A prototype of an adaptive Chinese pronunciation training system

Hsien-Cheng Liao, Ying-Hua Guan, Jia-Jang Tu, Jiang-Chun Chen

Article info
Article history:
Received 5 February 2013
Received in revised form 14 April 2014
Accepted 23 April 2014
Available online

Keywords:
Computer-assisted language learning
Computer-assisted pronunciation training
Automatic speech recognition
Pronunciation assessment
Tone assessment

ABSTRACT

Many beginning-level Chinese learners, particularly those whose mother tongue is not a tone language, find it challenging to learn Chinese pronunciations and tones. Yet pronunciation practice in formal language class settings is limited. Therefore, it has become necessary to develop computer-assisted and adaptive training systems to help learners practice Chinese pronunciation outside of class. This article reports on the development and verification of a prototypical adaptive Chinese pronunciation training system that specifically focuses on pronunciation errors related to aspiration, retroflexion, and tones. We describe how we built and trained the pronunciation error detection system, and how the system detected pronunciation errors and determined the timing of pronunciation exercises. The verification results of the system’s performance demonstrated that it achieves a relatively high accuracy rate in detecting pronunciation errors and it has potential to help learners overcome specific weaknesses in Chinese pronunciation. We conclude with suggestions for further research on and development of an adaptive Chinese pronunciation training system.

1. Introduction

Many beginning-level Chinese learners find that learning Chinese pronunciation and tones is a difficult task, which requires a great deal of time and effort outside of formal instructional classes (Guo & Tao, 2008; Ross, 2001). In a learning environment where learners rarely have a chance to practice speaking with native Chinese speakers, a Chinese computer-assisted pronunciation training (CAPT) system could be helpful. However, current pronunciation training systems on the market for Chinese learning cannot actually meet learners’ needs. Most of the systems provide scores instead of the corrective feedback that is given by teachers in a traditional class setting. Learners usually do not know how to improve their pronunciation if they receive only scores on their performance. Therefore, a CAPT system providing remedial feedback is a much more helpful tool for learners. In this article, we discuss the difficulties that arise in learning Chinese pronunciation, explain how speech technology has contributed to the design of a pronunciation training system that is able to provide corrective feedback, and describe how this prototype of an adaptive Chinese pronunciation training system addresses learners’ needs.

* Corresponding author. Tel.: +886 2 2363 6573x11.
E-mail addresses: hcliao@itri.org.tw (H.-C. Liao), yhguan@ntnu.edu.tw, profyhguan@gmail.com (Y.-H. Guan), santu@itri.org.tw (J.-J. Tu), jtcchen0901@itri.org.tw (J.-C. Chen).
1 The term “Chinese” used throughout this paper only refers to Mandarin Chinese (or Putonghua).
2. Literature review

2.1. The components of Chinese syllables

Understanding where certain Chinese pronunciation errors occur frequently requires understanding the structure of a Chinese syllable. Duanmu (2007) suggests that a syllable in standard Chinese should at least contain a nucleus that may consist of a vowel or a diphthong. Some syllables may have an onset in the beginning of the syllable and/or a coda after the nucleus. Furthermore, “the maximal size of a syllable in Standard Chinese is either CGVV or CGVC, where C is a consonant, G is a glide, and VV either a long vowel or a diphthong.” (Duanmu, 2007, p. 71). The syllable [kwan], for example, has a CGVC structure, and the syllable [kwai] has a CGVV structure. Researchers, however, have different opinions on whether C, the medial glides (i.e., [j], [w], [ɥ]), should belong to the initial or the final (for more information, see Duanmu, 2007, pp. 79–81). Duanmu (2007) contends that CG shares one onset slot that corresponds to the initial sound of a syllable. Nevertheless, Norman (1988) suggests that

the initial is the consonantal onset to the syllable; in both medieval and modern Chinese this always consists of a single consonant. The final is the remainder of the syllable minus the tone; it may be further broken down into a medial, a main vowel and an ending; of these elements only the main vowel is obligatory. (pp. 138–139)

In the field of Chinese language pedagogy, all the Chinese language textbooks (e.g., He, Jiao, Shao, & Livaccari, 2008; Liu, Yao, Bi, Ge, & Shi, 2009) consider the medial glides a part of the final, not a part of the initial,2 which is in line with the definitions suggested by Norman (1988). Therefore, the syllable [kwan] is said to have an initial [k] and a final that consists of a medial glide [w] and a rhyme [an]. Following the ways for classifying Chinese syllable structures in these Chinese language textbooks, syllables that only consist of a final have a structure of (G)V(X3) and may appear in one of the following forms:

- V, a single vowel (e.g., [a], [u])
- VV, a diphthong (e.g., [ai], [au])
- VC, a vowel and a consonant (i.e., the coda, e.g., [an], [in])
- GV(X), a medial glide followed by one of the three syllable-forms just mentioned (e.g., [ja], [wai], [qan]).

Moreover, syllables that comprise an initial and a final are, for example, [fan] and [tsʰan].

In addition to initial and final, another important component of a Chinese syllable is its tone. Duanmu (2007) states that “most Chinese words are stressed syllables…. Stressed syllables carry lexical tones and are longer than unstressed syllables, which do not carry lexical tones” (p. 72). Chao (1948) describes four lexical tones for stressed syllables and a neutral tone for unstressed syllables. He uses a five-point tonal scale (1 = low, 2 = half-low, 3 = middle, 4 = half-high, 5 = high) to describe the pitch of the four tones. The first tone is high-level (tonal scale value 55), the second is mid-rising (tonal scale value 35), the third is low-dipping (tonal scale value 214), and the fourth is high-falling (tonal scale value 51). It is noteworthy that the third tone is actually pronounced as the half third tone (tonal scale value 21) when it is followed by another tone other than the third tone. The neutral tone (also called the fifth tone) is used when a syllable is pronounced in a short and unstressed manner (e.g., it could occur in the morpheme at the end of a sentence, in morphemes with multiple syllables, or in interjections). Its pitch value varies with that of the preceding adjoining syllable and can be pronounced in many different ways (Chao, 1948; Norman, 1988).

2.2. Pronunciation and tonal errors made by learners of Chinese

Since Chinese is a tone language, learners whose first language (L1) is not a tone language usually find Chinese tones very difficult to acquire (Chen, 1974; Shen, 1989). Several studies have investigated tonal errors at the perceptual level (Chen, 1997; Wang, Jongman, & Sereno, 2003; Wang, Spence, Jongman, & Sereno, 1999) or the production level (Chen, 1974, 1997; Miracle, 1989; White, 1981). Such errors often occur when learners pronounce a tone that does not reach the expected pitch value or frequency (i.e., too low for a high tone or too high for a low tone). In particular, the second and the third tones are easily confused, even by many advanced learners of Chinese (Shen & Lin, 1991; Wang et al., 2003; Xing, 2010).

The pronunciation of some Chinese phonemes may be a bit similar to that of the phonemes in learners’ L1, but the pronunciations across languages differ in terms of the manner or place of articulation (e.g., [tʃʰ] sounds a bit similar to [tʃ] in English). Learners usually have difficulty in distinguishing between phonemes that have similar pronunciation and tend to use their L1 to imitate the pronunciation of the second language (L2) (Harrison, Lo, Qian, & Meng, 2009; Lo, Zhang, & Meng, 2010; Meng, Lo, Wang, & Lau, 2007), which results in negative transfer from L1 to L2. Thus, many pronunciation errors can be found in the articulation of phonemes (initials/finals).

2 The status of the medial is not discussed here further because it is not the focus of this paper. In our study, we always take the medial glide as a part of the final for the corpus analysis and for providing comprehensible feedback to learners of Chinese via our system because it is the standard way to teach Chinese sounds.

3 Elements in the parentheses are optional; X can be V or C.
Among the errors made in the pronunciation of Chinese phonemes, many errors of initials are related to aspiration and retroflexion, which are distinctive features in Chinese and, therefore, important for learners of Chinese to learn. Liu and Qi (2004) point out that French learners of Chinese often mix up the aspirated sounds with the unaspirated ones. For instance, they tend to mispronounce \([t^h]\) as \([t]\), \([ts^h]\) as \([ts]\), \([tc]\) as \([te]\), and \([p]\) as \([ph]\) or vice versa depending on the following vowel. However, English learners of Chinese tend to pronounce unaspirated sounds with voicing; for example, \([p]\) is pronounced as \([b]\), and \([t]\) is pronounced as \([d]\). These kinds of errors do not affect the comprehensibility in communication; therefore, they are corrected less often. Retroflexion errors, on the other hand, are more diverse. English learners of Chinese often mispronounce \([ts]\) as \([dz]\), \([ts^h]\) as \([ťʃ]\), \([g]\) as \([j]\), and \([z]\) as \([a]\) (Ni & Wang, 1992), which are typical examples of negative transfer from L1 to L2, whereas Korean learners often mispronounce \([ts^h]\) as \([ťʃ]\), \([ts]\) as \([ťʃ]\), and \([a]\) as \([s]\) (Chen, 2008). There are also various types of pronunciation errors related to finals. In general, finals like \([u]/[y]\), \([an]/[an]\) or \([aŋ]/[an]/[in]\), and \([aŋ]/[aŋ]/[in]\) are easily confused with each other by learners speaking different L1s. Assisting learners of Chinese to overcome their Chinese pronunciation problems was a central goal of the study; therefore, we developed a Chinese computer-assisted pronunciation training system.

2.3. CAPT systems for foreign language learning

Several studies have examined the advantages of applying speech technology in CAPT systems (Demenko, Wagner, & Cylwik, 2010; Ehsani & Knodt, 1998; Eskenazi, 1999; Hincks, 2003; Kartal, 2006; Neri, Cucchiarini, Strik, & Boves, 2002; Wachtowicz & Scott, 1999). Speech technologies, such as automatic speech recognition (ASR) and speech signal processing techniques, enable language learners to practice pronunciation through pronunciation training systems in a stress-free environment. These technologies are particularly helpful for learners who may be shy or feel uncomfortable participating in oral presentations in a traditional class setting. Moreover, the systems allow learners to undertake specific pronunciation drills at their convenience since it is difficult for learners and teachers to find the time for such practice in a formal classroom environment due to the specific needs of learners who speak different L1s (Hincks, 2003).

Despite these advantages, current pronunciation training systems cannot meet all learners’ needs. Many systems suffer from inaccurate speech recognition, resulting in erroneous feedback (Neri, Cucchiarini, & Strik, 2008; Neri et al., 2008; Yeh et al., 2004). Furthermore, most systems give only summative feedback to learners such as whether their pronunciation is right or wrong (Witt & Young, 2000) or what scores they get on a set activity (Chen, Jang, & Tsai, 2007; Neumeyer, Franco, Digalakis, & Weintraub, 2000). They fail to pinpoint how or why a specific error is made and provide possible remedies. In contrast, most learners need formative feedback to know if their pronunciation is correct and, if not, how to improve their pronunciation (Neri et al., 2002).

2.4. Problems with current CAPT systems for Chinese learning

Although some Chinese pronunciation training software and systems are now available commercially, they are far from perfect. These systems fail to detect tonal or phonemic differences—such as retroflexion (\([s]\) vs. \([z]\), \([ts]\) vs. \([ts]\), \([ts^h]\) vs. \([ts^h]\)), or aspiration (\([t]\) vs. \([t^h]\), \([k]\) vs. \([k^h]\), \([tc]\) vs. \([tc^h]\)), or tonal differences (this could happen between any two tones depending on the performance of a given system)—and do not provide corrective feedback (L Labs, 2012; Rossetta Stone, 2013; Tell Me More, 2013). Score-based systems (e.g., Tell Me More and MyCT) give feedback based on the correctness of words and sentences and on prosody, fluency, and voice volume; however, they do not provide corrective feedback. Moreover, some of the systems are even not sensitive enough to detect pronunciation errors or may give false alarms when no errors exist, thereby misleading and confusing the learners. For these reasons, the system developed for this study exploits speech technologies to help detect pronunciation errors made by learners of Chinese and give them corrective feedback on their errors. In contrast to most score-based pronunciation training systems and the error networks used for scoring English pronunciation (Witt & Young, 2000), our system uses decision-tree-based tone assessment and focuses on the phone models of easily confused phoneme pairs to enhance the error detection function. Because easily confused phoneme pairs only differ in one phonemic feature (e.g., aspiration or retroflexion), learners of Chinese often mispronounce them due to their phonetic similarities or negative transfer from their L1 to Chinese (Harrison et al., 2009). Therefore, we employ strict speech recognition networks that differ by one frequently mispronounced phoneme to capture those phoneme errors and provide corresponding corrective feedback. Moreover, a decision-tree-based tone assessment technique identifies tonal errors by traversing the decision tree and provides corrective feedback based on the fundamental frequency (\(F_0\)) features that cause the errors.

Finally, we introduce an adaptive procedure that manipulates the pronunciation training flow to provide specific pronunciation exercises for sounds that learners find difficult. More specifically, we utilize human knowledge to design discriminative exercises and apply them in the system to help discover pronunciation errors. Based on the errors detected, our system provides feedback and a series of tailor-made exercises that focus on those errors to help each learner correct his/her own specific pronunciation errors.

3. Methods for detecting phoneme and tonal errors

We divided the analysis of pronunciation samples into three parts: phoneme verification, phoneme error detection, and tone assessment. Phoneme verification was used to exclude out of vocabulary (OOV) pronunciation, phoneme error detection
was performed by ASR, and tone assessment was based on a decision-tree scheme. After the phoneme verification step, we checked the pronunciation for phoneme and tonal errors.

### 3.1. Speech recognizer

We used the MAT2000 corpus (Wang, Seide, Tseng, & Lee, 2000), which contains 77,035 utterances, to train our speech models for the Chinese speech recognizer. There are 50 context-independent phone models, one short pause model, and one silence model. The models were trained as hidden Markov models (HMMs) with the maximum likelihood (ML) criteria by using the Hidden Markov Model Toolkit (HTK; Young et al., 2000). The phone model contained three states, each of which had 32 Gaussian mixtures; the short-pause model had one state comprising 64 Gaussian mixtures; and the silence model had three states consisting of 64 Gaussian mixtures in total. Thirty-nine mel-frequency cepstral coefficients (13 MFCCs and their first and second time derivatives) were computed with a window size of 20 ms and a frame shift of 10 ms. Moreover, feature domain cepstrum mean subtraction, variance normalization, and auto-regression moving-average (ARMA) filtering (Chen & Bilmes, 2007) were applied to partially reduce the distortion caused by the channel, handset, and background noise. The pronunciation of a phoneme was recognized by matching the observed feature vectors with the phone models to obtain a sequence of phone units.

### 3.2. Phoneme verification

We adopted a vocabulary-independent approach (Kawahara, Lee, & Juang, 1998; Sukkar & Lee, 1996) to verify the input pronunciation. First, we segmented (forced alignment) the pronunciation into phonemes according to the characters to be practiced. Then, the phone models were used to perform hypothesis testing on every segment. Next, we computed the verification score based on its corresponding log likelihood ratio (LLR), as shown below:

$$\text{LLR}_n = \frac{\log P(O|H_0) - \log P(O|H_1)}{l_o}, \quad (1)$$

where $O$ is the observed pronunciation segment, $H_0$ is the null hypothesis (model) in which the phoneme $n$ is presented in the observed pronunciation segment, and $H_1$ is the alternative hypothesis (model) in which the phoneme $n$ is not in the pronunciation segment $O$. Because the phone models are statistical models, the verification score represents the likelihood that a phoneme $n$ exists in the observed pronunciation segment. The score is normalized by the duration $l_o$ and then combined with the scores of the other phonemes. If the total score, LLR, exceeds a predefined threshold, the pronunciation is accepted; otherwise, it is rejected. For example, if an insertion occurs, a low LLR of one or more phonemes would be found and yield a total score that would be lower than the threshold and the pronunciation would be rejected as incorrect. The threshold for the accept/reject decision is usually predetermined such that false acceptance and false rejection have equal probability. It should be noted that the determination of the threshold should consider the influence of learners’ accents. The higher the threshold, the harder it will be for learners to pass the test, which might frustrate them. However, if the threshold is too low, some OOV pronunciation would pass the verification test and unreliable corrective feedback would be generated. As a result, learners may lose confidence in the pronunciation training system.

### 3.3. Phoneme error detection and corrective feedback design

The proposed system, as mentioned in §2.3, mainly considers three types of pronunciation errors (i.e., aspiration, retroflexion, and certain vowels) that are often made by learners of Chinese and are concerned with easily confused phoneme-pairs (see Table 1). Based on these types of pronunciation errors, we used easily confused phoneme-pair recognition-nets to distinguish the correct pronunciation of a given phoneme from its corresponding confusing one. An easily confused phoneme-pair recognition-net is a recognition network used by the speech recognizer in conjunction with phone models to output a phone-level transcription of the learner’s speech. Except for a correct phoneme sequence, we added another sequence to substitute the given phoneme for its corresponding confusing one. For example, when determining the correctness of the pronunciation of “當 [dàng4] [təŋ]”, we adopted a phoneme pair recognition-net [t] and [th⁴], as shown in Fig. 1.

---

Table 1

<table>
<thead>
<tr>
<th>Phoneme pairs that are easily confused by learners of Chinese.</th>
<th>Easily confused phoneme pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspiration (unaspirated vs. aspirated)</td>
<td>[n][ŋ][p⁰]; [t][t⁰]; [k][k⁰]</td>
</tr>
<tr>
<td>Retroflexion (retroflex vs. non-retroflex)</td>
<td>[ʂ][ʂ⁰]; [ɕ][ɕ⁰]; [ʃ][ʃ⁰]</td>
</tr>
<tr>
<td>Certain vowels</td>
<td>[ou][oʊ]; [ʊ][u]; [an][ʌŋ]</td>
</tr>
</tbody>
</table>

---

4 dàng and many other words used later in this paper are written in Hanyu Pinyin, the official system for transcribing Chinese characters into Latin script.
algorithm (a) adjusts the parameters of the \( t \) phone model to were selected to adjust the parameters of the two phone models and if the input pronunciation of \( t \) is correct, then the GPD tracking (RAPT; Talkin, 1995) to extract the fundamental frequency (F0); then, we used a appropriate feedback based on the results of the traversal was provided to the learners. Associated with the constructed decision tree. Finally, to assess a test utterance, the decision tree was traversed, and the tone models for the decision tree. Based on human knowledge about tone production, a set of detailed comments were Bälter, 2007), which sounds natural and should be easily taken up by learners. Pronunciation. This is the most common and effective type of corrective feedback provided in a classroom setting (Engwall & Lee, 2010). We used the decision tree, which is often intuitive and instructive in error analysis, to model the pitch patterns of correctly produced (good) and incorrectly produced (bad) tones. The decision tree (Breiman, Friedman, Olshen, & Stone, 1984) is a flexible model that enables experts and teachers to determine the key attributes of bad tones and then add appropriate feedback to the decision tree’s nodes, leaves, and paths to learners. That is, when bad tones are identified, the learners are given feedback that corresponds to the traversed path to help them improve.

Fig. 2 shows a block diagram of the tone assessment procedure. Unlike phoneme errors, tonal errors cannot be categorized into different types; therefore, each tone’s error patterns must be derived from the mispronunciation data. We used a corpus comprised of good/bad labels of tone production (see §6.1) for this task. We first extracted tone features from the corpus in the training phase. Then, we combined the features with the good/bad labels and used the C4.5 algorithm (Quinlan, 1993) to build the tone models for the decision tree. Based on human knowledge about tone production, a set of detailed comments were associated with the constructed decision tree. Finally, to assess a test utterance, the decision tree was traversed, and appropriate feedback based on the results of the traversal was provided to the learners.

The tone features contained the pitch information about each syllable. First, we used the robust algorithm for pitch tracking (RAPT; Talkin, 1995) to extract the fundamental frequency (F0); then, we used a five-point moving average filter to smooth the pitch contour. Because of the differences in the mean F0 of speakers, F0 must be normalized across speakers to facilitate meaningful comparisons. The process was based on a method widely used in Mandarin Chinese tone studies (Rose, 1987; Wang et al., 2003); that is, if \( x \) is the observed raw pitch value, the tone can be normalized based on the following formula:

\[
p(x) = 4 \frac{\log x - \log \text{Min}}{\log \text{Max} - \log \text{Min}} + 1,
\]

where Max and Min are, respectively, the highest and lowest F0 over all the syllables of each speaker after smoothing. This puts F0 on a common five-point scale that was originally proposed by Chao (1948).

If the likelihood of path 1 is greater than that of path 2, the pronunciation is considered correct; otherwise, it is deemed to be a mispronunciation.

The error detection task becomes a classification problem when using easily confused phoneme-pair recognition-nets. However, models trained with the ML criteria are not guaranteed to achieve the best classification accuracy. Such models rely on estimates of the probability distribution; therefore, they are suboptimal. This is usually because the form of probability distribution for various speech signal dimensions is unknown and any assumed distribution could be incorrect. Therefore, we used a discriminative training method called the generalized probabilistic descent (GPD) algorithm (Juang, Chou, & Lee, 1997), which minimizes the number of classification errors, to detect pronunciation errors. The goal of the GPD algorithm is to derive the best recognition/classification results rather than fit the distributions to the data. The GPD algorithm adjusts the parameters of the phone models during training, then increases the discriminative abilities of models, and reduces the classification error rate. Since the MAT2000 corpus was recorded from native speakers, we accepted them all as correct pronunciations that were then used in the GPD training. For example, when utterances that contain the phoneme [t] or [\( \text{th} \)] were selected to adjust the parameters of the two phone models and if the input pronunciation of [t] is correct, then the GPD algorithm (a) adjusts the parameters of the [t] phone model to increase the likelihood of that pronunciation and (b) adjusts the parameters of the [\( \text{th} \)] phone model to reduce the likelihood of its pronunciation. Consequently, the models are better able to discriminate between correct and incorrect pronunciation.

Based on the phoneme-error detection function, our system exploited human knowledge to provide corrective feedback that is frequently used in Chinese classes and easily comprehended by most learners. For example, if a learner pronounces an unaspirated [t] as an aspirated [\( \text{th} \)], our system gives the following corrective feedback that is often given in Chinese classes: “Do not produce a puff of air when producing the initial sound; try again.” Subsequently, the system demonstrates the correct pronunciation. This is the most common and effective type of corrective feedback provided in a classroom setting (Engwall & Bälter, 2007), which sounds natural and should be easily taken up by learners.

### 3.4. Decision-tree-based tone assessment and corrective feedback design

The proposed CAPT system incorporated the decision-tree-based tone assessment technique (Liao, Chen, Chang, Guan, & Lee, 2010). We used the decision tree, which is often intuitive and instructive in error analysis, to model the pitch patterns of correctly produced (good) and incorrectly produced (bad) tones. The decision tree (Breiman, Friedman, Olshen, & Stone, 1984) is a flexible model that enables experts and teachers to determine the key attributes of bad tones and then add appropriate feedback to the decision tree’s nodes, leaves, and paths to learners. That is, when bad tones are identified, the learners are given feedback that corresponds to the traversed path to help them improve.

Fig. 2 shows a block diagram of the tone assessment procedure. Unlike phoneme errors, tonal errors cannot be categorized into different types; therefore, each tone’s error patterns must be derived from the mispronunciation data. We used a corpus comprised of good/bad labels of tone production (see §6.1) for this task. We first extracted tone features from the corpus in the training phase. Then, we combined the features with the good/bad labels and used the C4.5 algorithm (Quinlan, 1993) to build the tone models for the decision tree. Based on human knowledge about tone production, a set of detailed comments were associated with the constructed decision tree. Finally, to assess a test utterance, the decision tree was traversed, and appropriate feedback based on the results of the traversal was provided to the learners.

The tone features contained the pitch information about each syllable. First, we used the robust algorithm for pitch tracking (RAPT; Talkin, 1995) to extract the fundamental frequency (F0); then, we used a five-point moving average filter to smooth the pitch contour. Because of the differences in the mean F0 of speakers, F0 must be normalized across speakers to facilitate meaningful comparisons. The process was based on a method widely used in Mandarin Chinese tone studies (Rose, 1987; Wang et al., 2003); that is, if \( x \) is the observed raw pitch value, the tone can be normalized based on the following formula:

\[
p(x) = 4 \frac{\log x - \log \text{Min}}{\log \text{Max} - \log \text{Min}} + 1,
\]

where Max and Min are, respectively, the highest and lowest F0 over all the syllables of each speaker after smoothing. This puts F0 on a common five-point scale that was originally proposed by Chao (1948).
After being normalized on a five-point scale, we divided each syllable’s F0 into three equal segments, as shown in Fig. 3. Since tone 5 (the neutral tone) has been widely regarded as the cause of unstable pitch patterns in tone production, we did not consider it in this study. We adopted the mean value of F0 for each segment and the differences between the mean values as feature vectors for each syllable. This yielded a six-dimension feature vector for each segmented syllable.

We divided the feature vectors into several right-context-dependent (RCD) categories to train the decision tree. Two RCD tones were regarded as different if their immediate right tones are different. For example, tone 1 before tone 2 in a syllable is defined as “tone 1 + tone 2” in the RCD context, which is still a type of tone 1. The number of RCD tone models is much larger than the number of context-independent tone models. For example, tone 1 has five models: tone 1 + tone 1, tone 1 + tone 2, tone 1 + tone 3, tone 1 + tone 4, and tone 1 + silence, which means a total of 20 RCD tone models for the four tones. However, because our good/bad tones were extremely biased (described in §6), the C4.5 algorithm cannot minimize the false acceptance and false rejection rates at the same time. Solving this problem involved assigning different costs to different errors. We used a method called MetaCost (Domingos, 1999) to help us convert C4.5 into a cost-sensitive tool to solve such unbalanced data problems. We used a well-known machine learning tool called Weka (Witten & Frank, 2005) for the entire training procedure that enabled us to construct the tone models quickly.

The constructed tone models were used to examine the attributes of each leaf node that classifies a syllable as a bad tone. As shown in Fig. 4, path1 (solid arrows) represents a bad tone 2. According to the criteria determined by C4.5 algorithm automatically on the traversed path, we found that the samples classified to this leaf node had two problems: an extremely high pitch level at the beginning and a negative slope during tone production. Therefore, we could give corrective feedback like “lower your pitch in the beginning” or “increase your pitch gradually” on the associated nodes on path1. Exploiting the decision tree in this way, we were able to examine each node with regard to the criterion used to split the node and then

---

**Fig. 2.** Block diagram of the decision-tree-based tone assessment procedure in the training (upper) phase and the testing (lower) phase.

**Fig. 3.** Feature vector extraction for tone assessment.
added instructions to help the learner correct the errors. We could also summarize such errors in a systematic manner that allowed individual learners to improve their tone production.

4. Architecture of the adaptive Chinese pronunciation training system

Fig. 5 shows a block diagram of our adaptive pronunciation training system. This system contains two parts: the server side and the client side. We implemented the web-based graphic user interface (GUI) on the client side in JavaScript. Its main function is to display practice texts and feedback information. Fig. 6 shows the appearance of the interface in the current system. We display the practice text and corresponding pinyin in region 1 while illustrations and feedback information are shown in region 2. Learners use the “record” button (region 3) to start the course. Any tonal or phonemic errors are listed in region 4 so that the learners can see the problem(s) to be corrected. The client-side GUI communicates with the server via the Internet. One main PHP script controls the entire learning procedure and communicates with other components (i.e., tone assessment, phoneme error detection, phoneme verification, and text-to-speech [TTS]) and course databases. Once the input
speech has been evaluated, the assessment results and the corrective feedback are sent back to the client side. Most corrective feedback is read by means of TTS.

After each learner starts the course, this responsive system adjusts practice materials according to the errors made by the learner. The input pronunciation is first verified for correctness. If the pronunciation is incorrect, then the system asks the learner to repeat the exercise until the pronunciation is verified. As shown in Fig. 7, at the beginning of the course, the system asks learners to practice the sentences that contain the vocabulary they have to learn (Fig. 8a). The system detects the main
categories of tonal and phoneme errors during this step. If no errors are detected, the system moves to the next exercise. When errors are detected, the system analyzes the detection results and selects disyllabic to quadrisyllabic morphemes that contain the same kinds of phonemes and tones from a course database for further practice (Fig. 8b). After repeating this procedure, the system detects the major tonal or phoneme errors made by the learner. Those errors are stored for future reference, and the system provides the learner with disyllabic (Fig. 8c) or even monosyllabic (Fig. 8d) exercises that focus on
the saved errors one-by-one to help the learner improve. If the learner cannot complete a monosyllabic exercise successfully, the error is saved and related exercises will be presented to the learner when next using the system.

5. Corpus evaluation

The speech data in the National Taiwan Normal University (NTNU) corpus, which contained errors made by eight English students of Chinese, were used in training and testing the phoneme error detection and tone assessment systems. We
compiled our corpus by focusing on three types of phoneme errors and set up nine phoneme pairs (see Table 1) that are easily mispronounced by learners of Chinese. Then, based on the nine target phoneme pairs, we created monosyllable and disyllable pairs for the corpus. For example, for the initial phoneme /p/, we chose the syllable 比 (bǐ [pi]) and the syllable 披 (pī [pi]) to form a target monosyllable-pair because, with the exception of the tones, the two syllables differ only in their initials. As well, the initials only differ in their manner of articulation, namely, aspiration in this case. Next, we created a disyllable-pair by inserting the target monosyllable-pair into the first and second positions in the disyllable pair, respectively. For instance, to form a contrasting syllable pair, we put the syllable 比 (bǐ [pi]) in the first position of a disyllabic morpheme like 比方 (bǐfāng [pi,faŋ]) and the syllable 披 (pī [pi]) in the first position of a disyllabic morpheme like 披風 (pīfēng [pi, fəŋ]). Accordingly, the homophone of the target monosyllable pair was put in the second position of other disyllabic morphemes, such as 封筆 (fēngbǐ [foŋ, pi]) versus 橫批 (héngpī [hoŋ, phi]). The objective of this manipulation was to assess the effect of the positions on the pronunciation of target monosyllable-pairs in the disyllabic context.

The NTNU corpus contains 1900 syllables that correspond to morphemes with their lengths ranging from 1 to 4 syllables. The pronunciations of the monosyllabic and disyllabic morphemes were collected only to assess the performance of the phoneme error detection system; the pronunciations of monosyllabic to quadrisyllabic morphemes were used for training and testing the tonal error detection system. We divided the 1900 syllables into 20 subsets, each comprised of 95 syllables forming morphemes of different lengths. The distribution of morphemes and their corresponding number of syllables in each subset was as follows: morphemes 1–10 were monosyllabic, 11–35 were disyllabic, 36–40 were trisyllabic, and 41–45 were quadrisyllabic. When collecting the speech data, each morpheme was displayed twice on the computer showing the Chinese morpheme in traditional form along with its pinyin transcription. In the first presentation, the students were required to read the morpheme by themselves; in the second presentation, the students were asked to speak after the pronunciation demonstration (i.e., shadowing) given by the computer. The pronunciation demonstrations were recorded in advance by a professional Chinese language teacher. We asked the students to pronounce each morpheme twice for two reasons: more data would be collected and we could assess whether shadowing helped improve their immediate pronunciation as compared to the slightly delayed pronunciation. Overall, our speech corpus contained 30,400 (8 × 1900 × 2) pronounced syllables collected from eight subjects, each of whom pronounced the 1900 syllables twice; half of the pronounced syllables were spoken by the students after the pronunciation demonstrations.

We then asked six raters (three professional and three preservice Chinese language teachers) to evaluate the quality of the pronunciations. After listening to the pronunciation of each syllable, they registered their decision via an online evaluation interface. The corpus was divided into 16 sets, each of which contained 20 subsets. The raters had to finish evaluating a subset completely in order to proceed to the next one. Fig. 9 shows the evaluation interface.

Evaluation of the pronunciation of each syllable required the raters to mark where the error(s) occurred (initial position, final position, or both). The tonal evaluation involved a five-point Likert scale (1 = worst, 2 = bad, 3 = mediocre, 4 = good, 5 = best). We used this scale instead of a binary (right vs. wrong) scale for tone evaluation because judging the correctness of tone pronunciation is sometimes difficult and the judgments often vary from person to person. Although some tone pronunciations may not be completely correct, they are still acceptable to some raters but not to others. Thus, a five-level scale provides raters the ability to make more fine-grained judgments and allows any differences between their evaluations to be indicated more precisely.

![Fig. 9. The online interface for evaluating pronunciation.](image)
6. Results

6.1. Corpus evaluation

We used Fleiss’ kappa to calculate the level of agreement among the six raters for their evaluations of the initials and finals. The results ($k = 0.29$ for initials and $k = 0.21$ for finals) showed only a fair agreement between the two ratings (Landis & Koch, 1977). Therefore, we adopted a voting process to further determine bad pronunciations: a phoneme’s pronunciation was rated as bad if more than three raters classified it as bad; otherwise, it was deemed to be good. Table 2 shows the percentages of good phoneme pronunciations as evaluated by the six raters.

Tables 3 and 4 show the error ratios and total number of the nine selected phoneme-pairs listed according to the first and second pronunciations. Other phonemes were not included because they were not relevant to the current study. Error ratio was the error frequency of a given phoneme divided by its total number in our corpus.

Although the pronunciation of tones was rated on a five-point Likert scale, to emphasize the characteristics of the syllables that a majority of the raters marked as bad and to exclude extremely subjective judgments, we simplified the annotation data as follows. A tone’s pronunciation was only classified as bad (i.e., wrong) if more than three raters ranked it as 1 or 2; otherwise, the pronunciation was classified as good. These binary labels were also needed for the subsequent decision-tree training. Table 5 shows the ratios of tonal errors in the first and second pronunciations. We performed Wilcoxon Signed Ranks Tests to determine whether there were overall significant differences between the error rates in the first and the second pronunciations. The results showed that significant differences were found between the two pronunciations for the initials ($Z = -2.31$, $p < 0.05$) and the finals ($Z = -2.21$, $p < 0.05$) but not for the tones. The results seemed to indicate that shadowing could instantly help learners improve their pronunciation of initials and finals, but it failed to improve their pronunciation of tones. However, these results should be interpreted very carefully because the sample sizes observed here were relatively small. Whether shadowing had an immediate and significant effect on learners’ pronunciation is inconclusive, but the results (Tables 3–5) generally showed a declining trend in the error rates in the second pronunciation.

We also used the inter-rater correlation to evaluate the consistency of the ratings given to the speakers. Table 6 summarizes the inter-rater correlations for all the syllables. The average correlation between raters was 0.68, which indicates that the agreement was not by chance.

<table>
<thead>
<tr>
<th>Rater</th>
<th>First/second pronunciation (%)</th>
<th>Initial</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>95.76/96.79</td>
<td>95.73/97.31</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>97.36/97.85</td>
<td>91.87/93.06</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>98.19/98.99</td>
<td>97.30/98.53</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>98.16/98.65</td>
<td>95.69/96.64</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>92.54/93.13</td>
<td>86.03/87.74</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>98.66/98.97</td>
<td>98.07/99.20</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2
Percentage of good phoneme pronunciations in the NTNU corpus.

<table>
<thead>
<tr>
<th>Rater</th>
<th>First/second pronunciation (%)</th>
<th>Initial</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>95.76/96.79</td>
<td>95.73/97.31</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>97.36/97.85</td>
<td>91.87/93.06</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>98.19/98.99</td>
<td>97.30/98.53</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>98.16/98.65</td>
<td>95.69/96.64</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>92.54/93.13</td>
<td>86.03/87.74</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>98.66/98.97</td>
<td>98.07/99.20</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3
Errors in the initials of the first and second pronunciations.

<table>
<thead>
<tr>
<th>Initial (written in IPA)</th>
<th>Error ratio (%) (first/second pronunciation)</th>
<th>Total number (first/second pronunciation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[p]</td>
<td>0/0</td>
<td>536/536</td>
</tr>
<tr>
<td>[pʰ]</td>
<td>0/0</td>
<td>232/232</td>
</tr>
<tr>
<td>[t]</td>
<td>0/0</td>
<td>648/648</td>
</tr>
<tr>
<td>[tʰ]</td>
<td>0/0.26</td>
<td>384/384</td>
</tr>
<tr>
<td>[k]</td>
<td>0.25/0</td>
<td>400/400</td>
</tr>
<tr>
<td>[kʰ]</td>
<td>0/0</td>
<td>272/272</td>
</tr>
<tr>
<td>[ʦ]</td>
<td>20.23/11.59</td>
<td>440/440</td>
</tr>
<tr>
<td>[ʦʰ]</td>
<td>11.32/4.25</td>
<td>320/320</td>
</tr>
<tr>
<td>[s]</td>
<td>11.32/6.25</td>
<td>592/592</td>
</tr>
<tr>
<td>[ʦ]</td>
<td>13.26/7.29</td>
<td>344/344</td>
</tr>
<tr>
<td>[ʦʰ]</td>
<td>12.51/11.67</td>
<td>264/264</td>
</tr>
<tr>
<td>Total*</td>
<td>2.51/1.67</td>
<td>12.368/12.368</td>
</tr>
</tbody>
</table>

* Includes initials not listed in this table.
6.2. System performance on phoneme error detection

We employed the mispronunciation corpus to evaluate our system’s performance in detecting phoneme and tonal errors. Since the corpus contained a small percentage of bad phoneme-pronunciations, we used half of the good pronunciations of a phoneme as bad pronunciations of the phoneme that could be easily confused with the target phoneme. For example, there were 815 good pronunciations of [t], so we divided them into 408 good pronunciations of [t] and 407 bad pronunciations of [th]. Table 7 shows False Reject (FR), False Alarm (FA), Precision, Recall rates, and F-measure using phone models trained on ML and then improved by GPD training. Because more than one phone was involved in final pronunciation errors like [ɑʊ], [oʊ], [ɑn], and [ən], the corresponding HMM models could not be adjusted by the GPD algorithm; therefore, our system still used the ML-trained models of [ɑʊ], [oʊ], [ɑn], and [ən]. The results demonstrate that the accuracy improved, particularly in relation to retroflexion. The FR and FA rates were extremely biased, which shows that the original retroflex and nonretroflex models were so similar that the models could not discriminate this phonemic feature with a high degree of accuracy. The GPD algorithm successfully adjusted the parameters of the phone models, improved the discriminative ability of the models, and reduced the classification error rate.

6.3. System performance on tone assessment

We verified the reliability of tone modeling for good and bad tone productions using a 10-fold cross-validation. A recognition result is considered correct if it is consistent with the rating given by human experts. Table 8 shows the tone recognition results for 20 RCD tone models and demonstrates that tone recognition was reliable and its accuracy was comparable to that of related studies (Chen et al., 2007; Tang & Yin, 2007). The trend of the results reflects the production difficulties of the four tones to a certain degree. For example, the accuracy rate for tone 3 was relatively low due to variations of its pitch during pronunciation; therefore, tone 3 could not be modeled easily.

7. Conclusion

We have presented a prototype for an adaptive Chinese pronunciation training system developed by the Industrial Technology Research Institute and National Taiwan Normal University. To improve the performance of the Chinese pronunciation training system, we used state-of-the-art speech technologies to design a speech recognition net based on easily
confused phoneme pairs and a decision-tree-based tone assessment method. The detection results demonstrate that our system can accurately identify the pronunciation errors that learners usually make and then provide appropriate feedback and strategies to help them correct their errors. These functions are rarely available in existing pronunciation training systems. The corrective feedback, which is based on the experiences of professional classroom teachers, sounds natural and is easy to follow.

Although the accuracy rates of this system’s error-detection functions are relatively high, an extensive usability test should be conducted in the future to evaluate the system’s pedagogical value. In addition, a survey of learners’ opinions about the GUI and the training procedures could provide guidelines for configuring an even more user-friendly interface so that learners can operate the system and practice pronunciation more easily and efficiently. Finally, at present, the decision-tree-based tone models have to be further trained with mispronunciation data. More samples of pronunciation errors, especially the common errors made by learners who speak different mother tongues, would facilitate the training of a more accurate and powerful error-detection system to meet the needs of learners from different countries.

Acknowledgments

This study was partially supported by Project 103-EC-17-A-24-0617 conducted by ITRI and NTNU’s Aim for the Top University Project 100J000151 under the sponsorship of Taiwan’s Ministry of Economic Affairs and Ministry of Education respectively. The authors would like to thank Prof. Larry D. Yore (University of Victoria) and Sharyl A. Yore for their mentoring assistance in relation to this article.

References


Table 7

<table>
<thead>
<tr>
<th>Table 7</th>
<th>Accuracy evaluation of phoneme error detection (%)</th>
<th>ML</th>
<th>GPD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FR FA Precision Recall F-measure</td>
<td>FR FA Precision Recall F-measure</td>
<td></td>
</tr>
<tr>
<td>[ʦʰ]</td>
<td>80.78 54.66 88.16 36.22 51.34</td>
<td>20.00 23.90 79.57 80.00 79.78</td>
<td>58.45</td>
</tr>
<tr>
<td>[ʦʰ]</td>
<td>10.63 68.11 53.16 89.38 66.67</td>
<td>23.75 29.73 68.93 76.25 72.40</td>
<td>17.21</td>
</tr>
<tr>
<td>[ʈʂ]</td>
<td>84.23 69.51 97.66 45.77 60.61</td>
<td>28.87 15.67 86.59 71.13 77.69</td>
<td>43.37</td>
</tr>
<tr>
<td>[ʈʂʰ]</td>
<td>9.26 90.01 53.99 90.74 67.70</td>
<td>17.13 25.09 71.60 82.87 76.82</td>
<td>28.24</td>
</tr>
<tr>
<td>[ʂ]</td>
<td>47.27 1.95 97.62 52.73 68.48</td>
<td>13.83 12.68 91.16 86.17 88.60</td>
<td>63.82</td>
</tr>
<tr>
<td>[ʂʰ]</td>
<td>1.94 45.51 58.72 98.06 73.45</td>
<td>12.14 11.86 83.03 87.86 85.38</td>
<td>44.91</td>
</tr>
<tr>
<td>[p]</td>
<td>21.13 4.57 90.00 78.87 84.07</td>
<td>16.49 5.11 89.50 83.51 86.40</td>
<td>14.65</td>
</tr>
<tr>
<td>[pʰ]</td>
<td>3.49 20.51 89.97 96.51 93.13</td>
<td>4.30 11.79 93.93 95.70 94.81</td>
<td>24.46</td>
</tr>
<tr>
<td>[ʈ]</td>
<td>21.82 9.07 77.71 78.18 77.95</td>
<td>14.55 10.54 76.63 85.45 80.80</td>
<td>12.95</td>
</tr>
<tr>
<td>[ʈʰ]</td>
<td>8.35 12.73 94.67 91.65 93.13</td>
<td>9.58 8.48 96.34 90.42 93.28</td>
<td>2.17</td>
</tr>
<tr>
<td>[k]</td>
<td>12.00 9.52 88.71 88.00 88.35</td>
<td>11.20 9.86 88.45 88.80 88.62</td>
<td>2.31</td>
</tr>
<tr>
<td>[kʰ]</td>
<td>7.14 16.06 87.22 92.86 89.95</td>
<td>7.14 15.66 87.50 92.86 90.10</td>
<td>1.48</td>
</tr>
<tr>
<td>[u]</td>
<td>11.04 6.91 94.26 88.96 91.53</td>
<td>– – – – –</td>
<td>–</td>
</tr>
<tr>
<td>[ou]</td>
<td>6.65 10.81 87.10 93.35 90.12</td>
<td>– – – – –</td>
<td>–</td>
</tr>
<tr>
<td>[y]</td>
<td>7.26 3.92 94.32 92.74 93.52</td>
<td>5.59 4.31 93.89 94.41 94.15</td>
<td>9.71</td>
</tr>
<tr>
<td>[u]</td>
<td>7.03 6.15 95.58 92.97 94.26</td>
<td>7.03 5.59 95.97 92.97 94.44</td>
<td>3.26</td>
</tr>
<tr>
<td>[an]</td>
<td>17.40 9.84 87.92 82.60 85.18</td>
<td>– – – – –</td>
<td>–</td>
</tr>
<tr>
<td>[ən]</td>
<td>10.78 14.80 87.39 89.22 88.30</td>
<td>– – – – –</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 8

<table>
<thead>
<tr>
<th>Table 8</th>
<th>Assessment results of 20 RCD tone models in the NTNU corpus (%)</th>
<th>Tone1</th>
<th>Tone2</th>
<th>Tone3</th>
<th>Tone4</th>
</tr>
</thead>
<tbody>
<tr>
<td>*+tone1</td>
<td>95.57</td>
<td>77.73</td>
<td>85.97</td>
<td>89.16</td>
<td></td>
</tr>
<tr>
<td>*+tone2</td>
<td>91.32</td>
<td>86.48</td>
<td>88.95</td>
<td>94.80</td>
<td></td>
</tr>
<tr>
<td>*+tone3</td>
<td>91.64</td>
<td>81.11</td>
<td>84.67</td>
<td>90.90</td>
<td></td>
</tr>
<tr>
<td>*+tone4</td>
<td>92.99</td>
<td>87.78</td>
<td>72.26</td>
<td>91.04</td>
<td></td>
</tr>
<tr>
<td>*+silence</td>
<td>91.89</td>
<td>85.19</td>
<td>78.83</td>
<td>91.43</td>
<td></td>
</tr>
</tbody>
</table>


