CASTLE: a Computer-Assisted Stress Teaching and Learning Environment for Learners of English as a Second Language

Jingli Lu[1,2], Ruili Wang[1,2], Liyanage C De Silva[3,1], Yang Gao[2], Jia Liu[4]

[1]SEAT, Massey University, Palmerston North, New Zealand
[2] State Key Laboratory for Novel Software Technology, Nanjing University, China
[3]Faculty of Science, University of Brunei Darussalam, Brunei Darussalam
[4]Department of Electronic Engineering, Tsinghua University, China

R.Wang@massey.ac.nz

Abstract

In this paper, we describe the principle and functionality of the Computer-Assisted Stress Teaching and Learning Environment (CASTLE) that we have proposed and developed to help learners of English as a Second Language (ESL) to learn stress patterns of English language. There are three modules in the CASTLE system. The first module, individualised speech learning material providing module, can provide learners individualised speech material that possesses their preferred voice features, e.g., gender, pitch and speech rate. The second module, perception assistance module, is intended to help learners correctly perceive English stress patterns, which can automatically exaggerate the differences between stressed and unstressed syllables in a teacher's voice. The third module, production assistance module, is developed to help learners to be aware of the rhythm of English language and provide them feedback in order to improve their production of stress patterns.

Index Terms: Computer-Assisted Pronunciation Training (CAPT), English stress, Prosody.

1. Introduction

Recent studies in applied linguistics have shown that stress contributes greatly to intelligible spoken English [7]. Stress can be defined as the relative emphasis given to certain syllables in words in order to make them more prominent than others. Despite its importance, stress has been overlooked and marginalized in ESL teaching. Currently, pronunciation teaching in ESL more focuses on teaching individual sounds of vowels and consonants, especially those sounds that do not exist in learners' first language.

For some ESL learners, English stress is difficult to master. They tend to convey the stress patterns of their first language into English, which are not always appropriate for English. For example, Asian ESL learners are more likely to pronounce every syllable with the same length, no matter whether they are stressed or not [12].

For some ESL learners, it is also challenging to perceive English stress patterns correctly. The phonological properties of a listener's first language may influence his/her speech perception. For example, native speakers of French, as well as Finnish and Hungarian [13], encounter difficulties in distinguishing stress contrasts in English.

Stress perception and speech production are related to each other. Empirical evidences showed that improvements in perception could lead to improvements in speech production [8]. Although learners can perceive stress patterns relatively quickly, in order to produce them correctly, they need to have thorough and systematic learning and practice of the usage.

Considering the importance and difficulties of learning English stress, we have proposed and developed a Computer-Assisted Stress Teaching and Learning Environment (CASTLE) that is intended to help ESL learners correctly use stress patterns in English. In this paper, we describe the principle and functionality of CASTLE.

2. Three modules of CASTLE system

As mentioned before there are three modules in CASTLE: an individualised speech material providing module, an exaggeration-based perception assistance module, and a production assistance module.

2.1. Individualised speech material providing module

Imitation is a commonly used method to improve pronunciation, and also considered as one of the most effective methods [4]. However, it is not clear whether different voices (e.g., different genders, pitch medians, or speech rates) which produce the same learning material make a difference in pronunciation learning. In other words, what is the golden voice for a language learner to imitate?

2.1.1. Previous research on the golden voice

Some studies have suggested that learners can benefit from listening to their own voice producing native-like utterances since it may be easier for them to perceive differences between their own utterances and their native-like utterances [3]. 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 a learner's voice. However, this prosody transferring techniques keep segmental errors (e.g., mispronounced phonemes) of the learner's voice intact.

Voice conversion techniques [5] can potentially be used to modify a teacher's voice to make it sound as being produced by a learner, which transform a source speaker's voice to a target speaker's voice. However, the aim of voice conversion is to make a voice sound as if it was produced by the target speaker. Thus, the converted speech also preserves the accent of the target speaker, such as the foreign accent of a 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 and without errors. However, not all learners can speak accurately and fluently.

Also, some language educators advocate that Computer-Assisted Pronunciation Training (CAPT) systems should have multiple speakers' voices for learners to select, listen to and imitate, which cover a wide range of pitch medians and speech rates [14]. By listening to and imitating their favorite voices, learners might have a better perception of pronunciation. The survey conducted by Probst [14] shows that same gender, a reasonable speed and clarity of the speech are the most commonly mentioned criteria of selecting preferred learning utterances by ESL learners. However, providing multiple

teachers' voices increases the workload of recording speech learning materials and the storage space. Moreover, no matter how wide the range of the prosodic features of the teachers' voices covers, they cannot meet all learners' needs.

2.1.2. *Our research on searching of the golden voice*

It is important to provide learning material with learners' preferred voice features (e.g. gender, pitch and speech rate) since learners' preferred voice may help to create a positive learning environment and boost their learning interests. As indicated by Arnett [2], if a teacher speaks with a smooth, easy and pleasant voice, his/her students try to imitate his/her voice.

We have investigated what voice features (i.e., gender, pitch and speech rate) make a teacher's voice preferable for language learners to imitate. Our experimental results show that different people have different imitation preferences, and a teacher's voice, which has a similar pitch median and 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., their English background and their language proficiency.

2.1.3. *Prosody modification in CASTLE*

In order to provide learning material with learners' preferred voice features (e.g. gender, pitch and speech rate), CASTLE can modify the prosodic features of teacher's voices according to learners' preferences. The prosody modification in CASTLE is different from the prosody conversions in previous CAPT systems that map the prosodic features of a teacher's voice to a learner's voice. The prosody modification in CASTLE is based on teacher's voices. Thus, the resynthesized utterances can be free from segmental errors.

The prosody modification in CASTLE includes both pitch modification and duration modification, which are implemented by time domain Pitch-Synchronous Overlap and Add (PSOLA) algorithms [11]. The pitch of a voice is also related to the formants of the vocal tract that produces the voice. In general, the formants of a female voice are higher than those of a male voice. In order to keep the resynthesized utterances natural, if the pitch median of a voice is changed from a male (female) pitch range to a female (male) pitch range, the formants of the utterance will be increased (decreased) correspondingly.

The individualised speech learning material providing module of CASTLE also allows learners to adjust the pitch medians and speech rates of teacher's voices according to their preferences. By changing the pitch median of a voice, learners can also change the gender of the voice perceived.

2.2. Perception assistance module

In order to help ESL learners correctly perceive stress patterns in English, the CASTLE system has an exaggeration-based perception assistance module that can automatically enlarge the differences between stressed and unstressed syllables in a teachers' voice.

Prosodic features (i.e. pitch, duration and intensity) are important acoustic cues to indicate English stress [17]. Stressed syllables tend to be pronounced longer, louder, and with significant pitch movements. In the perception assistance module, there are three basic stress exaggeration methods: (i) pitch-based exaggeration, (ii) duration-based exaggeration, (iii) and intensity-based exaggeration. Also, (iv) a combined stress exaggeration that combines the three basic exaggeration methods is integrated into the CASTLE system.

2.2.1. *Pitch-based stress exaggeration*

Both pitch level and pitch movement can contribute to stress syllables. A stressed syllable may relate to a higher (or lower) pitch level, or a significant pitch movement. We have developed two types of pitch-based stress exaggeration techniques: (i) exaggeration based on pitch level, and (ii) exaggeration based on pitch movements.

In the stress exaggeration based on pitch level, the pitch level of a stressed syllable is increased (or decreased), if the syllable is stressed by a higher (or lower) pitch level. The pitch level of a stressed syllable is exaggerated by multiplying its pitch contour with a positive pitch-changing factor, Δf ,

$$\begin{aligned} newPitch &= oldPitch * \Delta f \\ \Delta f &= newMedian / oldMedian \end{aligned} \quad (1)$$

where Δf is the ratio of the new pitch median to the old pitch median. For a stressed syllable with a high pitch accent, Δf is set to be greater than 1, in order to increase its pitch level. For a stressed syllable with a low pitch level, Δf is set to be less than 1, in order to make its pitch level further lower. The pitch level is changed by multiplying old pitch values with pitch-changing factor Δf . This is to simulate the human auditory perception of pitch, which is more closely related to the logarithmic value of the frequency than to the frequency itself.

In the technique of stress exaggeration based on pitch movements, the pitch range of a stressed syllable is expanded by multiplying its pitch range with a scale factor, Δp , that is greater than 1.

$$newPitch = PitchMedian + (oldPitch - PitchMedian)\Delta p \quad (2)$$

The exaggeration based on pitch movements, vertically stretches the pitch contour of a syllable, and in the meantime, remains its pitch median as it is. The exaggeration based on pitch movements makes the stressed syllable have a wider pitch range.

In CASTLE system, the pitch-based stress exaggeration is either based on pitch level or based on pitch movement depending on the ToBI labels [16] of the stressed syllable. For a stressed syllable with H* or L* ToBI label, it is exaggerated based on pitch level. For a stressed syllable with other types of ToBI label, it is exaggerated based on pitch movements.

2.2.2. *Duration-based stress exaggeration*

The duration-based stress exaggeration is a technique that elongates the durations of stressed syllables and shortens the durations of unstressed syllables to make the differences between stressed and unstressed syllables more noticeable. Since stress has more influence on the duration of a syllable nucleus than the durations of a syllable onset and a syllable coda [1], in the CASTLE system, the duration exaggeration of the syllable nucleus is set to be greater than the duration exaggerations of the syllable onsets and syllable codas. For stressed syllables, we set the duration enhancing factor of syllable nuclei as 1.5, and it linearly reduces to 1 for the syllable onsets and the syllable codas. For unstressed syllables, we set the duration reducing factor of syllable nuclei as 0.8, and it linearly increases to 1 for the syllable onsets and the syllable codas.

2.2.3. *Intensity-based stress exaggeration*

Since the intensity differences between stressed and unstressed syllables lie in the high frequency band (i.e. above 500Hz) [17], in our intensity-based stress exaggeration, the intensity of a stressed syllable is increased by amplifying its high

frequency band, and the intensity of an unstressed syllable is decreased by compressing its high frequency band.

For stressed syllables, the intensity-based stress exaggeration is implemented by increasing 9dB of the high frequency band, which is greater than 500Hz. Increasing 9dB of the high frequency band is equivalent to multiplying the values in that region by a factor of 2.82, which is 10^(9/20). To reduce abrupt changes at the edge of the high frequency band, we smooth the increase by going from 1 to 2.82 linearly within a frequency band of 100Hz. For unstressed syllables, the intensity-based stress exaggeration is implemented by compressing 5dB of the frequencies above 500Hz.

2.2.4. Perceptual experiment results

We have conducted perceptual experiments to evaluate whether the stress exaggeration methods can help learners perceive stress patterns. Fifteen ESL learners voluntarily participated in our test. The experimental results show that all the three basic stress exaggeration methods, along with the combined stress exaggeration, have improved the accuracy of listeners' stress pattern labeling. Among the three basic exaggeration methods, the duration-based exaggeration is more effective than the other two methods. The exaggerated utterances generated by the combined method are more helpful to improve the listeners' stress perception than the exaggerated utterances generated by all the three basic exaggeration methods. The combined exaggeration method and the duration-based exaggeration method helped the listeners improve their stress labeling accuracy at the significance level of 0.01 for a Student's t-Test. The pitch-based and intensity-based exaggeration methods helped the listeners improve their stress labeling accuracy at the significance level of 0.05.

2.3. Production assistance module

In order to help ESL learners produce stress patterns correctly, we have designed a production assistance module that includes a clapping-based pronunciation practice assistance (CPPA) model and a stress-error feedback model.

2.3.1. Clapping-based practice model

Stress pattern is closely related to rhythm that is composed of regular occurrences of stressed syllables. Therefore, stress pattern learning can benefit from rhythm learning. Clapping has been used in classrooms to help ESL learners recognise the rhythm of English language. Given a teacher's utterance, our CPPA model can automatically resynthesise a clapping-based teacher's utterance by adding a clap to every stressed syllable of the original teacher's utterance. The syllable-level time alignments and stress labels of the original teacher's utterances are obtained by using automatic phoneme alignment [9] and stress detection techniques [15]. We also proposed and developed a linear regression based ensemble model for automatic phoneme alignment, which is more suitable for stress detection [10].

2.3.2. Stress-error feedback model

Stress is conventionally treated as a categorical concept, in which a syllable is either stressed or unstressed. However, stress is a subjective concept. Given an utterance to label the stress pattern of each syllable in it (i.e., stressed or unstressed), even for trained linguists, there is no guarantee that their answers would be exactly the same. The categorical representation of stress patterns is insufficient to represent the subjective nature of stress. Taylor argues that it is pointless to define and try to find the strict boundaries of suprasegmentals (e.g., stress) that are underlying continuous phenomena [18].

In order to represent the uncertainty of stress, we propose to extend the categorical representation of stress to a fuzzy representation. A fuzzy set of *stressed syllables* can be expressed as follows. Let S is a collection of syllables, for each object s in S, the membership value of s belonging to the fuzzy set of *stressed syllables* is given in Eq. (3)

$$\mu_{stressed}(s) = f(s) \tag{3}$$

where $f(s) \in [0,1]$. The higher the membership value of a syllable belonging to the fuzzy set of *stressed syllables* is, the more likely the syllable is *stressed*.

The stress differences between a teacher's utterance and learners' corresponding imitations are useful feedbacks to help learners realise and correct their stress errors in speech. However, it would be possibly destructive to provide feedback for every minor deviation since learners may be confused and quickly become discouraged.

In order to protect and boost learners' learning interests and provide more useful feedback, CASTLE is intended only to indicate learners' stress errors that have a more noticeable stress divergence from teachers' utterances. Thus, the CASTLE system only pinpoints the words or syllables, of which the membership value divergence between the teacher's utterance and the learner's utterance is greater than threshold σ .

$$W = \{s \mid \mid \mu_{stressed}(s)_t - \mu_{stressed}(s)_s \mid \geq \sigma\} \tag{4}$$

where $\mu_{stressed}(s)_t$ and $\mu_{stressed}(s)_s$ are the membership values of the teacher's utterance and the learner's utterance, respectively.

This feedback model is suitable to be employed to provide a various levels of feedback to learners. For example, a lower threshold can be used to provide strict feedback to advanced learners, while a higher threshold is more suitable for novices to get feedback for their serious stress errors in speech.

In order to avoid incorrect feedback, following Fluency [6], a pronunciation training system developed in Carnegie Mellon University, CASTLE system does not pass judgment onto learners, and only indicates specific words for learners to work on.

3. The CASTLE system

The flowchart of the CASTLE system is illustrated in Figure 1. The speech learning material in CASTLE can be any utterance, together with its word transcription. For each sentence in the learning material, in order to listen to an individualised teacher's voice that has similar voice features to the learner's own voice, the learner needs to read this sentence and record his/her voice. By analysing the learner's utterance, the CASTLE system gathers the learner's voice characteristics (i.e., pitch median and speech rate). Then by prosody modification, the CASTLE system can provide an individualised teacher's voice producing this sentence, which has a similar pitch median and a similar speech rate to the learner's voice. In order to resynthesise the learner's favourite voices for her/him to imitate, the CASTLE system also allows the learner to control the prosody modification by adjusting the pitch and speech rate changing factors.

If learners have difficulties in identifying the stress patterns of teacher's utterances, the CASTLE system can provide stress-exaggerated teacher's utterances that enlarge the differences between stressed and unstressed syllables in the teacher's utterances.

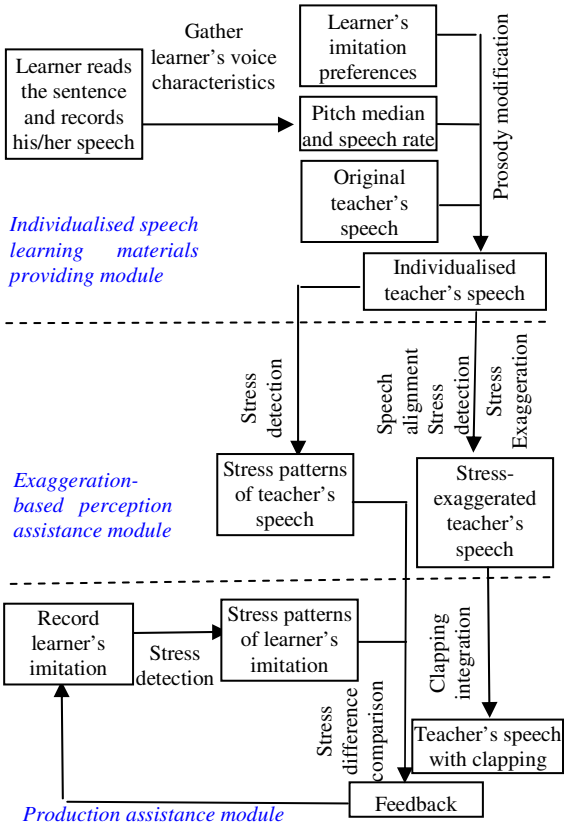

Figure 1: *Flowchart of the CASTLE system.*

In order to help learners to be familiar with the rhythms of English language and to train them to get used to placing more emphasis on syllables, which are supposed to be stressed, the CASTLE system resynthesises clapping-based teacher's utterances by adding a clap sound to every stressed syllable of the original teacher's utterances. The learners can then listen to and imitate the clapping-based teacher's utterances. Also, when they practice, they can clap their hands simultaneously with the clapping sound in the teacher's utterances. The stress detector in the CASTLE system can automatically obtain the stress patterns of the learner's imitation. By comparing the stress pattern differences between the teacher's utterances and the learner's imitations, the CASTLE system can then indicate specific words or syllables for learners to work on in order to improve their pronunciation. A screencopy of the CASTLE system is given in Figure2.

4. Conclusions and future work

In order to help ESL learners correctly use English stress patterns, we have proposed and developed the CASTLE system. We presented the three modules of CASTLE, which provides learning assistances from speech perception to speech production, in addition to offering individualised speech learning material possessing learners' preferred voice features. With the assistance of the CASTLE system, learners are expected to see an improvement on their language stress perception and speech production ability.

We have conducted preliminary tests of the individualised speech learning material providing module and the perception assistance model. The results are very encouraging. Learners found that the individualised speech learning material providing module is very helpful. The exaggeration-based perception assistance module significantly improved learners stress perception ability at the significance level of 0.01. Our future work is to refine our learning strategy and test the CASTLE system as a whole.

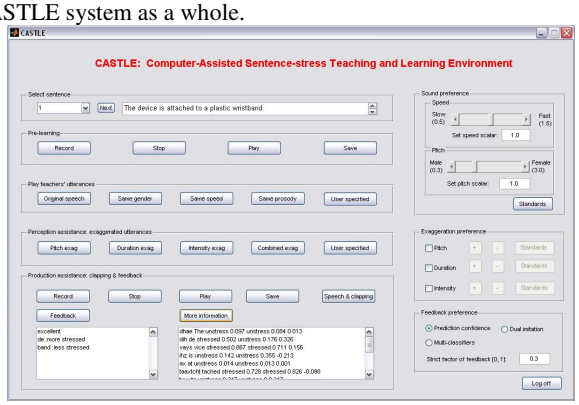

Figure 2: *Screencopy of the CASTLE system.*

5. References

[1] Ananthakrishnan, S., Narayanan, S., (2008). Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. IEEE Trans. Audio, Speech and Language Proc. 16(1) 216-228.

[2] Arnett, M.K., (1952). Does the Elementary Teacher Have Time to Teach Speech? Journal of the Southern States Communication Association, 17(3), 203-208.

[3] Bissiri, M. P. and Pfitzinger, H. R. (2009). Italian speakers learn lexical stress of German morphologically complex words. Speech Communication, 51 (10), 933-947.

[4] Ding, Y. (2007). Text memorization and imitation: The practices of successful Chinese learners of English. System, 35 (2), 271-280.

[5] Erro, D. and Moreno, A. (2007). Weighted frequency warping for voice conversion. In Proc. EuroSpeech.

[6] Eskenazi, M., Ke, Y., Albornoz, J., Probst, K., (2000). The fluency pronunciation trainer: update and user issues. In: Proc. InSTiLL, Dundee, Scotland, pp. 73–76.

[7] Field, J. (2005). Intelligibility and the listener: The role of lexical stress. Tesol Quarterly, 39(3), 399-423.

[8] Hincks, R. (2002). Speech synthesis for teaching lexical stress. TMH-QPSR, 44, 153-156.

[9] Hosom, J.-P. (2002). Automatic phoneme alignment based on acoustic-phonetic modeling. In Proc. International Conference on Spoken Language Processing, 357-360.

[10] Lu, J., Wang, R., De Silva, L. C. and Gao, Y. (2009). Syllable nucleus durations estimation using linear regression based ensemble model. In Proc. ICASSP, Taipei, Taiwan, 4849-4852.

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

[12] Nation, I. S. P. and Newton, J. (2008). Teaching ESL/EFL listening and speaking. New York: Routledge.

[13] Peperkamp, S. and Dupoux, E. (2002). A typological study of stress 'deafness'. Laboratory Phonology, 7, 203-240.

[14] Probst, K., Ke, Y. and Eskenazi, M. (2002). Enhancing foreign language tutors - in search of the golden speaker. Speech Communication, 37 (3-4), 161-173.

[15] Rosenberg, A. and Hirschberg, J. (2009). Detecting Pitch Accents at the Word, Syllable and Vowel Level. NAACL-HLT, 81-84.

[16] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., 1992. ToBI: a standard for labeling english prosody. In: Proc. ICSLP, 867–870.

[17] Sluijter, A. M. C., van Heuven, V. J. and Pacilly, J. J. A. (1997). Spectral balance as a cue in the perception of linguistic stress. J. Acoust. Soc. Amer, 101, 503-513.

[18] Taylor, P. (2000). Analysis and synthesis of intonation using the tilt model. J. Acoust. Soc. Amer. 107 (3), 1697-1714.