Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 53 (2011) 175–184

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Investigation of golden speakers for second language learners from imitation preference perspective by voice modification

Ruili Wang*, Jingli Lu

School of Engineering and Advanced Technology, Massey University, Palmerston North, New Zealand
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

Received 19 December 2009; received in revised form 20 August 2010; accepted 23 August 2010

Abstract

This paper investigates what voice features (e.g., speech rate and pitch-formants) make a teacher's voice preferable for second language learners to imitate, when they practice sentence pronunciation using Computer-Assisted Pronunciation Training (CAPT) systems. The CAPT system employed in our investigation uses a single teacher's voice as the source to automatically resynthesize several sample voices with different voice features based on the features of a learner's voice. Our approach is different from that in the study conducted by Probst et al. which uses multiple native speakers' voices as sample voices [Probst, K., Ke, Y., Eskenazi, M., 2002. Enhancing foreign language tutors—in search of the golden speaker. Speech Communication 37 (3–4), 161–173]. Our approach can reduce the influence of characteristics of teachers' voices (e.g., voice quality and clarity) on the investigation. Our experimental results show that a teacher's voice, which has similar speech rate and pitch-formants to a learner's voice, is not always the learner's first imitation preference. Many factors can influence learners' imitation preferences, e.g., background and proficiency of the language that they are learning. Also, a learner's preferences may change at different learning stages. We thus advocate an automatic voice modification function in CAPT systems to provide speech learning material with a wide variety of voice features, e.g., different speech rates or different pitch-formants. Learners then can control the voice modifications according to their preferences.

Keywords: Computer-assisted language learning (CALL); Computer-Assisted Pronunciation Training (CAPT); Voice modification; Pitch; Speech rate

1. Introduction

The importance of pronunciation in second language learning has been recognized by teachers and learners (Derwing, 2003) since verbal communication between people from different countries are becoming frequent with the development of economic globalization. Good pronunciation can make listeners understand easily, while bad pronunciation may become a barrier to verbal communication, or even break down conversations. Thus, language learners are encouraged to improve their pronunciation at least to the intelligible level (Hişmanoğlu, 2006). In the traditional teacher-student-based language learning model, imitation is the most commonly used method to improve pronunciation, and also considered as one of the most effective methods (Ding, 2007).

With the development of speech processing technologies and the popularity of personal computers, Computer-Assisted Pronunciation Training (CAPT) is playing an increasingly important role in pronunciation learning (Eskenazi, 2009). CAPT can provide a private and stress-free learning environment, and allows learners to learn anytime and anywhere, where a computer is available. Moreover, CAPT can also provide individualized learning material and prompt feedbacks. Since CAPT can provide individualized learning material and give learners more autonomy, a question is raised whether different voices which produce same learning material make a difference

* Corresponding author.
E-mail address: r.wang@massey.ac.nz (R. Wang).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.08.015

for pronunciation learning. In other words, what voices are suitable for language learners to imitate? Some previous research has attempted to answer this question.

1.1. Hearing your own voice

Some studies have suggested that language learners can benefit from listening to their own voices producing native-like utterances since it may be easier for them to perceive differences between their own utterances and their native-like utterances (Sundström, 1998; Bissiri and Pfitzinger, 2009). Also, speech synthesis technologies have been developed to synthesize native-like utterances with learners' voice characteristics (Nagano and Ozawa, 1990; Sundström, 1998; Hirose, 2004; Bissiri and Pfitzinger, 2009; Felps et al., 2009).

In order to correct prosodic errors of a learner's voice, prosody conversion techniques have been used to transfer the prosodic features of a teacher's voice to the learner's voice (Nagano and Ozawa, 1990; Sundström, 1998; Hirose, 2004). However, this prosody transferring keeps the segmental errors (e.g., mispronounced phonemes) in the learner's voice intact. The segmental errors of the learner's voice, which are unavoidable in a learner's speech, are then inherited into the prosody modified learners' voices. Because of the segmental errors, practicing with the prosody modified learners' voices goes against the objective of CAPT, which is to help learners produce more native-like utterances in a second language. Thus, these resynthesized utterances by mapping the prosody of a teacher's voice onto a learner's voice are not suitable for learners to imitate.

The foreign accent conversion proposed in (Felps et al., 2009) is claimed to be able to correct both prosodic and segmental errors. However, this foreign accent conversion lowered the voice quality to 2.67 on a 5-point scale due to the distortion generated in the conversion process, in which a score of 1 means bad voice quality and a score of 5 means excellent voice quality. Thus, the voice quality of the foreign accent conversion needs to be improved before it can be applied into CAPT systems.

Voice conversion techniques (e.g., Erro and Moreno, 2007), which transform a source speaker's voice to a target speaker's voice, can potentially be used to modify a teacher's utterance to make it sound as being produced by a learner. However, the aim of voice conversion is to make a voice sound as if it is being produced by the target speaker. Thus, the converted speech also preserves the accent of the target speaker, such as a foreign accent of a language learner. Moreover, voice conversion needs to record a set of the teacher's utterances, as well as the learner's utterances, which have to be fluent, without errors, and being recorded in good quality (Black, 2007), e.g., in a studio-like environment with a high quality microphone. Recording a learner's voice in such good quality is not an easy task since not all learners can speak accurately and fluently, and not all learners' learning environments can meet the studio-like requirements. Thus, more research needs to be done to make the learner's voice more native-like through voice conversion techniques.

Apart from the immature speech synthesis technologies to make a learner's voice more native-like, there are also some negative opinions about the idea of "hearing your own voice speaking". For example, (Black, 2007) claimed that it may be the novelty of this idea impresses language learners and makes it useful, and moreover not everyone likes to listen to his/her own voice. Also, to some learners, hearing their own voices could be distracting, and could hinder them from improving their pronunciation.

1.2. Hearing multiple speakers' voices

Some language educators and teachers advocate that CAPT systems should have a number of speakers' voices for users to select, listen to and imitate. They should also cover different genders, and a wide range of pitch and speech rate (Probst et al., 2002; Dyck, 2002; Lee, 2008). By listening to and imitating their favorite voices, learners might have a better perception of pronunciation. Moreover, hearing multiple voices might also help learners to generalize pronunciation skills that they have gained. This can result in more robust learning.

Lee's study (2008) shows that learners found it difficult to catch each word and imitate utterances when the speech rates of the utterances were high. Thus, the learners would like to control the speed of speech material. Hearing fast speech might increase learners' cognitive load, thereby impeding their interpretation and production of speech in a second language. It is understandable that it may be difficult for novices to imitate utterances of fast speakers, as their efforts might be concentrated on how to speed up their speech rather than how to pronounce each word correctly (Lee, 2008).

Also, in (Dyck's, 2002) review of "Tsi Karhakta: At the Edge of the Woods" (a CAPT system of Mohawk language), Dyck indicated that a slow version of the pronunciation of longer words and sentences would be helpful to novices, and the speech learning material in a system should be produced at least by a male and a female speakers, so that learners could be exposed to more variations in speech. Although slow speech might be beneficial to novices, it is worth to note that slow speech might be detrimental over a long-term course of second language learning, since the objective of second language learning is to perceive and produce natural speech with a regular speed.

However, providing multiple teachers' voices multiplies the workload of recording speech learning material and the storage space. Moreover, no matter how wide the range of the prosodic features of the teachers' voices covers, they cannot always meet all learners' needs. Also, the characteristics of the multiple teachers' voices, such as voice quality and clarity, might also have an impact on the learners' performances.

Although some CAPT systems can provide multiple speakers' voices, the question of which voice is the "golden voice" for a language learner to imitate is still a research issue open to discussion. The pioneer study that is intended to answer this question is conducted by Probst et al. (2002). The survey conducted by Probst et al. (2002) shows that same gender, reasonable speed and clarity are the most commonly mentioned criteria of selecting preferred learning utterances by second language learners. Thus, Probst et al. suggested that CAPT systems should provide multiple teachers' voices producing same learning material in order to select the "golden speaker" for different learners. The study conducted by Probst et al. (2002) investigated the "golden speaker" from the pronunciation improvement perspective. In their study, the measurements to evaluate the effectiveness of different teachers' voices were the reductions of phone error and duration error from pretest to posttest. The subjects were randomly divided into three groups. Given six native speakers' voices, Group 1 subjects were allowed to choose one speaker's voice to imitate by themselves. Group 2 subjects imitated the voices that were the most similar to their own voices in term of pitch and speed, which were automatically chosen by the CAPT system, FLUENCY (Eskenazi and Hansma, 1998). Group 3 subjects imitated the voices that were the least similar to their own voices, which were chosen by FLUENCY. Probst et al. (2002) found that Group 2 improved their pronunciation slightly more than Group 3, and more significantly than Group 1. In their experiment, learners could practice each sentence as many times as desired. It was noticed that on average Group 1 subjects practiced each sentence (3.5 times) fewer times than Group 2 subjects (4.5 times) and group 3 subjects (4.8 times). Probst et al. (2002) argued that whether the less practice was one of the reasons for the poor performance of Group 1 needed to undertake further test. They also claimed that it might be beneficial for CAPT systems to automatically choose the voice that is the most similar to a learner's voice for the learner to imitate.

The study conducted by Probst et al. (2002) investigated the "golden speaker" from the pronunciation improvement perspective. There is no doubt about the importance of pronunciation improvements since the ultimate goal of pronunciation learning is to improve pronunciation. However, pronunciation improvements can be influenced by many factors, such as learners' learning ability and proficiency of the language that they are learning, not only the acoustic features of learning material. Also, these factors make it difficult to directly investigate the relationship between speech learning material and pronunciation improvements.

1.3. Our research

In this paper, we study the "golden speaker" from the learners' imitation preference perspective. We investigate what voice features make a teacher's voice preferable for language learners to imitate since learners' preferred speech learning material may please them and increase their learning interests. As indicated by Arnett (1952), if a teacher speaks with a smooth, easy and pleasant voice, his/her students try to imitate his/her voice. Also, some learners may be more receptive to certain voices. For instance, as claimed by Jacob and Mythili (2008), children might be more receptive to their parents' or teachers' voices. A pleasant voice may also help to maintain a positive learning environment that plays an important role in a learning process.

In this paper, we focus on two voice features: speech rate and pitch-formants. In order to provide speech learning material with different voice features, CAPT system CASTLE (Computer-Assisted Stress pattern Teaching and Learning Environment) is employed in our investigation. CASTLE (Lu et al., 2010), a system that we have recently developed, is intended to help learners of English as a Second Language (ESL) to improve their abilities to correctly use stress patterns (both sentence stress and lexical stress). The learning material in CASTLE is in the form of sentences. To reduce the influence of characteristics of teachers' voices (e.g., voice quality and clarity), CASTLE uses a single teacher's voice as the source to automatically resynthesize several sample voices based on a learner's voice features (i.e., speech rate and pitch-formants) and the learners' imitation preferences.

Our voice modification transfers the voice features of a *learner's* voice to a *teacher's* voice, unlike previous prosody conversions, which transfer the prosodic features of a *teacher's* voice to a *learner's* voice. Because our voice modification is based on a teacher's voice, the resynthesized utterances can be free from segmental error. Previous prosody conversions are normally based on a learner's voice (e.g., in (Nagano and Ozawa, 1990; Sundström, 1998; Hirose, 2004; Bissiri and Pfitzinger, 2009)), which causes the resynthesized utterances to inevitably inherit the segmental errors (e.g., mispronounced phonemes) from the learner's utterances. Compared with a teacher's speech, a learner's speech is more likely to have segmental errors.

Moreover, unlike the approach in (Probst et al., 2002), which needs to record multiple teachers' voices in order to make the teachers' voices cover a variety of prosodic features, our approach only needs to record one teacher's voice. Based on the teacher's voice, our CAPT system, CASTLE, can resynthesize multiple sample voices with different prosodies by voice modification. Compared with recording multiple teachers' voices, providing multiple sample voices based on the voice modification reduces the workload of producing speech learning material and saves storage space in a computer. Also, the voice modification can resynthesize voices with any prosodic features that language learners may prefer. By investigating learners' imitation preferences, CAPT systems can be developed to provide learners' favorite voices, which may please the learners and promote their learning interests.

This paper is organized as follows. In Section 2, we present the voice modification techniques which were employed in our study to resynthesize sample voices with different voice features. Section 3 describes the setup of the experiments that we conducted to explore language learners' imitation preferences. Experimental results and discussions are provided in Section 4. Section 5 concludes our present work and discusses our future work.

2. Voice modification

Based on a teacher's voice, our CASTLE system resynthesizes sample voices with different voice features (i.e., speech rate and pitch-formants) by voice modification. In the following, we identify the teacher's utterances as *original teacher's utterances*, and identify the resynthesized utterances as *individualized teacher's utterances*. The individualized teacher's utterances are automatically resynthesized based on the original teacher's utterances and learners' preferences. Our voice modification is implemented on the Praat platform (Boersma and Weenink, 2009).

2.1. Duration modification

Pitch-Synchronous Overlap and Add (PSOLA) algorithms (Moulines and Charpentier, 1990) allow us to compress or stretch an utterance in the time domain, and in the meantime to maintain its pitch values. PSOLA algorithms can be implemented in both the time domain and the frequency domain. Considering the low computational complexity of the time-domain PSOLA, the duration modification in CASTLE is implemented in the time domain.

2.2. Pitch-formant modification

The pitch-formant modification is composed of two steps: pitch modification and formant modification. Pitch modification changes the pitch median (*fMedianOri*) of an original teacher's utterance to a new pitch median ($fMedian_{Ind}$) according to a learner's preference. The new pitch values of the individualized teacher's utterance are calculated by multiplying the pitch values of the original teacher's utterance by Δf that is the ratio of the new pitch median and the old pitch median, as it is shown in Eq. (1).

$$\begin{aligned} newPitch &= oldPitch * \Delta f \\ \Delta f &= fMedian_{Ind} / fMedian_{Ori} \end{aligned} \tag{1}$$

The multiplication is to simulate the human auditory perception of pitch, which is more closely related to the logarithm of frequency (e.g., the semitone scale) than to frequency itself (Nolan, 2003). The following relationship between an old pitch value and a new one is to maintain the shape of a pitch contour in a perception scale (i.e., log scale),

$$\begin{aligned} Log(newPitch) = Log(oldPitch) + Log(fMedian_{Ind}) \\ - Log(fMedian_{Ori}) \end{aligned} \tag{2}$$

Thus, the relationship between the old and new pitch values in Eq. (2) can be expressed as follows:

$$newPitch = oldPitch * \frac{fMedian_{Ind}}{fMedian_{Ori}} \tag{3}$$

The pitch modification, as the reciprocal process of the duration modification, is also implemented by the time-domain PSOLA.

Generally speaking, the pitch median of a female voice is higher than that of a male voice. There is no consensus of the pitch boundary between female voices and male voices. Meszaros et al. (2005) indicated that the pitch cutoff point between female voices and male voices is around 140–170 Hz. A voice with a pitch median below 140 Hz is usually perceived as a male voice. A voice with a pitch median above 170 Hz is usually perceived as a female voice. Also, a voice with a pitch median between 140 Hz and 170 Hz can be a male voice or a female voice.

The pitch of a voice is also related to the formants of the vocal tract which produces the voice. Formants are the concentrations of acoustic energy around particular frequencies in a speech. According to the *source-filter* theory (Fant 1960), a speech signal is generated by a source signal passing through a filter. The source signal is a sequence of vibrations of a vocal cord. The filter is a vocal tract, which spectrally shapes the various sounds of speech. The vocal tract lengths of different people vary. Normally, the length of a woman's vocal tract is shorter than that of a man's. The formants of a shorter vocal tract are higher than these of a longer vocal tract. Thus, the formants of a female voice tend to be higher than these of a male voice. Therefore, formants modification can contribute to the perceived gender of an utterance.

In order to keep the resynthesized utterances natural, in CASTLE, if the pitch median of an utterance is changed from the female pitch range to the male pitch range, the formants of the utterance need to be decreased correspondingly. Similarly, if the pitch median of an utterance is changed from the male pitch range to the female pitch range, the formants of the utterance need to be increased.

In order to demonstrate the procedure of formants modification in our system, we take a male-to-female voice change as an example. The change from a female voice to a male voice can be processed in a similar way. As indicated by Clark et al. (2007, pp. 242), the length of a woman's vocal tract is about 80–90% of a man's vocal tract. Then the formants of a woman's vocal tract are about 1.1–1.25 times of the formants of a man's vocal tract. We take 1.2 as an example in the following description.

There are four steps in our formants modification. In the first step, the sampling frequency of an original teacher's utterance (U_0^1) is overridden by 1.2 times of its original sampling frequency. It makes the duration of U_1 compressed to 1/1.2 times of the duration of U_0, and shifts the formants and pitch values of U_1 to 1.2 times of these of U_0. In the second step, the duration of U_1 is lengthened to the duration of U_0, which can be implemented by the duration modification (refer to Section 2.1). In the third step, the pitch values of U_2 shift to their corresponding pitch values of U_0, which can be implemented by the pitch modification (refer to Section 2.2). In the last step, U_4 is generated by re-sampling U_3 at the sampling frequency of U_0. Then, the duration and pitch values of utterance U_4 are the same as these of U_0, but the formants of U_4 are 1.2 times of these of U_0.

Thus, the pitch-formant modification will change the perceived gender of a voice, if the pitch median of the voice is changed from the female pitch range to the male pitch range or vice-versa. The formant scale factor used in a formant modification has a linear relationship with the pitch median of the utterance, which is calculated by CASTLE system.

3. Setup of the experiments

The experiments are to investigate how the voice features (i.e., speech rate and pitch-formants) of teachers' voices influence learners' imitation preferences. We tested the following two hypotheses: (i) whether language learners prefer to imitate voices that sound like being produced by the same genders as themselves and possess similar pitches to their own voices; (ii) whether language learners prefer to imitate voices with speech rates close to their own voices. We expected that learners express a preference to imitate voices that are similar to their own voices in terms of gender and speech rate.

3.1. Speech material

The learning material was selected from the Boston University Radio News Corpus (BU-RNC) (Ostendorf et al., 1995), which consisted of continuous speech produced by FM radio news announcers associated with WBUR, a public radio station. Two paragraphs PRLP2, PRLP4 uttered by female native speaker F1A, were selected as the learning material. The utterances of the two paragraphs were segmented into short portions. The duration of each segmented short portions ranged between 2 s and 3 s. Ten utterances of sentences, which were selected from the segmented short portions, were taken as the original teacher's utterances. One utterance was used as an example to demonstrate the experiment procedure to subjects, and the other nine utterances were used as test utterances.

3.2. Subjects

Fifteen university students speaking English as a second language voluntarily participated in the test. Seven of them were male and eight were female. Seven of the subjects were aged between 20 and 29 years old; another seven of the subjects were aged between 30 and 39 years old, and the other one subject was older than 49 years old. Their first languages were: Hindi, Japanese, Persian, Spanish, Malay, Urdu ($N = 2$), and Chinese ($N = 8$). They had a history of learning English for between 5 and 30 years. The duration of their living in English speaking countries ranged between 4 months and 15 years.

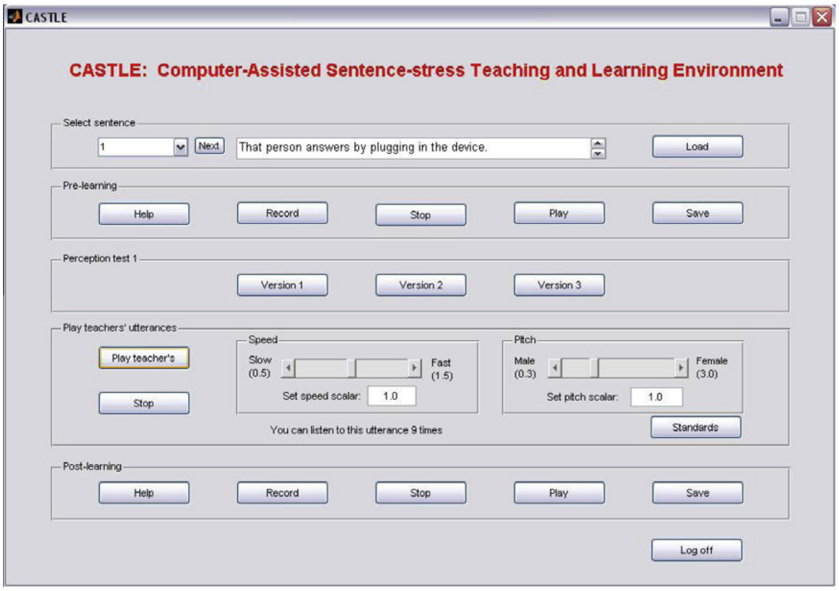
Before the test, the subjects were given a questionnaire about their English backgrounds. Twelve of them ranked their English speaking proficiencies no more than 3, on a 5-point scale, in which a score of 1 means very poor and a score of 5 means very good. Most of them ranked their English reading and listening capabilities higher than their English speaking capabilities. Ten of the subjects had been using imitation to improve their pronunciation. Three of the ten subjects claimed that when they chose imitation speech material, they preferred to imitate voices from the same gender of themselves, while the other seven subjects ranked reasonable speed as their first preference.

3.3. Procedures

The experiments were conducted on the CAPT system, CASTLE. A screenshot of CASTLE system is illustrated in Fig. 1.

For each sentence, the subjects were given its prompt, and asked to read it and record their utterances. CASTLE then detected the pitch median and speech rate of each subject's utterance. For each subject, CASTLE resynthesized individualized teacher's utterances based on the pitch median and speech rate of the subject's voice and the original teacher's utterance which was produced by native speaker F1A. There were three types of individualized teacher's utterances: speed similar and gender different utterances (SpS_PFD, in which PF stands for pitch-formant), speed different and gender same utterances (SpD_PFS), and speed similar and gender same utterances (SpS_PFS). For each subject, the SpS_ PFD utterances had similar speech rates to the subject's utterances and were perceived as being produced by a speaker whose gender is opposite to the subject. The SpD_PFS utterances sounded like being produced by a speaker whose gender was the same as the subject, and had similar pitch medians and different speech rates with the subject's utterances. Also, the SpD_PFS utterances were faster (slower) than the subject's voices, if the subject produced this sentence slower (faster) than the speed of the original teacher's utterance. The SpS_PFS utterances had similar pitch medians, formants and speech rates to the subject's utterances. Since the pitch medians and formants of SpS_PFS utterances and the subject's utterances were same, SpS_PFS utterances sounded like being produced by a speaker whose gender was the same as the subject. Sample utterances can be found at (WWW, 2010), which include original teacher's utterances and individualized teacher's utterances.

The orders of presenting the three individualized teacher's utterances are different for each learning sentence. The presenting order is in a loop format. For the first sentence of the learning material, the order was SpS_PFD, SpD_PFS and SpS_PFS. For the second sentence, the order was SpS_PFS, SpS_PFD and SpD_PFS. For the third sentence, the order was SpD_PFS, SpS_PFS and



CASTLE: Computer-Assisted Sentence-stress Teaching and Learning Environment

That person answers by plugging in the device.

Set speed scalar: 1.0

Set pitch scalar: 1.0

You can listen to this utterance 9 times

Fig. 1. Screenshot of CASTLE system providing individualized teacher's utterances with different voice features.

SpS_PFD. Then, for the fourth sentence, the order was the same as the order of the first sentence, and so on and so forth. The subjects were blind to the order of presenting the three individualized teacher's utterances.

Given each sentence in the speech learning material, each subject was asked to choose utterances that they most and least wanted to imitate from the three types of individualized teacher's utterances. For the "most wanted" label, the subjects were also allowed to choose *one* (*more than one* or none) utterances as the most wanted to be imitated, if there were one (more or none) individualized teacher's utterances favored by them to imitate. It is same for the "least wanted" label. The subjects could choose *one*, *more than one* or *none* utterances as the least wanted to be imitated.

For each sentence, when the subjects labeled their "most wanted" and "least wanted" to imitate utterances, they could select their "most wanted" individualized teacher's utterances to listen to and imitate. They could practice each sentence as many times as they desired. The subjects were also allowed to freely adjust the pitch medians and speech rates of the teacher's utterances by dragging the sliders or inputting scale factors as it is illustrated on the fourth panel in Fig. 1. When the pitch median was changed from the male (female) pitch range to the female (male) pitch range, the perceived gender of the resynthesized individualized teacher's utterances would change from male (female) to female (male). By allowing subjects to change pitch median and speed of the individualized teacher's utterances, we could observe whether their imitation preferences of a sentence change along with their practices.

4. Experimental results and discussion

The distributions of the most and least wanted to be imitated utterances labeled by the subjects are given in Fig. 2(a). Since for the three types of resynthesized individualized teacher's utterances of each sentence, a learner could label none, one or more than one utterance as the most (or least) wanted speech, totally there are 146

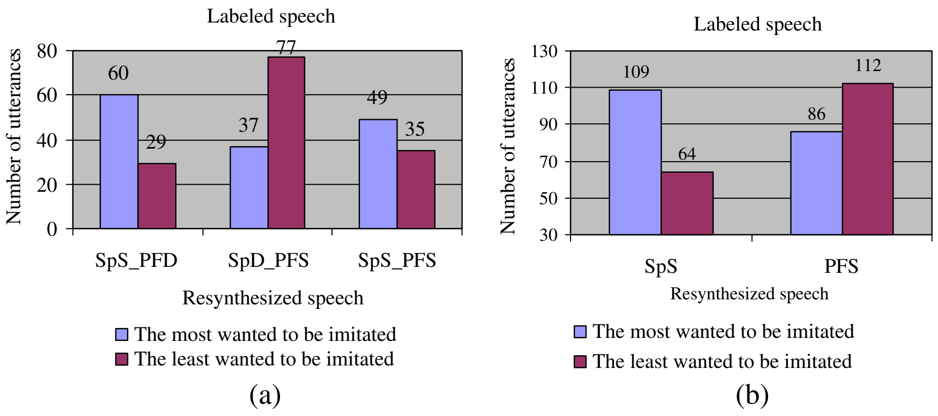


Fig. 2. Distributions of the most and the least wanted to be imitated speech.

utterances being labeled by the subjects as the most wanted to be imitated, and 141 utterances labeled as the least wanted to be imitated. Among the utterances being labeled as the most wanted to be imitated, 60 utterances are SpS_PFD utterances, which are more than the numbers of SpD_PFS utterances (37) and SpS_PFS utterances (49) labeled as "the most wanted utterances". In the utterances being labeled as the least wanted to be imitated, 29 are SpS_PFD utterances, 77 are SpD_PFS utterances, and 35 are SpS_PFS utterances.

From Fig. 2(a), we can see that in the SpS_PFD utterances, the number of utterances labeled as "the most wanted" is as twice as the number of the utterances labeled as "the least wanted". It demonstrates that voices, which have a similar speech rate to the subject's voice but sound as being produced by a speaker whose gender is opposite to the subject, are more positive on the subjects' imitation preferences. In the SpD_PFS utterances, the number of the utterances labeled as "the most wanted" is only about half of the number of the utterances labeled as "the least wanted". It shows that voices, which have the same gender as the subjects, but dissimilar speeds to the subjects' voices, cannot increase all learners' imitation interests. In the SpS_PFS utterances, the number of "the most wanted" utterances is slightly higher than the number of "the least wanted" utterances. Thus, not all learners always prefer to imitate voices with same gender (i.e., similar pitch and formants) and speech rates to their own voices.

We also noticed that all the SpD_PFS utterances resynthesized in our experiments were faster than the subjects' voices, since all the subjects' voices recorded in the pretest were slower than the original native speaker's utterances. This might be because (i) the native speaker F1A was a radio news announcer, who spoke fluently, and (ii) the subjects did not pronounce the learning material very fluently since the material was new to them. Since all the SpD_PFS utterances resynthesized in our experiments were faster than the subjects' voices, in the following we identify SpD_PFS utterances as having similar pitches and formants to the subjects' voices but higher speeds than the subjects' voices.

The influences of speech rate and pitch-formants of speech learning material on learners' imitation preferences are illustrated in Fig. 2(b). SpS refers to the resynthesized individualized teacher's utterances possessing similar speech rates to learners' utterances, which include both SpS_PFD and SpS_PFS utterances. Also, PFS are the resynthesized individualized teacher's utterances having the similar pitch medians and formants to the learners' utterances. The perceived gender of the learners' utterances and their corresponding PFS utterances are same. PFS includes both SpD_PFS and SpS_PFS utterances. From Fig. 2(b), we can see that in the SpS utterances, the number of utterances with "the most wanted" label is nearly as twice as the number of utterances with "the least wanted" label. This means that reasonable speed has a significant positive impact on the subjects' imitation preferences. Voices possessing similar speeds to the subjects' voices are more pleasant to be mimicked by them. In contrast, in the PFS utterances, the number of the utterances with "the most wanted" label is slightly lower than the number of the utterances with "the least wanted" label. It shows that similar pitch and formants between a teacher's voice and a learner's voice has a slightly negative influence on the subjects' imitation preferences.

Although the experimental results show that the subjects, as a whole, are more willing to imitate voices produced by an opposite gender to themselves with similar speeds to their own voices, the imitation preferences of different subjects also have diversities in the preference of the gender of the produced speech learning material. Five subjects preferred to imitate opposite gender voices. They labeled more than six (in nine) opposite gender voices as their "the most wanted" voices. Three of the five were female and two of them were male. Some subjects of the five claimed that to them, voices of the opposite gender sounded clearer than the voices of the same gender. Also, one of the five subjects stated that opposite gender voices were friendlier and less overwhelming. However, there are also two subjects preferred to imitate the same gender voices: one female and one male. The female subject labeled all the opposite gender voices as the "least wanted" voices, and a male subject labeled more than half of the opposite gender voices as the "least wanted" voices. The other *eight* participants did not show an obvious imitation preference on the speaker's gender of the speech learning material.

Learners' English backgrounds may have an influence on their imitation preferences. Four subjects chose to imitate voices which had a slightly faster speed than their own speech rates. Two of them were from Pakistan, one was from India, and the other one was from Japan. They all had good English listening proficiencies. For instance, the subjects from India and Pakistan, although their first languages were not English, they used English in some formal occasions in their home countries. They did not have any problem understanding radio news or TV programs. In contrast, seven subjects who labeled more than six fast teacher's utterances as "the least wanted" had a medium English proficiency. Except the eleven (4 + 7) subjects having preference to the speeches that were either slightly faster or slower than their own voices, the other four subjects did not have clear imitation preferences on speech rates.

There is no significant difference in imitation preferences between the subjects who had experience using imitation to improve their second language pronunciation and the subjects who had not. In the ten subjects who had experience using imitation for pronunciation learning, three of them showed preference to voices (SpS_PFD) of the opposite gender to themselves. Two of the ten subjects showed preferences to fast speed voices (SpD_PFS). Another two of the ten subjects showed preferences to the voices (SpS_PFS) which have the same gender and a similar speed to their own. Also, the other three of the ten subjects did not show an obvious imitation preference. In the five subjects who had never used imitation to improve their pronunciation, two of them preferred voices (SpS_PFD) of the opposite gender to themselves. One of the five subjects preferred voices (SpD_PFS) having a same gender and a faster speed to their own voices. One of the five subjects preferred voices (SpS_PFS) having a same gender and a similar speed to their own voices. Also, the other one of the five subjects showed no obvious imitation preference.

Also, in the experiments we find that some subjects changed their imitation preferences of an utterance along with their familiarity of the utterance increasing. In the experiments, for each sentence, after labeling their imitation preferences, the subjects were asked to imitate a resynthesized teacher's utterance, and they were also allowed to change the pitch and speech rate by dragging the sliders or inputting scale factors. We notice that at the beginning some subjects slowed down the speed of utterances and after several times practice, then they sped it up a little or changed it back to its normal speed. After experiments, they were interviewed by our experimenter. Some of them claimed that slowing down an utterance could help them catch the pronunciation features in the utterance, such as linking, assimilation and elision. When they were aware how the utterance was produced and could pronounce it fluently in a slow speed, they changed it back to the normal speed in order to imitate a more natural speech. Thus, a learner's speed requirement for speech learning material may change at different learning stages. However, further study is needed to investigate how the familiarity of speech material may influence learners' imitation preferences.

In order to analyze each subject as an individual, Table 1 lists the average of the absolute deviations from the mean of the number of utterances which were labeled as "the most wanted to be imitated", and the average of the absolute deviations from the mean of the number of utterances which were labeled as "the least wanted to be imitated". The average of the absolute deviations from the mean is calculated by Eq. (4)

Table 1
The average of the absolute deviations from the mean.

Subject No.	Average of the absolute deviation		Mean
	Most wanted	Least wanted	
1	2.00	4.00	3.00
2	2.00	3.33	2.67
3	1.56	3.78	2.67
4	2.22	3.11	2.67
5	2.67	2.00	2.33
6	2.00	2.44	2.22
7	2.44	1.78	2.11
8	2.00	2.00	2.00
9	1.78	2.00	1.89
10	1.78	1.78	1.78
11	2.00	1.33	1.67
12	1.78	1.56	1.67
13	2.00	1.11	1.56
14	1.78	1.33	1.56
15	0.67	1.33	1.00

$$AveDeviation = \sum_{i=1}^{n} |x_i - \bar{x}| \quad (4)$$

where $\bar{x}$ is the mean of $x_{1\ldots n}$. For example, if a subject chooses 3 SpS_PFD utterances, 6 SpD_PFS utterances and 0 SpS_PFS utterance as his/her the most wanted to be imitated utterances. The average of the absolute deviations from the mean (which is 3) is 2. The higher the average absolute deviation is, the stronger the imitation preference of the subject is. From Table 1, we can see that some subjects showed very strong imitation preference, such as subject No. 1 whose average absolute deviation for "the least wanted" label is 4. On the contrary, some subjects almost did not show any imitation preference. For example, for subject No. 15, the average absolute deviations for "the most wanted" label and "the least wanted" label are 0.67 and 1.33, respectively, which are very low.

Note that the quality of the resynthesized individualized teacher's utterances in our study is fairly good. After the experiments, the subjects were asked by our experimenter if they realized that all the listening material was generated from the same speaker. All of them answered no. When the subjects were told that all the teacher's utterances provided by CASTLE were generated by resynthesizing one female native speaker's voices, they were amazed. Few subjects realized that there were some minor distortions in the individualized teacher's utterances, but they felt that the individualized teacher's utterances were all still on an acceptable level. Their preferences of imitation seem unaffected by those unobvious distortions.

Our research findings are *consistent* with previous research. Eskenazi et al. (2000) observed that the golden voices chosen by some learners were very different from their own voices in terms of pitch and speech rate. For example, in the experiments of (Eskenazi et al., 2000), some female learners chose male voices as their golden voices, and some people with a low average pitch chose voices with much higher pitches as their golden voices. The experiments conducted by Probst et al. (2002) also show that given a choice, most learners chose voices dissimilar to their own voice to imitate. These observations reported by Eskenazi et al. (2000) and Probst's et al. (2002) confirm our research finding that voices with similar pitch-formants and speech rates to learners' own voices are not always what they prefer to imitate.

5. Conclusions and future work

In this paper, we have investigated what voice features (i.e., speech rate and pitch-formants) make a teacher's voice be a "golden voice" that is preferable for a language learner to imitate.

Our approach of searching the "golden voice" is different from the study conducted by Probst et al. (2002). Probst et al. investigated the "golden voice" from learners' pronunciation improvement perspective, while we investigated the "golden voice" from learners' imitation preference perspective. Providing learners' favorite voices is important for CAPT systems since it may help to develop a pleasant learning environment and increase learners' learning interests. Also, (Probst et al., 2002) takes multiple native speakers' voices as teacher's voices to provide to ESL learners for pronunciation learning. In our study, based on a single teacher's voice and learners' imitation preferences, individualized teacher's voices with different voice features were automatically resynthesized. Since the individualized teacher's voices in our study were generated from one teacher's voice, our approach can reduce the influence of characteristics of teachers' voices (e.g., clarity and accent) on the investigation.

Our experimental results show that a teacher's voice, which has same gender and similar speed to a learner's voice, is not always the learner's first imitation preference. Learners' imitation preferences can be influenced by many factors, e.g., English background and proficiency. Four out of fifteen subjects in our experiments preferred to listen to normal or fast voices. The possible explanation might be that the relatively strong English backgrounds of those four subjects contribute to their preferences of normal or fast voices. In our experiments, we also observed that learners might change their speed preferences of an utterance at different learning stages. Seven out of fifteen subjects in our experiments preferred a slow version of the speech material to catch their unfamiliar pronunciation features (e.g., linking, assimilation and elision). These seven subjects had a medium or low English proficiency. Thus, their relatively low English proficiency may be one of the reasons that they prefer a slow version of the voices to imitate. In our experiments, we also noticed that some of the subjects, who preferred a slow version of speech material, tended to speed up the speech material a little or switch it back to the normal speed, when they had caught the pronunciation features in these utterances. This tendency reflects the fact that their objectives of second language learning are to perceive and produce natural speech with a regular speed. Also, a number of (5 out of 15) subjects in our experiments were more willing to listen to voices produced by a speaker whose gender is opposite to themselves rather than by a speaker whose gender is the same as themselves, while few (2 out of 15) subjects were more willing to listen to voices produced by the same gender of themselves. A subject in our experiments claimed that voices of the opposite gender were more pleasant and less overwhelming. Thus, we conclude that different people may have different imitation preferences, and their imitation preferences of an utterance may change with their familiarity of the utterance increasing.

In order to meet learners' different imitation needs, we advocate an automatic voice modification function in CAPT systems to provide speech material with a wide variety of voice features (e.g., different speeds and different genders). For a CAPT system, automatic voice modification can be used to resynthesize speech learning materials with learners' preferred voice features. Learners then can have an opportunity of listening to voices with more variations.

In our present experiments, the subject group is relatively small. Thus, in our future work, we intend to expand the number of subjects and recruit subjects with a wider range of second language experience. Also, the subjects can be grouped according to their English backgrounds, English proficiencies, ages, etc, in order to investigate how these factors influence their imitation preferences.

Another topic worth studying is the relationship between learners' imitation preferences and their pronunciation improvement. We have investigated the relationship between learners' imitation preferences and two voice features (i.e., speech rate and pitch-formants). However, there is a lack of a clear evidence of how providing learners' preferred voices can actually help them to improve their pronunciation. Considering that pronunciation improvement is the ultimate goal of pronunciation learning, it is essential to investigate the relationship between learners' imitation preferences and their pronunciation improvement.

References

Arnett, M.K., 1952. Does the elementary teacher have time to teach speech?. J. Southern States Comm. Assoc. 17 (3) 203–208.

Bissiri, M.P., Pfitzinger, H.R., 2009. Italian speakers learn lexical stress of German morphologically complex words. Speech Comm. 51 (10), 933–947.

Black, A., 2007. Speech synthesis for educational technology. In: Proc. ISCA ITRW SLaTE Workshop on Speech and Language Technology in Education, Farmington, PA.

Boersma, P., Weenink, D., 2009. Praat: doing phonetics by computer (Version 5.1.05). <http://www.praat.org/> (retrieved 01.05.09.).

Clark, J., Yallop, C., Fletcher, J., 2007. An Introduction to Phonetics and Phonology, third ed. Blackwell Publishing.

Derwing, T.M., 2003. What do ESL students say about their accents? Can. Mod. Lang. Rev. 59 (4), 546–566.

Ding, Y., 2007. Text memorization and imitation: The practices of successful Chinese learners of English. System 35, 271–280.

Dyck, C., 2002. Review of Tsi Karhakta: At the edge of the woods. Lang. Learn. Technol. 6 (2), 27–33.

Erro, D., Moreno, A., 2007. Weighted frequency warping for voice conversion. In: InterSpeech 2007 – EuroSpeech.

Eskenazi, M., 2009. An overview of spoken language technology for education. Speech Comm. 51 (10), 832–844.

Eskenazi, M., Hansma, S., 1998. The Fluency pronunciation trainer. In: Proc. of Speech Technology in Language Learning 1998, pp. 77–80.

Eskenazi, M., Ke, Y., Albornoz, J., Probst, K., 2000. The fluency pronunciation trainer: update and user issues. In: Proc. InSTiLL Workshop on Speech Technology in Language Learning, Dundee, 1047 Scotland, pp. 73–76.

Fant, G., 1960. Acoustic Theory of Speech Production. Moutons'Gravenhage.

Felps, D., Bortfeld, H., Gutierrez-Osuna, R., 2009. Foreign accent conversion in computer assisted pronunciation training. Speech Comm. 51 (10), 920–932.

Jacob, A., Mythili, P., 2008. Developing a child friendly text-to-speech system. Advances in Human–Computer Interaction, Article ID 597971, 6 pages.

Hirose, K., 2004. Accent type recognition of Japanese using perceived mora pitch values and its use for pronunciation training system. In:

Proc. Internat. Symp. on Tonal Aspects of Languages (TAL), Beijing, pp. 77–80.

Hişmanoğlu, M., 2006. Current perspectives on pronunciation learning and teaching. J. Lang. Linguist. Studies 2 (1), 101–110.

Lee, S.T., 2008. Teaching pronunciation of English using computer assisted learning software: An action research study in an institute of technology in Taiwan. PhD thesis, Australian Catholic University.

Lu, J., Wang, R., De Silva, L.C., Gao, Y., Liu, J., 2010. CASTLE: a computer-assisted stress teaching and learning environment for learners of English as a second language. InterSpeech2010, Makuhari, Japan, pp. 606–609.

Meszaros, K., Vitez, L., Szabolcs, I., et al., 2005. Efficacy of conservative voice treatment in male-to-female transsexuals. Folia Phoniatr. Logo. 57, 111–118.

Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Comm. 9 (5–6), 453–467.

Nagano, K., Ozawa, K., 1990. English speech training using voice conversion. In: 1st Internat. Conf. on Spoken Language Processing (ICSLP 90), Kobe, Japan, pp. 1169–1172.

Nolan, F., 2003. Intonational equivalence: An experimental evaluation of pitch scales. In: Proc. 15th Internat. Congress of Phonetic Sciences, Barcelona, pp. 771–774.

Ostendorf, M., Price, P.J., Shattuck-Hufnagel, S., 1995. The Boston University Radio News Corpus. Boston Univ., Boston, MA, Tech. Rep. ECS-95-001, Mar.

Probst, K., Ke, Y., Eskenazi, M., 2002. Enhancing foreign language tutors—in search of the golden speaker. Speech Comm. 37 (3–4), 161–173.

Sundström, A., 1998. Automatic prosody modification as a means for foreign language pronunciation training. In: Proc. ISCA Workshop on Speech Technology in Language Learning (STILL 98), Marholmen, Sweden, pp. 49–52.

WWW, 2010. Voices materials on <http://www.box.net/shared/oa5ov0f1r7>.