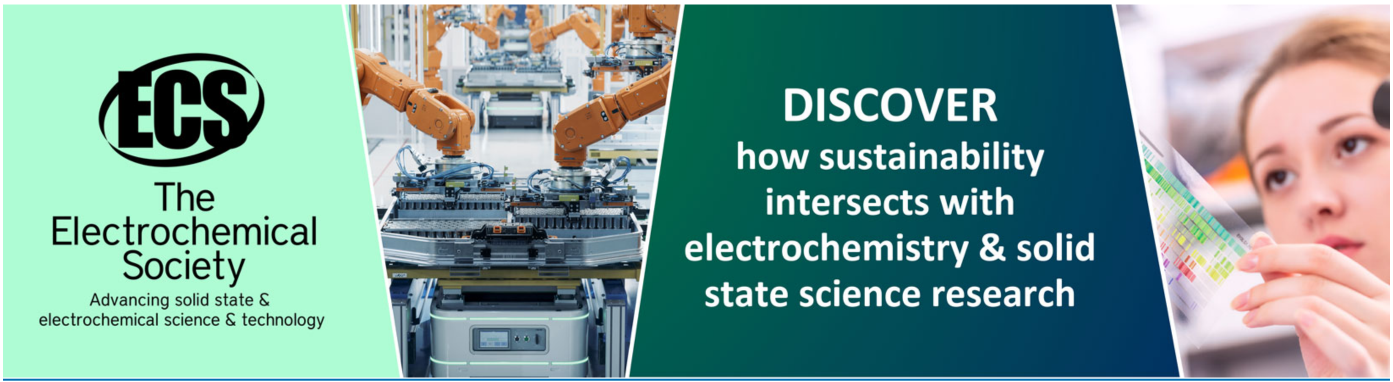
Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

English Pronunciation Quality Evaluation System Based on Continuous Speech Recognition Technology for Multi-Terminal

To cite this article: Xianxian Wu *et al* 2023 *J. Phys.: Conf. Ser.* **2632** 012024

View the article online for updates and enhancements.



English Pronunciation Quality Evaluation System Based on Continuous Speech Recognition Technology for Multi-Terminal

Xianxian Wu[1, *], Yan Zhang[2], Bin Feng[2]

[1] School of Foreign Languages, Taishan University, Taian Shandong China

[2] School of Information Science and Technology, Taishan University, Taian Shandong China

[*] Corresponding author's e-mail address: wuxianxian1980@163.com

Abstract: This paper presents a novel approach for evaluating the pronunciation quality of English speech using continuous speech recognition technology. The research focuses on the application of artificial intelligence in speech recognition, utilizing web browsers on various terminal devices such as computers, mobile phones, and tablets to allow users to read the provided text aloud. The web program captures audio input from the microphone, records it in MP3 format, and uploads it to the server. The server employs the Whisper model to transcribe the audio into semantic text, which is then compared with the displayed text. By calculating the semantic distance and assessing the accuracy of pronunciation, the system provides an evaluation of pronunciation quality, marking correct and incorrect words. To achieve real-time processing, the compact tiny model is employed, and further optimization is performed using Ctranslate 2, resulting in significant performance improvements.

1. Introduction

In the context of globalization, there is a growing demand for learning a second language, especially English, as the number of non-native speakers has already exceeded that of native speakers. The importance of English language proficiency has been increasing, and pronunciation is an essential part of language learning [1, 2]. However, the teaching of oral reading and phonetics is still severely lacking, and evaluating the quality of pronunciation has always been a challenge for language learners and teachers. The main reason for the lack and challenge of pronunciation teaching comes from the difficulty in implementing evaluation and feedback on pronunciation learning [3]. The pronunciation learning in the traditional classroom is greatly influenced by the teacher's pronunciation level. It is also not possible to provide targeted learning suggestions to learners promptly [4]. However, when learners practice on their own, there is a lack of guidance.

With the development of speech recognition technology and Artificial Intelligence [5], it is now possible to evaluate pronunciation quality more accurately and easily. Computer-assisted language learning (CALL) systems have emerged as a solution to this problem. These systems use speech recognition software to provide objective evaluations of a student's speech. They also provide personalized feedback and offer a wider range of educational resources, which are not limited by teacher availability [6]. Additionally, CALL systems can be integrated into various devices, making practicing oral communication less intimidating for students. Compared to traditional language learning methods, CALL systems have several advantages. They offer more objective evaluations and personalized feedback, provide more educational resources, and offer a friendlier learning environment.

This paper proposes an English pronunciation quality evaluation system based on continuous automatic speech recognition technology, which can achieve pronunciation quality evaluation on multiple terminals. The evaluation process includes the pronunciation collection process and the evaluation feedback process, and testers can complete personalized pronunciation quality assessments anytime and anywhere.

2. Technical foundations

This English pronunciation quality assessment system involves the following key components and technologies.

2.1. Acoustic feature extraction

Acoustic feature extraction is a critical step in speech processing. It involves extracting relevant information from the audio signal that captures the characteristics of speech. Commonly used acoustic features include Mel Frequency Cepstral Coefficients (MFCCs), filter banks, pitch, and energy. These features provide representations of the speech signal that can be used for further analysis and modeling.

2.2. Continuous speech recognition (CSR)

CSR is a technology that enables the conversion of spoken language into written text. It utilizes machine learning algorithms, such as deep learning models like recurrent neural networks (RNNs) or transformers, to process and transcribe continuous speech [7]. CSR models are trained on large amounts of speech data paired with their corresponding transcriptions, allowing them to learn the mapping between audio features and textual representations.

Whisper is an attention mechanism-based model developed by OpenAI [8], which uses neural network architecture to selectively focus on the relevant parts of the input signal when processing speech data using attention mechanisms. The operating mechanism of this model is shown in Figure 1, where the sound is converted into a Log Mel spectrogram and transmitted to the encoder, which predicts the text. Compared with traditional speech recognition models, Whisper has faster processing speed, stronger adaptability to accents, more accurate and efficient recognition, and is more suitable for speech-to-text applications.

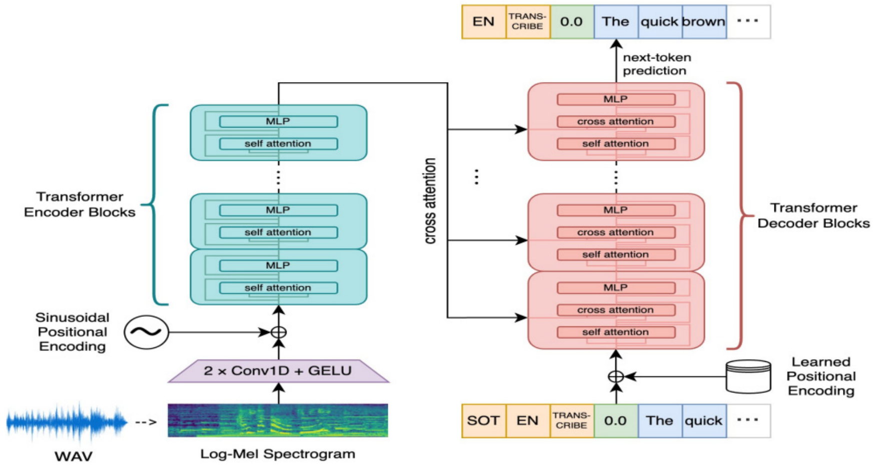


Figure 1. Schematic diagram of Whisper model operation execution process

2.3. Distance metrics

To evaluate pronunciation quality, distance metrics are employed to measure the dissimilarity between reference and test pronunciation. The traditional algorithms first extract the speech features and time positions, use Dynamic Time Warping (DTW) algorithm to align the features and calculate the speech quality scores based on feature distance. This type of algorithm has poor robustness and is also difficult to deploy. The acoustic model based on AI can directly extract high-level semantics from the original audio features and complete the evaluation at the semantic level. As shown in Figure 2, the semantic distance between the recognized text and the target text is calculated using a semantic comparison algorithm to determine pronunciation quality and provide feedback on incorrect words and their positions in the recognized text.

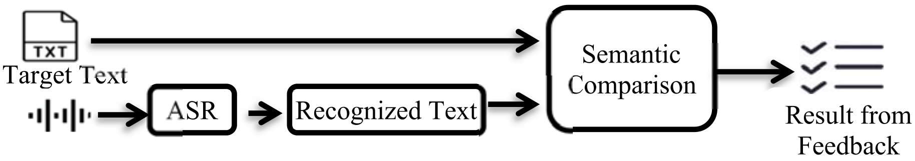


Figure 2. Speech Evaluation Based on Speech Recognition

Semantic comparison usually requires the use of natural language processing algorithms. However, in speech evaluation systems, the vocabulary, grammar, and context of the text are relatively certain, so they can be simplified as string comparisons. LCS algorithm is used to find the longest common subsequence of two texts, which can determine the similarity between two texts.

2.4. Web-based implementation

To enable real-time processing and accessibility across multiple terminal devices, a web-based implementation is often adopted. This involves developing a web application that allows users to record their speech using a microphone. The recorded audio is then sent to the server for processing. The server performs speech recognition using a Whisper model, computes the semantic distance, and provides feedback on pronunciation quality.

3. Physical setup

3.1. User devices

Users interact with the system through their devices, such as computers, laptops, tablets, or mobile phones. These devices should have a functional web browser that supports the necessary web technologies used in the system. The devices should also be equipped with a built-in or external microphone for recording the user's speech. User devices need to be connected to the internet to upload data. A display output, such as a screen or a monitor, is needed too.

3.2. Server

A server is required to handle the processing and evaluation tasks of the system. The server can be a dedicated machine or a cloud-based infrastructure. It needs to have sufficient computing power and memory to run the Whisper model and perform the required computations. The server should also have a reliable internet connection for communication with user devices and data transmission.

3.3. Web browse

A compatible web browser is a crucial component of the physical setup. The web browser enables users to access the system's web application, interact with the user interface, and perform tasks such as text input, speech recording, and result visualization. The web browser should support the necessary web technologies used in the system, including HTML5 and JavaScript for audio recording.

4. System scheme design

The overall structure is shown in Figure 3. The Whisper model runs on the server side. The database system is used to store data such as text and voice. The latest network security requirements require the use of HTTPS protocol for recording operations; the LCS algorithm needs to be implemented to evaluate pronunciation quality.

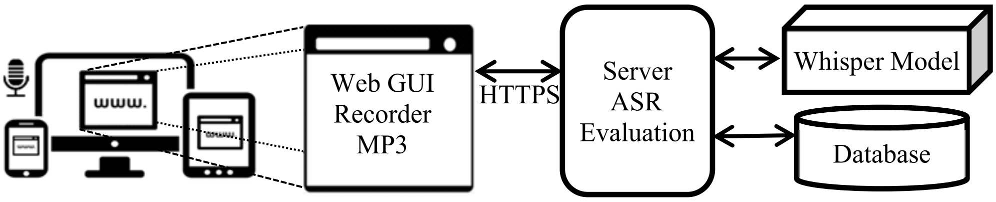


Figure 3. Structural diagram of Speech Evaluation Based on Speech Recognition

The browser side implements functions such as GUI, recording, and connecting to servers. Compressing recorded data into MP3 format can greatly reduce data transmission and achieve real-time recognition.

5. System Implementation

5.1. Server-side

Whisper provides models such as micro, basic, small, medium, and large, which can be directly localized for deployment, saving a lot of model training time. Although larger models have better recognition rates, they take longer and require more memory. To improve the real-time performance, this system chooses the tiny model, and the usage effect is not reduced.

Converting the model to CTranslate 2 format can further improve recognition speed. CTranslate 2 is an OpenNMT library for transformer model acceleration [9]. Meanwhile, using int8 quantization can further reduce time consumption. The experiment shows that after conversion, the time consumption is reduced by about 50% and the speed can be increased by four times.

We evaluate pronunciation quality by using the LCS algorithm to calculate the similarity between texts [10, 11]. The basic idea of the LCS algorithm is Dynamic programming. It involves constructing a two-dimensional table to determine the length of the longest common subsequence. The algorithm fills the table iteratively and calculates the length of the longest common subsequence, which is the number of correctly pronounced texts. Find text with correct pronunciation through backtracking.

5.2. Browser side

In the browser, the HTML5 Media Recorder API can obtain audio data in real time from the microphone. PCM format data can be recorded in its data available event processing function. However, due to the large size of PCM data, it is necessary to use the LAME [12] library to compress it into MP3 format. The LAME library is in C++ format by default, and to facilitate JavaScript processing, it needs to be compiled into the Web-Assembly module. Finally, place the data in the blob for transmission.

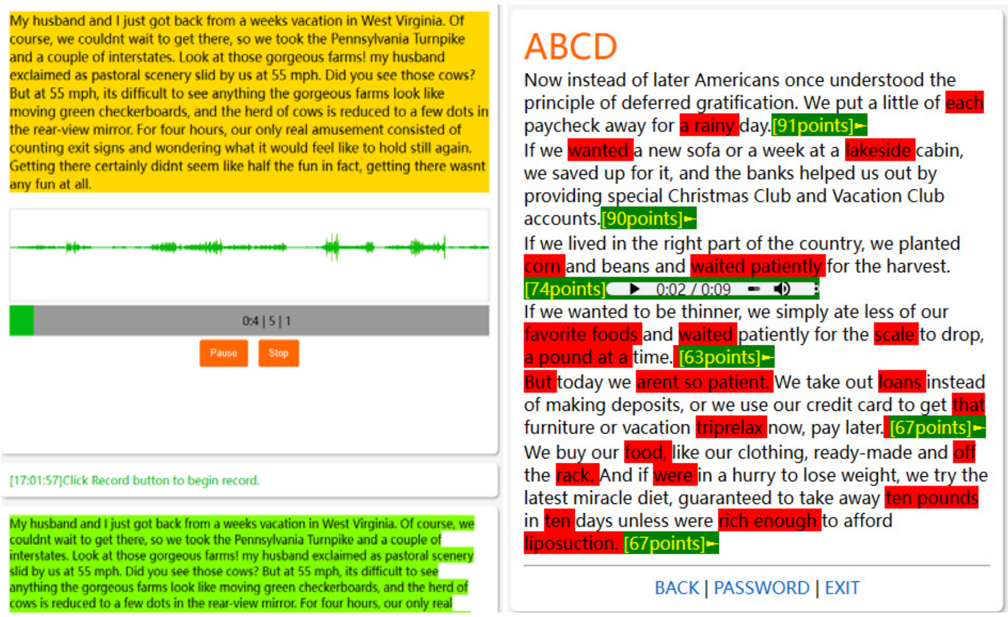
5.3. Data transmission

The browser displays the test text, records the pronunciation, uploads the data through XMLHttpRequest [13], and then receives the evaluation information. After receiving the MP3 data, the server decodes it into WAV and submits it to the speech recognition model for processing. At the same time as providing feedback to the browser, the record is saved in the database.

6. System testing

6.1. Browser side testing

On the left side of Figure 4 is the interface for testers to read text and record sound, while on the right side is the interface for viewing evaluation results. From the final effect, the recording responds quickly. The feedback time varies depending on the length of the text and meets the real-time requirements. Both correct and incorrect texts can be marked, and the score matches the actual pronunciation quality.



My husband and I just got back from a weeks vacation in West Virginia. Of course, we couldnt wait to get there, so we took the Pennsylvania Turnpike and a couple of interstates. Look at those gorgeous farms! my husband exclaimed as pastoral scenery slid by us at 55 mph. Did you see those cows? But at 55 mph, its difficult to see anything the gorgeous farms look like moving green checkerboards, and the herd of cows is reduced to a few dots in the rear-view mirror. For four hours, our only real amusement consisted of counting exit signs and wondering what it would feel like to hold still again. Getting there certainly didnt seem like half the fun in fact, getting there wasnt any fun at all.

[17:01:57]Click Record button to begin record.

My husband and I just got back from a weeks vacation in West Virginia. Of course, we couldnt wait to get there, so we took the Pennsylvania Turnpike and a couple of interstates. Look at those gorgeous farms! my husband exclaimed as pastoral scenery slid by us at 55 mph. Did you see those cows? But at 55 mph, its difficult to see anything the gorgeous farms look like moving green checkerboards, and the herd of cows is reduced to a few dots in the rear-view mirror. For four hours, our only real

ABCD

Now instead of later Americans once understood the principle of deferred gratification. We put a little of each paycheck away for a rainy day.[91points]⇤

If we wanted a new sofa or a week at a lakeside cabin, we saved up for it, and the banks helped us out by providing special Christmas Club and Vacation Club accounts.[90points]⇤

If we lived in the right part of the country, we planted corn and beans and waited patiently for the harvest. [74points]

If we wanted to be thinner, we simply ate less of our favorite foods and waited patiently for the scale to drop, a pound at a time.[63points]⇤

But today we arent so patient. We take out loans instead of making deposits, or we use our credit card to get that furniture or vacation triprelax now, pay later.[67points]⇤

We buy our food, like our clothing, ready-made and off the rack. And if were in a hurry to lose weight, we try the latest miracle diet, guaranteed to take away ten pounds in ten days unless were rich enough to afford liposuction. [67points]⇤

Figure 4. Browser interface

6.2. Server side testing

Table 1 shows the comparison of the max memory and time consumption when using the OpenAI library and Ctranslate 2 library to load tiny and basic models to recognize the same 30 seconds of speech. The test program is executed with 4 threads on an Intel (R) Core (TM) i5-3470 CPU. It can be seen that the combination of the Ctranslate 2 library and the tiny model has fewer resources and performs better.

Table 1. Model optimization comparison

Model	OpenAI Model Size	CTranslate 2 Model Size	OpenAI Max Memory	CTranslate 2 Max Memory	OpenAI Cost Time	CTranslate 2 Cost Time
Tiny	72 MB	74 MB	544 MB	232 MB	10.3 s	5.5 s
Base	138 MB	147 MB	743 MB	338 MB	23.3 s	8.6 s

7. Conclusion

The English pronunciation quality evaluation system based on continuous speech recognition technology allows users to evaluate their pronunciation quality on various devices such as computers, mobile phones, and tablets. This multi-terminal approach enhances accessibility and convenience, enabling users to practice pronunciation anytime and anywhere. The system utilizes the Whisper model to recognize speech and provides real-time feedback on pronunciation quality by calculating semantic distance and identifying incorrect words. This immediate evaluation allows users to receive instant feedback and make necessary corrections, enhancing the efficiency and effectiveness of language learning. The focus of this study is to achieve real-time processing through the use of the tiny model and to further optimize it using Ctranslate 2. This optimization significantly improves the system's performance in terms of processing speed and recognition accuracy.

The system has important practical implications for English language learning. It addresses the challenges in traditional pronunciation teaching methods by providing objective evaluations, personalized feedback, and a friendly learning environment. The system can offer guidance and support to learners even in the absence of a teacher, enhancing their learning experience and progress.

It suggests the possibility of extending the system's language evaluation capabilities to languages other than English. Additionally, it encourages further research on optimizing models and improving system performance, paving the way for continuous advancements in speech recognition-based language learning systems.

Acknowledgments

The study is supported by the Teaching Reform Research Project of Taishan University (No. JG202173).

References

[1] Liu N. (2022). Automatic English Pronunciation Evaluation Algorithm Based on Sequence Matching and Feature Fusion. Mathematical Problems in Engineering. doi:10.1155/2022/4785355.

[2] Khan R. M. I., Kumar T., Benyo A., Jahara S. F. (2022). The Reliability Analysis of Speaking Test in Computer-Assisted Language Learning (CALL) Environment. Education Research International. doi:10.1155/2022/8984330.

[3] Li H. X. (2022). The Influences and Improvement Strategies of the Use of Accurate English Phonetics on English Listening for Chinese University Freshmen. World Journal of Educational Research(5). doi:10.22158/WJER.V9N5P111.

[4] John A. (1992).GERALD KNOWLES. Patterns of spoken English: An introduction to English phonetics. WORD(1). doi:10.1080/00437956.1992.12098290.

[5] Van A., Dinh S. L., Ha H. H. (2023). Improving the Accuracy of Speech Recognition Models for Non-Native English Speakers using Bag-of-Words and Deep Neural Networks. Scientific Review(2). doi:10.32861/SR.91.10.14.

[6] Nedjah N., Bonilla (2023). Automatic speech recognition of Portuguese phonemes using neural networks ensemble. Expert Systems With Applications. doi:10.1016/J.ESWA.2023.120378.

[7] Wang S. (2023). Recognition of English speech – using a deep learning algorithm. Journal of Intelligent Systems (1). doi:10.1515/JISYS-2022-0236.

[8] OpenAI Whisper. GitHub. https://github.com/openai/Whisper

[9] OpenNMT/CTranslate2. GitHub. https://github.com/OpenNMT/CTranslate2

[10] Chen Y. (2022). English Translation Template Retrieval Based on Semantic Distance Ontology Knowledge Recognition Algorithm. Mathematical Problems in Engineering. doi:10.1155/2022/2306321.

[11] Kenawy T., Abdel R. M., Bahig H. M. (2022). A Fast longest crossing-plain preserving common subsequence algorithm. International Journal of Information Technology(6). doi:10.1007/S41870-022-01038-0.

[12] The LAME Project. Sourceforge. https://lame.sourceforge.io

[13] Xian J. C., Xin Y. L., Yong S. Z. (2012). Design of WebGIS Application Based on Ajax. Advanced Materials Research(542-543). doi:10.4028/www.scientific.net/AMR.542-543.1282.