# IELS: A Computer Assisted Pronunciation Training System for Undergraduate Students

Jinyu Chen[1, 2], Lan Wang, Chongguo Li, Jin Hu, Sheng Li
[1] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[2] The Chinese University of Hong Kong, Hong Kong, China
{jy.chen, lan.wang, cg.li}@siat.ac.cn, {jin.hu, sheng.li}@sub.siat.ac.cn

***Abstract*—IELS, which abbreviates Interactive English Learning System, is a computer assisted pronunciation training (CAPT) system for Chinese learners of English whose mother language is Mandarin. The system provides instant feedback of mispronunciations of phoneme, word, lexical stress, and a score of the student's overall pronunciation quality. The system employs client-server architecture, in which the client provides friendly interface and audio I/O function, the server takes charge of speech processing, including HMM-based speech recognition, SVM-based lexical stress detection and speech scoring. IELS has been used by 52 freshmen in a Chinese University. A questionnaire survey result among a class of 31 students in the university shows that the system helps to improve their pronunciation.**

***Keywords-computer assisted pronunciaiton training, speech recognition, lexical stress detection, speech scoring***

## I. INTRODUCTION

IELS (Interactive English Learning System) is a computer-assisted pronunciation training (CAPT) system developed for undergraduate students. English pronunciation is still an obstacle for Chinese undergraduates even though they may have learned English for more than six years. Because of various dialects, limited English teaching resources, most of the undergraduate students have difficulties in oral English presentation. As the second language, English pronunciation is not easy for Chinese students due to the differences between two languages. First of all, Mandarin Chinese and English use different phoneme inventories. There are vowels and consonants for English pronunciations, but Chinese pronunciation is comprised of initials and finals. Some phonemes that are present in English are missing in Mandarin Chinese inventory. If a Chinese student could not pronounce the phoneme correctly; the pronunciation of the word would be surely incorrect. Second, the letter to sound rules for English is quite different from Chinese Pinyin. When a Chinese student encounters an unfamiliar word, he may not use the letter to sound rules, but guess the pronunciation by Chinese Pinyin. Last but no least is the different rhythm of the two languages. As a stress-timed language, there is perceived to be a fairly constant amount of time between consecutive stressed syllables in English pronunciation. To the contrary, Chinese is a syllable-timed language, in which every syllable is perceived as taking up roughly the same amount of time. Stress is critical to express emphasize, rhythm and intonation in English, but syllable-timed languages tend to give syllables approximately equal stress. So, new to the stress-timed language, few Chinese students can speak English with correct stress. Even though there are many difficulties, Chinese students can also improve their pronunciation if they have change to practice with native speakers. The key is that students are lack of instant feedbacks in high schools. Most of high schools focus on reading and writing ability, such as vocabulary, grammar, etc. Besides, the ratio between English teacher and student is 1 to 50; an English teacher has to take care of a whole class, and it is impossible for him/her to check the mispronunciation of every student. Furthermore, even if the teachers spare time to talk to students in English, they may not pay much attention to phoneme mistakes, but to the overall expressive ability.

Computer Assisted Pronunciation Training (CAPT) system provides a solution for student oriented pronunciation training. Similar systems include SPELL [1], SRI EduSpeak System [2], ISLE system [3], and PLASER system [4], etc. EduSpeak System gives a score of user speech. ISLE system detects the mispronunciation type and position. PLASER system provides assessment of phoneme and visualization of recognition results. Even though the above systems could assist the learners to improve pronunciation by a straightforward score of a phoneme or sentence, they are not clear enough to point out mistake type and methods to correct mispronunciation. In the paper, we present a system for undergraduate students to practice English pronunciation, which can detect mispronunciation at the level of phoneme, lexical stress, and give statistical analysis of frequently mispronounced words and phonemes. Our system also provides a module to score the speech, which is useful for self evaluation when undergraduate students need to take some oral English tests.

The system employs a client-server architecture, in which the client provides friendly interface and audio I/O function, the server takes charge of speech processing, including HMM-based speech recognition, SVM-based lexical stress detection and speech scoring. The server also provides statistical analysis of registered user's learning progress. The function modules in the client interact with the server for various services through communication message in XML format based on the TCP/IP socket. IELS has been used by 52 freshmen in a Chinese University for three months and a questionnaire survey for a class of 31 students in the univers-

978-1-4244-6370-1/$26.00 © 2010 IEEE

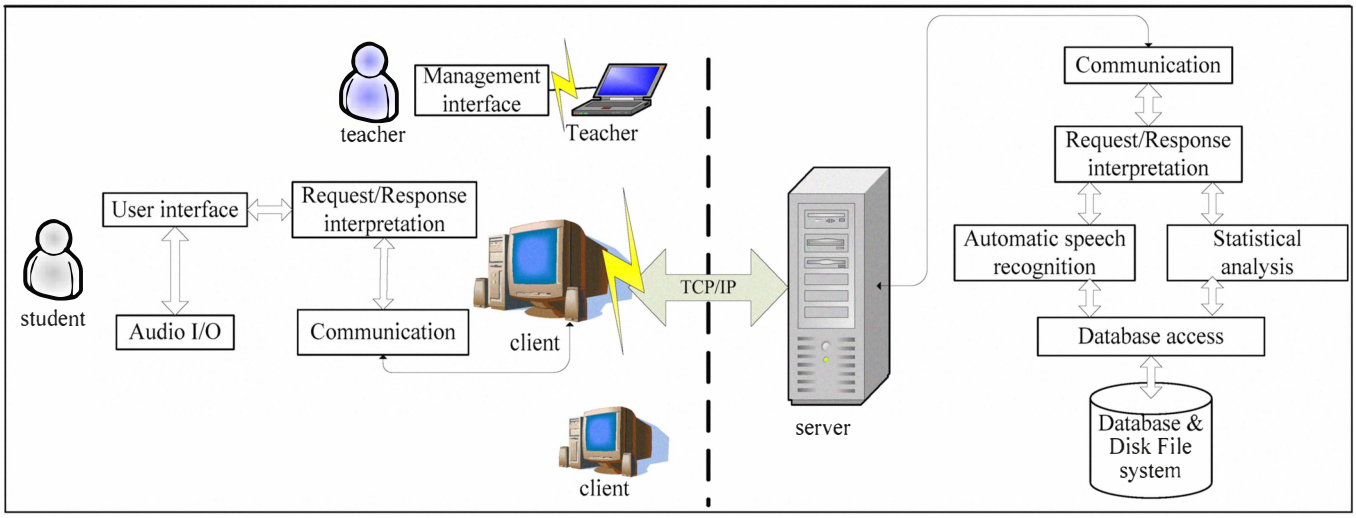


Figure 1. Overview of the IELS system architecture

ity shows that the system does improve their pronunciations.

The paper is organized as follows: in the second section, we present the architecture of the system, which is followed by a detailed design of function modules of the system in section III. Section IV describes the speech recognition technologies used in the IELS. Then the experiments and evaluations are described in V. Finally, the conclusion and future work.

## II. THE SYSTEM ARCHITECTURE

Figure 1 illustrates the IELS system architecture. The system employs client/server structure and clients and server communicate through TCP/IP socket. The client provides friendly user interfaces (UI) and audio Input and output (I/O) for users. The user interfaces include user information management UI, pronunciation training UI, and statistical analysis UI. As the main UI of the client, the pronunciation UI will be described in detail in Section III. The audio I/O includes the playing of reference speech, recording of user speech and pre-processing of audio before sending it to server for further processing. The clients and server communicate through message in Extensible Markup Language (XML) format base on TCP/IP socket. So both the client and server include request/response message interpretation module. The client focuses on providing friendly user interface, while the server saves all the learning resources and provides speech processing results. The modules in server include: (1) Resource storage, including learning materials such as audios, pictures, etc., user practice records and other needed log files; (2) Automatic speech recognition, including mispronunciation detection at phoneme level, lexical stress detection and correction, and speech scoring; (3) Statistical analysis, which provides analysis results from practicing records, such as most frequently mispronounced phonemes and words. (4) request/response interpretation and communication. The system also provides a special client terminal for teachers to manage the English learning materials in the server handily.

Before we develop the C/S architected IELS, we have developed a similar stand-along-system. The system performance is not desirable. The IELS is designed with client-server (C/S) architecture for the following considerations. Automatic speech recognition is computing consuming, which demands large amount memory and CPU resource. So we deploy a server to handle the speech processing computation, and the client for I/O and display. The client is simple to be installed in PC without extra software or libraries. In addition, the distributed client is efficient to balance the burden of the server by distributing part of speech pre-processing into the client. Moreover, user access is easy to control through client terminal, such as user access authority, user information management. Furthermore, the C/S architecture is convenient for data update. The model and algorithm of speech recognition can be easily updated in the server, without changes to the clients. The feasibility of system implementation requires the system to be well modularized with low coupling and good extensible. The functions modules of client and server in IELS are independent and the interfaces between modules are simple and the XML message is good at extensibility.

The server is placed in the administrative office and runs in a HP server under Ubuntu 8.04 for 24 hours every day. The client terminal software is installed in the computers in class room. The clients and server are connected through the intranet campus network. The database is MySQL5.0. The programming language in server is C++, however, in order to design and develop friendly user interface, C# based on Visual Studio is explored in the client software. The client runs in PC under Microsoft XP equipped with headset and microphone.

## III. THE FUNCTION MODULES OF THE IELS SYSTEM

### A. *Function modules in client*

The client includes user interface, audio I/O, message interpretation and communication.

#### 1) *User interface*

*a) User information management UI*

New user can register, login, logout, and update their profile. The system can be used in two modes, registered

users and guest. The practicing record is kept only for registered user so the statistical and progress analysis module is unavailable for guest.

*b) Pronunciation training UI*

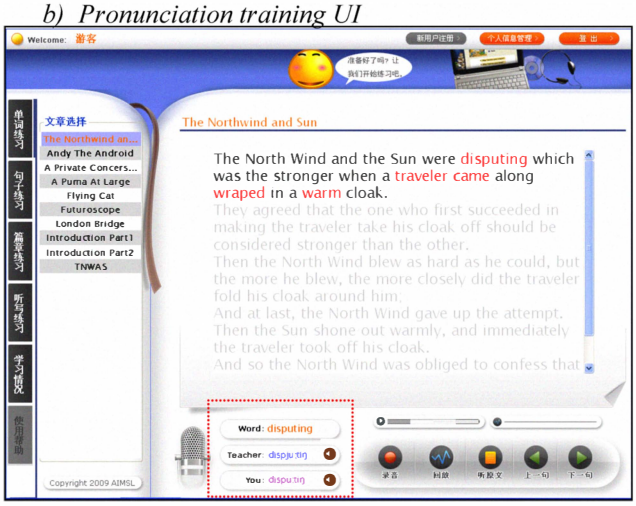


Figure 2. Practice UI

The Figure 2 shows the pronunciation training UI. As the main user interface of the client, the pronunciation training interface includes three features. First, there are three levels of practicing materials, words, sentences and paragraphs so that users can practice from simple words without context to paragraph. Second, all text, audios, and pictures are displayed dynamically in accordance with record in the server. To guarantee the instant response, the client downloads sources according to request. The client first gets the titles of all practicing materials, and then downloads the specific audios and references from the server when the client clicks the title of an exercise. The third is feedback visualization. After the user reads the prompts and records it, the client will visualize the feedback from the server, telling the mispronunciations and score by four highlights in the feedback visualization.

(1) The mispronounced words are highlighted in red, so that users can quickly find them.

(2) When clicking word in the red, the detailed information will display in the red dashed rectangle in Figure 2. The user can listen to the standard word pronunciation segmented from the whole sentence and her pronunciation as well.

(3) If the word is mispronounced, the user can compare the phoneme segments of "Teacher" and "You" to find out the difference.

(4) The overall score of the sentence is calculated.

*c) Statistical analysis UI*

The registered users can see the statistical analysis of their learning progress, for example the progress of average scores, the most frequently mispronounced phonemes and words according to their past practicing records. The statistic analyses are draw from analytical algorithm in the server based on user's practice record. The following Figure 3 shows the most frequently mispronounced phonemes for statistical analysis.

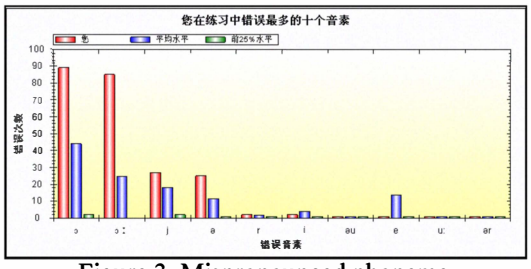


Figure 3. Mispronounced phoneme

*2) Audio I/O*

The client can play the reference speech recorded by native speakers while recording the student's speech. Moreover, the audio I/O also has the following functions: (1) Silence should be refused from acoustic feature extraction. (2) Noise, which is not regarded as input speech in accordance with the required prompts. The noise should be refused before mispronunciation detection. (3) Acoustic feature extraction. Due to the bandwidth and communication burden, the extracted acoustic features, rather than original waveform, are sent to the server for ASR because the size of feature file is much less than that for original waveform. After recording the waveform, the client should also extract acoustic features and then send them to the server for recognition. The above are pre-processing for recorded audio. The silence detection and acoustic feature abstraction are finished at client, and the noise refusal is computed in server at present. In the following section, we will state these speech technologies in detail.

*3) Request/response interpretation*

XML is known as a powerful format to define business logic, so the message is passed between the client and server in the format of XML. The message is labeled with operation type and descriptive semantic tags characterizing the arguments to pass which are agreed for both client and server terminals. Once the operation in client involves interaction with server, the client will send a XML message to the server, and expect a response. For example, the format of the speech recognition message for request and response messages are shown in the Figure 4.

Request format:
<recog><uid>guest@siat.ac.cn</uid><cont> ... </cont>
<type> sen </type> <fea> ... </fea> </recog>
Response format:
<recogres> <success>yes/no</success>
<score>86</score> <word>...</word> </recogres>

Figure 4. XML message format

The request message should contain the client's identification information and data required for this request, the response message contains the success of the request in the tag <success> and other data returned for this request.

*B. Function for system administrator or teacher*

The system is designed with good extendibility for English teachers. The learning materials can be easily uploaded on the system by teachers without extra changes to client and server software through management interface. The teacher only needs to prepare for the required files: native speech, text, reference transcription, etc. The request

XML message and materials will be uploaded to the server through administrator management interface. The Server interprets the message, receives the package, and finally, inserts the table in database automatically. After updating, the client can display the updated materials. This module enables the system open to new materials and allows the uploading conveniently.

### C. Function modules in server

The server is based on the socket listen program; once it receives a request from client it starts up a new thread and end the thread when the response is sent back.

#### 1) Message interpretation and response generation

The fast simple C++ XML parser CMarkup is used in the server software to handle interfaces with XML. The message interpretation follows the same format with client, which stated in the above section. The server must first get the message type to call the predefined function blocks.

#### 2) Resource storage in the Server

Learning materials and recordings of users are stored in the server through two methods: relational database and physical file system, whose directory is kept in database. There are diverse exercise materials and reference, including exercise in three levels: minimal pair, sentence, and passage. The contents are stored in records of database with unique identification number. The audio, video and text of learning materials are stored in the disk. User information is also stored in tables of the database. The record of user practice is stored in the database as original table and views for statistical analysis. Once receiving a request from client, the server will assess the database through the database access module.

#### 3) Automatic speech recognition

As the core function module of the system, the automatic speech recognition includes three blocks: mispronunciation detection at phoneme level, lexical stress detection and correction, and speech overall scoring. In addition, the noise rejection is also implemented in this part. The details of technologies will be explained in the following section. When the client requests ASR service of a sentence, the server will first find the resources, for example standard word labels, references stored in the disk by searching database with the identification number in the request message. Then the recognizer will be called to recognize the audio, followed by noise rejection. If the audio is appropriate for recognition, phoneme-level mispronunciation detection follows. If the word contains more two syllables, the lexical stress detection and correction module will be called. At last, the speech scoring module will be called to give a score based on the results of mispronunciation of phonemes and lexical stress, combined with other features. Afterwards, the reorganization results are kept in database, and the mispronounced phonemes and words are also summarized as records inserted in the tables, for statistical analysis.

Because automatic speech recognition is computing consuming, we optimize the recognition process to speedup the process. All of the shared resources, such as models, dictionaries, etc. are loaded when the server starts to work. Each client request is handled in a thread, and threads run in parallel.

#### 4) Statistical analysis

Based on records in the database, the server provides statistical analysis for registered users. Besides the statistical results of the user, the server also provides comparative analytical results, such as the average results of all registered users, and the result of the best record in the database for reference. By now, the system provides two types of analysis, intra-user analysis and inter-users analysis. The intra-user analysis is the line of day average score, which is the average value of all the practices scores of one day. The inter-user analysis includes the first ten most frequently mispronounced phonemes and words respectively.

#### 5) Communication of audio

Besides text message communication in XML, the server still needs to collection some audios from the clients for adaptation. Though only acoustic feature files are passed to the server for processing, there are two conditions that we want to collect the original speech in waveform. First is to improve system performance (increase recognition accuracy). The extended pronunciation dictionary should cover all mispronunciation variants, but the current version may be not complete, so the collected speech will help to build a comprehensive dictionary. Second, speeches that are regarded as refusal ones should be uploaded to server for human analysis. By collecting refused speech, the system can be improved to distinguish bad audio from verified speeches.

## IV. THE SPEECH RECOGNITION TECHNOLOGIES

Figure 5. shows the flowchart of the speech processing in IELS. The silence and acoustic feature extraction are finished in the client terminal. The rest speech processing are finished in server. The entire speech recognition system is implemented using the HTK toolkit [5] from Cambridge University. The HMM-based recognizer uses acoustic models based on 47 phonemes, including silence and short speech pause. The section will explain the involved details.

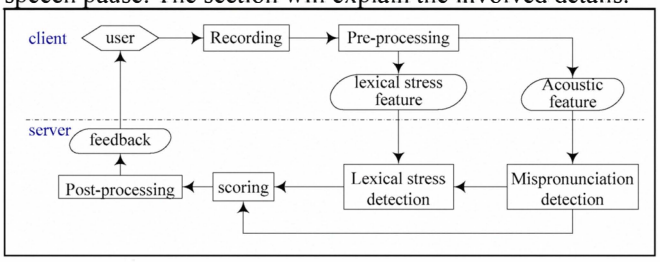


Figure 5. Flowchart of speech processing

#### 1) Pre-processing

The pre-processing of speech in the client includes silence detection, acoustic feature extraction and noise rejection. Various thresholds of energy are used to judge silence. The perceptual linear predictive (PLP) parameter for speech recognition, pitch and power for lexical stress detection, are extracted from the input waveform. There are two steps to judge whether the speech is noise or not. The first step checks the number of recognized words. If the number is much less than that of the transcription, the speech

is regarded as noise or incomplete speech. Then we will compare the average posterior score in a phone sequence with a threshold, the input with higher scores will be regard as verified speech and forward for mispronunciation detection, otherwise, it will be rejected.

*2) Mispronunciation detection of phonemes*

Based on language transfer effect, phonetic confusions of phonemes can be used to generate additional, erroneous pronunciation variants for word, and the extended pronunciation dictionary is used to produce confusion network in recognition. [6, 7, 8] The extended pronunciation dictionary is designed to cover all the predicted mispronunciation variations. With the dictionary and the acoustic model, the recognition hypothesis can be obtained by Viterbi decoding algorithm. Comparing with the correct reference transcription, the differences (mispronunciations) can be pointed out using dynamic programming algorithm.

*3) Lexical stress detection and correction*

The lexical stress refers to a syllable perceived as standing out from its environment in the form of primary or secondary lexical stress. The position of stressed syllable in a word is called stress pattern. In our system, word, rather than syllable is used as unit for lexical stress pattern detection. The feature of word is a composition of features of syllables in the words, and the feature of the current syllable is differential frame-averaged energy, pitch and duration over the preceding and following syllables. A support vector machine (SVM) is trained to predict the lexical stress pattern of polysyllabic words. At last, the correctness of lexical stress is given by comparison with reference.

*4) Automatic pronunciation scoring*

With the mispronunciation results at phoneme level and lexical stress at word level, posterior scores of HMM acoustic model, and features that can demonstrate the fluency, accuracy of phoneme, prosody and the completeness of speech [9], an overall score is calculated to evaluate the quality of input speech with reference speech. An SVM model is also trained for scoring.

## V. EXPERIMENTS AND EVALUATIONS

A beta version of IELS was tested by 52 freshmen over a period of 3 months in Shenzhen University. The practicing materials includes 50 groups of minimal pairs (a total of 206 words) to train the single phoneme, 200 sentences to train the fluency, and 9 passages to train reading in a context. A questionnaire survey was conducted to gauge the effectiveness of using IELS to learn English pronunciation, and to get comments and suggestions from teachers and students. We carried out the survey through a class of 31 students, and all of the questionnaires were collected and analyzed.

- All of the students believed that their pronunciations were improved after using IELS.
- 94% of the students are satisfied with the user interface, 70% of the students are satisfied with the system's stability and operability.
- 77% of the students think the mispronunciation detection is accurate.

Besides, the teacher also thinks that the system is helpful to undergraduate students, and would recommend her students to use IELS to train English pronunciation.

## VI. CONCLUSION AND FUTURE WORK

This paper describes the design and development of IELS, an English pronunciation training system for undergraduate students of China. We have explored the C/S architecture with speech technologies to assistant students to practice English pronunciation. The system wins good evaluation in a Chinese university.

Future works are planned in three aspects. Firstly, we would add more learning materials for practice. With the help of English teacher, more materials relevant to student's textbooks and their test requirements will be uploaded to the IELS. Secondly, we would work on the improvement of system performance. The accuracy of phone-level mispronunciation detection can be improved by including more variants into the extended pronunciation dictionary. The collected user speech, after being annotated by language experts, would be valuable source to generate mispronunciation variant to extended dictionary. We will also modify the acoustic model to make it more discriminative of phoneme. Thirdly, we also plan to develop a browser-server (B/S) architecture system, which will be more accessible to language learners.

## ACKNOWLEDGMENT

The work is supported by National Science Foundation of China (NSFC: 60772165).

## REFERENCES

[1] S. Hiller, E. Rooney, J. Laver, and M. Jack, "An automated system for computer-aided pronunciation teaching. Speech communication," Speech communication, Vol. 13, Dec. 1993, pp.463-473.

[2] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, and J. Butzberger, "The SRI EduSpeak(TM) System: Recognition and Pronunciation Scoring for Language Learning," Proc. Integrating Speech Technology in Language Learning (InSTILL),2000 , pp.123-128.

[3] M. Menzel, D. Herron, P. Bonaventura, and R. Morton, "Automatic detection and correction of non-native English pronunciaitions," Proc. Integrating Speech Technology in Language Learning (InSTILL), 2000.

[4] B. Mak, M. Siu, Ng. Mimi, Y.C. Tam, Y.C. Chan, K.W. Chan, K.Y. Leung, S. Ho, F.H. Chong, J. Wong, and J.Lo, "PLASER: Pronunciation Learning via Automatic Speech Recognition", Proc. HLT-NAACL 2003, pp.23-29.

[5] S. Yong, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, et al., HTKBook, Cambridge University Engineering Dept.

[6] H. Meng, Y.Y. Lo, L. Wang and W.Y. Lau, "Deriving salient learners mispronunciations from cross language phonological comparison", Proc. ASRU 2007, 2007, pp.437-442.

[7] L. Wang, X. Feng, and H. Meng, "Mispronunciation detection based on cross-language phonological comparisons," Proc. ICALIP2008, 2008, pp.307-311.

[8] L. Wang, X. Feng, and H. Meng, "Automatic generation and pruning of phonetic mispronunciations to support Computer-Aided Pronunciation Training", InterSpeech 2008.

[9] T. Cincarek, R. Gurhn, C. Hacker, and E.T Nöth, "Automatic pronunciation scoring of words and sentences independent from non-native's first language", Computer Speech and Language, 2009.