes it difficult for many Japanese learners to overcome the typical Japanese accent, known as the katakana accent in Japan.

### C. Overview of the Support Tool

In our support tool, users first press the record button, read out a word or sentence of their choice, and records it. The system then converts the recorded speech to match a native accent while preserving the user's voice quality. This generates a speech that makes it easier to recognize pronunciation differences. They can play back and compare their own speech with the converted speech. Furthermore, the phonetic symbols of the their recorded speech and the generated speech are extracted and displayed. This provides visual feedback on the pronunciation differences. The differences between the extracted phonetic symbols of the their pronunciation and the target pronunciation are highlighted in red. This allows beginner learners to recognize the differences between their own pronunciation and target pronunciation more easily, As a result of this, pronunciation learning becomes more efficient even for self-study.

Using phonetic symbols to show differences is only one effective method in pronunciation learning. Therefore, future developments of this system include implementing features to visualize differences in characteristics beyond phonemes, such as prosody.

## III. Generation of Model Speech with the Learner's Voice Quality

Using accent conversion technology with DNN models, we generate target speech with the learner's voice quality to make pronunciation differences easily recognizable.

### A. Conventional Method

Waris Quamer et al. [5] proposed a method that does not require reference speech for accent conversion. This method is based on Tacotron2[6] model that generates speech from text. Fig. 1 shows an overview of this system.

The system consists of six models: acoustic model, accent encoder, seq2seq model (translator), speaker encoder, seq2seq model (synthesizer), and vocoder. The flow of Waris Quamer et al.'s accent conversion method is as follows:

1) First, the acoustic model generates a phonetic-posteriorgram (PPG) that represents the linguistic content independent of the speaker. In Waris Quamer's method, the acoustic model was implemented using Kaldi[7]. Typically, the PPG is high-dimensional, but in this method, bottleneck features from the layer prior to the final softmax layer of the acoustic model are used instead of the PPG to extract low-dimensional linguistic features. The PPG has dimensions of time frames $\times$ 256.
2) The accent encoder generates a fixed-dimension embedding vector, which represents the accent from the native speech. This model was trained to maximize the cosine similarity between speech samples with the same accent. The accent embedding is a 256-dimensional feature which captures a speaker's accent.
3) The PPG and accent embedding are fed into the seq2seq model (translator), which adjusts the PPG based on the accent embedding to convert L2 PPG into L1 PPG.
4) The speaker encoder generates a fixed-dimension embedding vector which represents the speaker identity. This model is trained as speaker-verification and maximize the cosine similarity between utterances from the same speaker. The speaker embedding is a 256-dimensional feature which capture a speaker's identity.
5) The converted L1 PPG and speaker embedding are fed into the seq2seq model (synthesizer), which adjusts based on the speaker embedding to generate a mel-spectrogram with L1 accent while preserving L2 speaker's voice quality.

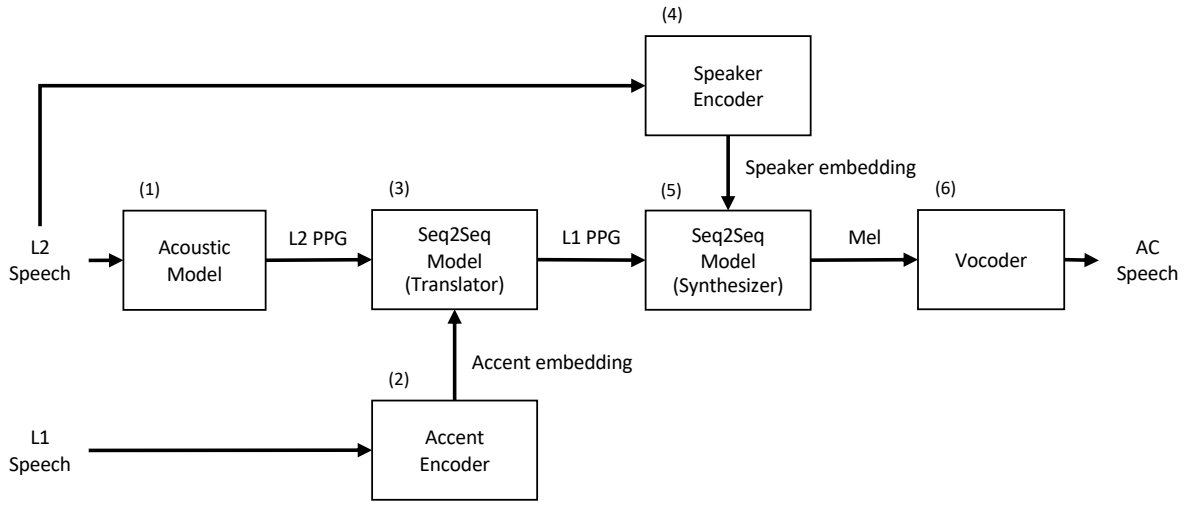


Fig. 1. Overview of conventional system

6) The generated mel-spectrogram is fed into the vocoder to generate the speech with L1 accent and L2 speaker's voice quality.

### *B. Proposed Method*

In the conventional method, the translator and synthesizer are LSTM-based models inspired by Tacotron2 whose architecture uses a location-sensitive attention mechanism, which is prone to fail with long utterances. Waris Quamer et al.[5] suggest that another direction for future research is to make the model robust for synthesizing long utterances.

Additionally, since the synthesizer is trained on data from 24 speakers and generates mel-spectrograms from PPGs, it produces high-precision outputs for the voice qualities of those 24 speakers. However, through several experiments, we found that it does not sufficiently retain the voice quality when generating mel-spectrograms for speakers not included in the training data. In the experiments, training the Tacotron2-based model using two NVIDIA Tesla V100 GPUs took a long time due to the use of LSTMs in both the encoder and decoder. LSTMs require sequential processing because the output at each time step depends on the output from the previous time step. This dependency makes parallel processing impossible. It causes long training time, therefore, fine-tuning with the conventional method is impractical. Based on the above considerations, we identified the following issues in prior research:

- Synthesis of long utterances
- Handling of speakers not included in the training data
  - Fine-tuning could be a solution, but their model makes this very time-consuming.

To address these issues, we propose replacing the translator and synthesizer with transformer-based models. The transformer uses self-attention mechanisms that allow it to retain and process information from earlier parts of long sequential data. This capability makes it robust for long utterance synthesis. Additionally, self-attention mechanisms enable parallel processing. They significantly reduce training time compared to Tacotron2 and make fine-tuning more feasible. We replace the Tacotron2-based translator and synthesizer models with transformer-based models, pre-train on data from numerous speakers, and fine-tune with a small amount of data from a single L2 learner. This approach aims to generate personalized "golden speaker" speech and solve the problems of the conventional model. The other models remain unchanged from the conventional method, and we use pre-trained models for experiments.

Fig. 2 shows the transformer-based translator model. Unlike the standard Transformer, which takes a sequence of words as input and converts each word to a vector, the input to the proposed model is already vectorized PPG, which is processed directly. The differences from a standard transformer are summarized below:

- Adding Tacotron2's pre-net to the input side of the encoder and decoder
- Concatenating the encoder's output with the accent embedding before inputting it to the decoder
- Adding linear projections to generate L1 PPG and predict stop tokens
- Adding a post-net, composed of five CNN layers, to the output side

The pre-net from Tacotron2 functions as a bottleneck. It compresses and extracts the input information, which contributes to learning efficiency. To convert the PPG into the target pronunciation based on the accent embedding, the encoder's output is concatenated with the accent embedding and input into the decoder. The stop token controls the output termination timing. The post-net improves PPG quality by predicting and adding residuals to the predicted PPG. The structure of the encoder and decoder follows the standard

transformer design. Multihead Attention consists of multiple attention layers which allows simultaneous focus on multiple parts. The Masked Multihead Attention in the decoder functions similarly but operates with masked information. The Feed Forward Network (FFN) is a layer that performs linear transformations.

The synthesizer also has a transformer-based structure similar to the translator. It takes PPG and speaker embedding as inputs and outputting a mel-spectrogram with dimensions of time frames $\times$ 80 at the final output layer.

Since the transformer-based model uses self-attention mechanisms, it can process long sequential data in parallel. This reduces training time compared to the LSTM-based Tacotron2 and makes fine-tuning easier. We aim for high-precision accent conversion by pre-training the translator and synthesizer on English speech data from many speakers, then fine-tuning with a small amount of data from one speaker. In the translator training, the model trained without Japanese speaker data is fine-tuned with a small amount of Japanese speaker data to adapt to the unique katakana pronunciation of Japanese speakers. In synthesizer training, fine-tuning is performed using one speaker's English utterances to retain the user's voice quality sufficiently. Although it is ideal to fine-tune using English utterances, since English beginners may not be able to speak English well, we also attempt fine-tuning using only Japanese utterances to synthesize the "golden speaker" speech.

## IV. Experiment

### A. Tranning Data

For the pre-training of the translator, we use the CMU ARCTIC [8] and L2-ARCTIC [9] datasets. The CMU ARCTIC database is constructed as a phonetically balanced single-speaker database of American English. L2-ARCTIC is a non-native English speech corpus designed for research in speech conversion, accent conversion, and mispronunciation detection. This corpus includes English read speech from 24 non-native speakers whose first languages (L1) are Hindi, Korean, Mandarin, Spanish, Arabic, and Vietnamese. In previous experiments using conventional methods, Japanese speakers were not included in the training data, and it is anticipated that these methods would not accommodate the unique katakana pronunciation characteristic of Japanese speakers. Therefore, in this study, to apply accent conversion to the distinctive katakana pronunciation of Japanese speakers, we collect and use the same reading sentences from a Japanese speaker as fine-tuning data for training.

For the pre-training of the synthesizer, we use the train-other-500 subset of the Librispeech dataset[10]. Additionally, to sufficiently maintain speaker characteristics and synthesize the speech of a "golden speaker," we collect data from two Japanese speakers for fine-tuning. The data from each speaker is used independently for fine-tuning. We conduct experiments using only the English utterances and only the Japanese utterances recorded from each speaker.

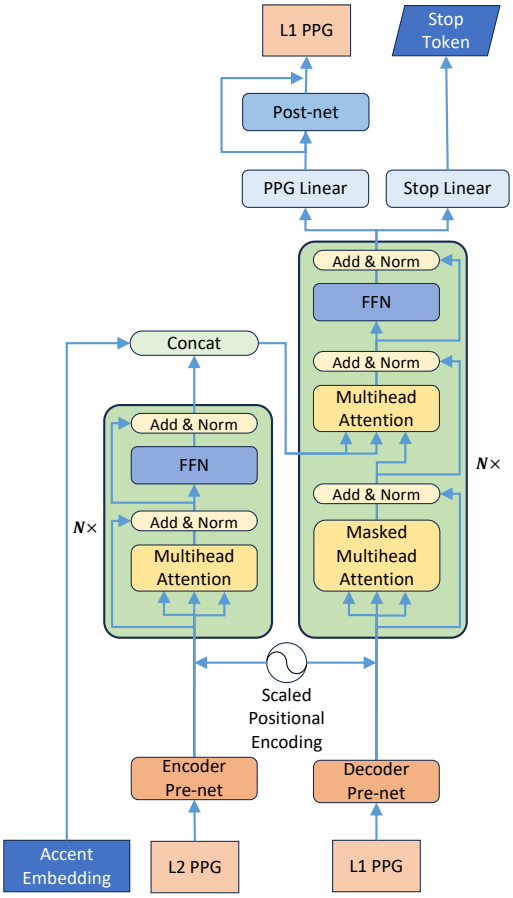


Fig. 2. The proposed model

TABLE I
TRANSLATOR EXECUTION GPU ENVIRONMENT SPECIFICATIONS

| Component | Specification |
|---|---|
| GPU | NVIDIA GeForce RTX 4070 Ti |
| GPU Memory | 12G |

### B. Experimental Environment and Model Parameters

Table I and Table II show the GPU of experimental environments for the translator and synthesizer, respectively. Due to the memory capacity limitations of the GPUs used, we set the number of layersN of the encoder and decoder to 1 and the number of heads to 2, as illustrated in Fig. 2. The same configuration is applied to the synthesizer.

### C. Training

*1) Pre-Training with Data Excluding Japanese Speakers:* In the translator experiments, pre-training is conducted using the CMU ARCTIC and L2-ARCTIC datasets, excluding Japanese speaker data. The input data includes the CMU ARCTIC speakers SLT, CLB, and RMS, as well as the non-native speakers from L2-ARCTIC (excluding NJS, YKWK, TXHC, and ZHAA). Unlike conventional methods that used only a

TABLE II
SYNTHESIZER EXECUTION GPU ENVIRONMENT SPECIFICATIONS

| Component | Specification |
|---|---|
| GPU | NVIDIA GeForce RTX 4090 |
| GPU Memory | 24G |

single native speaker's data from CMU ARCTIC as the target pronunciation for accent conversion, this study additionally incorporates the data of three native speakers used as inputs into the target pronunciation set. This expansion increases the combinations of input and target pronunciations. As a result of this, the variety of the training data is enhanced. The Adam optimizer is employed for training, with a batch size of 32 and 200 epochs.

*2) Fine-Tuning with Japanese Speaker Data:* The model trained with data excluding Japanese speakers is fine-tuned using data from a single Japanese speaker. The Adam optimizer is employed for fine-tuning, with a batch size of 32 and 100 epochs.

*3) Pre-training of the Synthesizer:* In the synthesizer experiments, we initially use approximately 30 utterances per speaker from the Librispeech train-other-500 subset for pre-training. The synthesizer is trained independently. Prior to this, PPG and speaker embeddings are generated using the acoustic model and speaker encoder, respectively. These are then input into the Synthesizer, which adjusts speaker characteristics based on the Speaker Embedding and predicts the mel-spectrogram from the PPG and speaker embedding. The Adam optimizer is employed for training, with a batch size of 24 and 200 epochs.

*4) Fine-Tuning of the Synthesizer:* The model pre-trained with Librispeech data is fine-tuned using data from Japanese speakers. The Adam optimizer is employed, and fine-tuning is performed over 400 epochs. Experiments are conducted using 25, 50, and 75 utterances of English speech and 25, 50, and 75 utterances of Japanese speech to investigate the impact of the amount of data used for fine-tuning on the accuracy of maintaining speaker's voice quality. The Adam optimizer is employed for these experiments, with a batch size of 24 and 400 epochs.

### D. Evaluation

*1) Evaluation Method for the Translator:* To assess the extent to which Japanese read-aloud speech can be converted to a native-like accent, we focus on the number of correctly accent-converted words within a single sentence, that is, the number of correctly accent-converted words in a sentence is divided by the total number of words in the sentence. For $n$ sentences, the average of these ratios is taken as the evaluation score. Note that the evaluation does not include metrics for audio quality or fluency.

*2) Evaluation Data for the Translator:* For evaluation data, we use 50 sentences read aloud by two non-native speakers whose native languages are Arabic and Mandarin, and two non-native speakers whose native language is Japanese.

TABLE III
EVALUATION RESULTS

| Native Language | Fine-tuning | |
|---|---|---|
| | Before | After |
| Arabic | 0.90 | 0.79 |
| Mandarin | 0.89 | 0.83 |
| Japanese (speaker1) | 0.63 | 0.81 |
| Japanese (speaker2) | 0.65 | 0.77 |

*3) Evaluation Results for the Translator:* Table 3 shows the evaluation results of the model before and after fine-tuning. Before fine-tuning, the model could convert about 90% of the words for speakers whose native languages are Arabic and Mandarin, but it could convert fewer than 65% of the words for speakers whose native language is Japanese. After fine-tuning, approximately 80% of the words were converted for speakers of all native languages, including Arabic, Mandarin, and Japanese. Fine-tuning to adapt to the unique katakana pronunciation of Japanese speakers reduced the conversion accuracy for speakers of Arabic and Mandarin. Inappropriate conversions included significantly different pronunciations of words after conversion, words that could not be recognized as words, and omissions in the output.

Additionally, we evaluated the pre-fine-tuning model using only 19 sentences composed of 11 to 13 words from the evaluation sentences. The results are shown in Table IV. The 50 sentences used for evaluation consisted of a maximum of 13 words. Comparing Tables III and IV, it can be said that the evaluation scores for speakers whose native languages are Arabic and Mandarin are not affected by the number of words.

*4) Evaluation Results for the Synthesizer:* Currently, fine-tuning of the synthesizer has only been conducted on data from two speakers, so quantitative and qualitative evaluations of speaker identity retention accuracy and other metrics have not been performed. Of the two speakers tested, one produced speech that seemed to better retain speaker identity when fine-tuned with both English and Japanese utterances compared to the conventional model. Additionally, it was observed that the number of data samples appeared to affect the audio quality.

For the other speaker, the fine-tuned model did not exhibit superior speaker identity retention compared to the conventional model. No significant impact of fine-tuning with either English or Japanese utterances was observed; both sets of data resulted in speech with a similar level of speaker identity retention. Moreover, the time required for fine-tuning the synthesizer in the experimental environment shown in Table II is presented in Table V. Even when fine-tuning with 75 data samples over 400 epochs, the process was completed in approximately 6 minutes. This demonstrates that fine-tuning a transformer-based model is feasible in practice.

### E. Discussion

In the pre-fine-tuning model for the translator, only 63% of the words in the data from speakers whose native language is Japanese were appropriately converted. This is likely due to the

TABLE IV
EVALUATION RESULTS FOR LONG SENTENCES

| Native Language | Before Fine-tuning |
|---|---|
| Arabic | 0.92 |
| Mandarin | 0.89 |
| Japanese (speaker1) | 0.52 |
| Japanese (speaker2) | 0.59 |

TABLE V
TIME REQUIRED FOR SYNTHESIZER FINE-TUNING

| Number of Data | Time [minutes] |
|---|---|
| 25 | 4.52 |
| 50 | 5.54 |
| 75 | 6.37 |

significant differences in pronunciation characteristics between the speakers in the training data and the unique katakana English pronunciation of Japanese speakers. This differences is thought not to lead to expected effective conversions. After fine-tuning, this issue was mitigated, with approximately 80% of the words being appropriately converted. However, since the fine-tuning data included only one Japanese speaker, and pronunciations can vary even among Japanese speakers, some words were still not converted correctly. To achieve a level of accuracy usable by Japanese beginners in English, it is essential to increase the training data to comprehensively cover the unique katakana pronunciations of Japanese speakers.

The transformer-based model proposed in this study did not show a decrease in evaluation scores for longer sentences compared to the overall score of the 50 sentences. This result indicates that the proposed method's conversion accuracy is not affected by sentence length. This addresses the issue reported in conventional LSTM-based models where longer sentences are more prone to conversion failures. In the pre-fine-tuning model, the evaluation scores for sentences composed of 11 to 13 words showed a decline for Japanese speakers, which is an acceptable result given that the training data did not include Japanese speakers.

## V. CONCLUSION

This paper first presented an overview of the English pronunciation learning support tool we aim to develop. The tool converts the learner's pronunciation into a target pronunciation. The target pronunciation is designed to retain the learner's voice quality while having the correct accent. For accent conversion, we proposed a transformer-based DNN model based on conventional methods. This study confirmed that fine-tuning with data from one Japanese speaker improved conversion accuracy for the unique katakana English pronunciations of Japanese speakers. Furthermore, switching to a transformer model helped address the issue found in conventional LSTM-based models, in which long sentences often resulted in conversion failures.

However, the conversion accuracy is still not at a satisfactory level, likely due to insufficient training data. Although we have not yet evaluated the retention of speaker characteristics by the synthesizer, it is expected that the characteristics of the speaker's voice, such as pitch, may influence retention.

Future challenges include collecting a large amount of English reading data from speakers whose native language is Japanese, conducting experiments to improve the accuracy of speaker identity retention, extracting phonetic symbols from English speeches, and building the actual English speech learning support tool.

## REFERENCES


[1] *The world' s largest ranking of countries and regions by english skills*, https://www.ef.com/wwen/epi/, (Accessed on 07/02/2024).

[2] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Communication*, vol. 51, no. 10, pp. 920–932, 2009, Spoken Language Technology for Education.

[3] S. Ding, C. Liberatore, S. Sonsaat, *et al.*, "Golden speaker builder – an interactive tool for pronunciation training," *Speech Communication*, vol. 115, pp. 51–66, 2019, ISSN: 0167-6393.

[4] K. Nagano and K. Ozawa, "English speech training using voice conversion," in *Proc. First International Conference on Spoken Language Processing (ICSLP 1990)*, 1990, pp. 1169–1172.

[5] W. Quamer, A. Das, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Zero-Shot Foreign Accent Conversion without a Native Reference," in *Proc. Interspeech 2022*, 2022, pp. 4920–4924.

[6] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[7] D. Povey, A. Ghoshal, G. Boulianne, *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[8] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. 5th ISCA Workshop on Speech Synthesis (SSW 5)*, 2004, pp. 223–224.

[9] G. Zhao, S. Sonsaat, A. Silpachai, *et al.*, "L2-arctic: A non-native english speech corpus," in *Proc. Interspeech*, 2018, pp. 2783–2787.

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