CASTLE: a Computer-Assisted sentence Stress Teaching and Learning Environment

A thesis presented
in partial fulfilment of the requirements
for the degree of
Doctor of Philosophy
in Computer Science
at Massey University, Manawatu
New Zealand

Jingli Lu

2010

Abstract

A Computer-Assisted sentence Stress Teaching and Learning Environment (CASTLE) is proposed and developed, in order to help learners of English as a Second Language (ESL) to perceive and produce English stress correctly.

Sentence stress plays an important role in English verbal communication. Incorrect stress may confuse listeners, and even break down a conversation. Stress is also challenging for ESL learners to master. It is neither easy for them to produce nor easy to perceive stress. Learners tend to transfer the stress patterns of their first language into English, which is not always appropriate. However, stress has been overlooked in English language teaching classes, due to the time limits of a class and teachers' lack of confidence of teaching stress. Thus, CASTLE is intended to help ESL learners to use sentence stress correctly.

There are three modules in CASTLE: an individualised speech learning material providing module, a perception assistance module and a production assistance module.

Through conducting an investigation into which voice features (i.e. gender, pitch and speech rate) makes a teacher's voice preferable for learners to imitate, we find that learners' imitation preferences vary, according to many factors (e.g. English background and language proficiency). Thus, the speech material providing module of CASTLE can provide individualised speech material for different learners, based on their preferred voice features.

In the perception assistance module of CASTLE, we propose *a set of* stress exaggeration methods that can automatically enlarge the differences between stressed and unstressed syllables in teachers' voice. These stress exaggeration methods are implemented by the manipulation of different prosodic features (i.e. pitch, duration and intensity) of the teachers' voice. Our experimental results show that all our proposed exaggeration methods could help ESL learners to perceive sentence stress more accurately.

In the production assistance module of CASTLE, we propose a clapping-based sentence stress practice model that is intended to help ESL learners to be aware of the rhythm of English language. By analysing the limitation of conventional categorical representation of stress, we propose a fuzzy representation which is intended to better represent the subjective nature of stress. Based on the fuzzy representation of stress, we propose three feedback models in order to help the learners correct their stress errors.

In addition to the development of CASTLE, we also propose an enhanced fuzzy linear regression model which can overcome the spreads increasing problem encountered by previous fuzzy linear regression models.

Dedicated to my parents for their love,
encouragement and endless support.

Acknowledgements

I would like to take this opportunity to express my appreciation and gratitude to those people who have supported me to achieve this qualification.

My first sincere thanks go to my supervisor, Dr. Ruili Wang, for his invaluable guidance and tremendous support throughout this research. Without his tireless directions and continuing encouragement, it would have been unfeasible for me to achieve my PhD degree. His intellectual rigour and logical way of thinking have had a remarkable influence on my academic career.

I would like to express my deep gratitude to my co-supervisors Dr. Liyanage C. De Silva and Dr. Helen Zhou, for the time and effort they have spent with me, during my PhD study. I appreciate their valuable suggestions and constructive comments.

I am also grateful to Dr. Shichao Zhang, my previous supervisor for my Master's degree, who introduced me to the field of Computer Science and put my footsteps onto the research path.

I thank Claire, Rosalind and all the participants for their help in the system evaluation. Thanks to Jason, Frank, Yan and June, and other friends at Massey University, for their support and friendship.

I gratefully acknowledge the funding from the *Foundation for Research, Science and Technology* towards my study and research.

Lastly, my special thanks go to my parents for their support, understanding and encouragement.

Contents

List of figures

List of tables

Chapter 1.

Introduction and Scope

1.1 Introduction

In this research, a Computer-Assisted sentence Stress Teaching and Learning Environment (CASTLE) is proposed and developed, which is intended to help learners of English as a Second Language (ESL) improve their ability to correctly use sentence stress. The development methods of CASTLE are also presented. Sentence stress is the relative emphasis given to certain syllables in words, in order to make them more prominent than others (Yavas, 2006). English speakers use stress to indicate new and contrastive information (Hahn, 2004), and to help them be understood.

Modern transportation and telecommunication make it more convenient, for people coming from different countries, to meet and talk with each other. Consequently, English has become an international language and verbal communication in English is playing an increasingly important role in today's society. Thus, pronunciation attracts more attention in ESL teaching and learning.

With the development of speech processing techniques and the popularity of personal computers, Computer-Assisted Pronunciation Training (CAPT) is becoming an alternative to the traditional classroom-based student-teacher model (Eskenazi, 2009). CAPT can provide a private and stress-free learning environment, and it also allows the learners to learn anytime and anywhere, where a computer is available (Neri, et al., 2002).

Current CAPT systems are more focussed on teaching *segmental features* (Derwing and Rossiter, 2002; Seferoğlu, 2003) that refer to the individual sounds of vowels and consonants (Dalton and Seidlhofer, 1994), since segmental features are the emphases of traditional foreign language teaching (Seferoğlu, 2005). Learners are asked to practise

individual sounds in isolated words, especially those sounds that do not exist in their first language.

However, in continuous speech, *suprasegmental features*, which refer to speech features stretching over more than one sound and up to whole utterances (Dalton and Seidlhofer, 1994), are also important. Suprasegmental features consist of stress, intonation and rhythm. *Stress* is the emphasis placed on a syllable, in order to make it more outstanding than others. *Intonation* (or *tone*) is the rising or falling pitch of a voice. Speakers use different intonations to convey different meanings. *Rhythm* is composed of the strong and weak elements in an utterance. Different aspects of suprasegmental features are related to each other. For example, some stressed syllables are perceived as capturing significant pitch movements, and rhythm consists of the occurrence of stressed syllables.

Studies in applied linguistics have found that *suprasegmental features*, especially stress, contribute greatly to mutual understanding (Bond, 1999; Field, 2005). According to Bond (1999), native listeners depend considerably more on stressed syllables than unstressed syllables to distinguish words. The study conducted by Bond and Small (1983) shows that a misplaced stress is three times more likely to break down communication than a mispronounced phoneme. Intelligibility can be damaged by a shift of stress, especially a rightward shift (Cutler and Clifton, 1984; Field, 2005) which is the stress shifting from a syllable to its right-hand side syllable, such as *wal*let changing into wal*let*[1].

Despite its importance, stress, as well as other aspects of suprasegmental features (i.e. intonation and rhythm), has been overlooked and marginalised in ESL teaching and learning. The time limits of a class (Eskenazi, 1999) may partially explain the lack of suprasegmental features teaching within classes. Also, some non-native English-speaking teachers are not used to (or reluctant to) teach suprasegmental features, due to their lack of professional knowledge and/or confidence in teaching suprasegmental features (Gong, 2002; Ilčiukienė, 2005).

[1] In this thesis, we use *Italic* font to indicate a syllable being stressed.

Thus, in this research, in order to help ESL learners to correctly use sentence stress, we propose and develop a computer-assisted sentence stress teaching system, CASTLE.

In addition to the development of CASTLE system, in this thesis, we also propose an enhanced fuzzy linear regression model that can overcome the *spreads increasing problem* encountered by previous fuzzy linear regression models.

1.2 Scope of this thesis

This thesis is organised as follows:

Chapter 2: *Motivation and Research Objectives*. In this chapter, we demonstrate the importance of sentence stress for successful English verbal communication, and we discuss the difficulties in learning stress, which are faced by ESL learners. We then propose the CASTLE system which is intended to help ESL learners improve their ability to correctly use sentence stress. The research issues and proposed solutions of developing CASTLE are also presented.

Chapter 3: *Speech Processing Techniques for CASTLE*. Two foundational speech processing techniques, for the development of CASTLE, are presented in this chapter, i.e. automatic phoneme alignment and stress detection. We review previous work on automatic phoneme alignment and stress detection. We also propose a linear-regression-based flexible boundary phoneme alignment algorithm, and develop an acoustic-feature-based stress detector, which are more suitable for the application of CASTLE.

Chapter 4: *Individualised Speech Material Module*. In this chapter, we investigate which voice features (i.e. gender, pitch and speech rate) make a teacher's voice preferable for language learners to imitate. We then advocate the use of prosody modification techniques, in order to automatically transform original speech learning material into individualised speech material for different learners, which have the learners' preferred voice features.

Chapter 5: *Exaggeration-based Perception Assistance Module*. In order to help ESL learners better perceive stress patterns of English speech, a set of stress exaggeration methods is proposed in this chapter, which can enlarge the differences between stressed and unstressed syllables in teachers' utterances. Perceptual experiments are conducted to evaluate the effectiveness of the proposed stress exaggeration methods.

Chapter 6: *Production Assistance Module*. In order to help ESL learners to become familiar with the stress patterns of English speech, a clapping-based pronunciation practice model is developed in the production assistance module. We also analyse the limitations of the conventional categorical representation of stress, and propose a fuzzy representation of stress. Based on this fuzzy representation, three feedback models are proposed in order to help learners correct their stress errors.

Chapter 7: *An Enhanced Fuzzy Linear Regression Model*. In addition to proposing and developing the CASTLE system, we also propose a fuzzy linear regression model with more flexible spreads. Our model can overcome the *spreads increasing problem* which has been encountered by previous fuzzy linear regression models.

Chapter 8: *Conclusions and Future Work*. This chapter summarises the main findings and conclusions of our present research. It also reviews our contributions and presents suggestions for future research directions.

Chapter 2.

Motivation and Research Objectives

In this chapter, we discuss, in detail, the reasons why our research focuses on computer-assisted sentence stress learning, and define our research objectives. Section 2.1 demonstrates the importance and difficulties of learning English stress. In Section 2.2, we propose to develop a Computer-Assisted sentence-Stress Teaching and Learning Environment (CASTLE), and we also present our research issues and propose solutions for the development of CASTLE.

2.1 Motivation

2.1.1 Importance of English stress

The importance of pronunciation, in foreign language learning, has been recognised by teachers and learners (Derwing, 2003), since verbal communication between people from different countries has become frequent, with the development of economic globalisation. Good pronunciation can enable listeners to understand more easily, while bad pronunciation may become a barrier to verbal communication, or even break down a conversation. Thus, language learners are encouraged to improve their pronunciation, at least to a certain intelligible level (Hişmanoğlu, 2006).

Intelligibility, instead of accent-free, has been recognised as an appropriate goal for pronunciation teaching and learning (Field, 2005). Foreign accents are quite common in second language speakers who begin to learn a second language post-puberty (Derwing and Munro, 2005). Although empirical studies have shown that a foreign accent could be reduced, by intensive pronunciation practice (Bongaerts, 1999), accent-free could only be achieved (extremely) rarely by adult second language learners (Flege, et al., 1995; Scovel, 2000). Moreover, a strong foreign accent does not necessarily make speech difficult to be understood, and the reduction of an accent does not always make speech easier to be understood (Munro and Derwing, 1999). Considering the difficulty

of achieving accent-free pronunciation, in addition to its function in verbal communication, intelligibility (i.e. to be understood) has been considered as the appropriate goal for second language pronunciation learning.

Studies in applied linguistics have shown that, compared with segmental features, suprasegmental features contribute considerably to the intelligibility of spoken English (Bond, 1999; Field, 2005). For a description of segmental and suprasegmental features, refer to Section 1.1. Segments, suprasegmentals and syllable structure all have a significant impact on pronunciation ratings, while suprasegmentals have the strongest effect (Anderson-Hsieh, et al., 1992). The study conducted by Derwing et al. (1998) showed that the group, which received suprasegmental instructions, achieved significant improvements in a spontaneous picture narrative task.

Among these suprasegmental features, stress performs an important role in English language intelligibility (Bond, 1999; Field, 2005). Hahn (2004) found that listeners tended to capture more details and rate speakers more highly, when the speakers used primary stress correctly. According to Bond (1999), stressed syllables play a more important role than unstressed syllables for native listeners to differentiate words. A stress error is more likely to break down communication than a phoneme error (Bond and Small, 1983, Field 2005). Also, the meanings of words may change depending on which syllable is stressed (Deshmukh and Verma, 2009). For example, *pre*sent is a noun, which means a gift, while pre*sent* is a verb, which means to give something to somebody. A shift of stress, especially a rightward shift, can damage the intelligibility of speech (Cutler and Clifton, 1984; Field, 2005).

The study conducted by Negrin-Cristiani (1997) shows that suprasegmental errors have a greater impact on verbal communication, probably due to native English speakers' lack of tolerances towards suprasegmental errors. Native speakers' early possession of suprasegmentals makes them believe that suprasegmentals are naturally understood. They do not recognise the importance of suprasegmentals, and they are not aware of the difficulties that non-native speakers have in using suprasegmentals. Thus, they have a low tolerance for suprasegmental errors. In contrast, native speakers have more experiences of learning vocabulary and the pronunciation of individual sounds, which make them more forgiving of segmental errors. As Field (2005) indicated, for

occasional segmental errors, native listeners could easily find the best match for the mispronounced words.

2.1.2 Difficulties in learning English stress faced by ESL learners

Despite its importance, stress, as well as other aspects of suprasegmentals (i.e. intonation and rhythm), has been overlooked and marginalised within ESL teaching and learning. A survey conducted by Derwing and Rossiter (2002) found that only eight learners from a group of the 100 adult intermediate ESL learners who had been enrolled in a full time ESL programme in a local college for extended periods of time, had received pronunciation instructions. The limits of time and the size of a class may partially explain the lack of pronunciation teaching in English language classes (Eskenazi, 1999). Even if pronunciation is taught in English language classes, it is more likely to focus on segmental features since practising segments has been the emphasis of traditional foreign language teaching (Seferoğlu, 2005). Thus, ESL learners have to learn English stress patterns by themselves.

It is challenging for ESL learners to produce English stress correctly. ESL learners tend to transfer the stress patterns of their first language onto the English language, and this is not always appropriate. It usually interferes with the learners' ability to produce English-like stress correctly (Hahn, 2004, and references in there). Asian ESL learners are more likely to pronounce every syllable with the same length, no matter if it is stressed or unstressed (Nation and Newton, 2008) since most Asian languages are *syllable-timed* languages, in which every syllable has roughly an equal weight and duration (Yavas, 2006). In contrast, English is a *stress-timed* language, in which stressed syllables occur at approximately equal time intervals (Yavas, 2006, p. 21). Then some syllables have to be spoken very quickly if there are several syllables between two stressed ones, while some syllables have to be spoken slowly if there are few syllables between two stressed ones. Todaka (1990) found that Japanese learners of English seemed unable to employ a sufficient pitch range to produce stressed syllables to the same degree as native English speakers. Juffs (1990) found that Chinese learners of English tended to stress every word in a message unit, no matter whether it had semantic importance or not. Wennerstrom (1994) found that Thai, Japanese, and

Spanish ESL learners used an inadequate pitch movement to produce new or contrastive information which needs to be emphasised. As indicated by Hahn (2004), two major problems of ESL learners were misplacing stress and stressing all words more or less equally in an utterance, without one prominent stress.

For some ESL learners, it is also difficult to correctly perceive English stress. As indicated by Peperkamp and Dupoux (2002), the phonological properties of a listener's native language could influence his/her speech perception. Research found that native speakers of French (Dupoux, et al., 1997), as well as Finnish and Hungarian (Peperkamp and Dupoux, 2002), encountered great difficulties in distinguishing stress contrasts in English. Wang (2008) found that all changes of pitch, duration and intensity had a considerable effect on native English speakers' stress perception, but it seemed that only pitch change affected Chinese learners' stress perception. Therefore, helping ESL learners correctly identify English stress is an important first step of helping them correctly produce stress.

These two major difficulties in learning English stress faced by ESL learners (i.e. stress perception and production) are related to each other. Empirical evidences showed that improvements in perception could lead to improvements in production (Akahane-Yamada, et al., 1996; Bradlow, et al., 1997; Hincks, 2002). Although learners can perceive stress relatively quickly, in order to produce it correctly, learners need to have thorough and systematic learning and practice (Jenkins, 1998).

2.1.3 Current computer-assisted pronunciation teaching

The importance of suprasegmentals in verbal communication is becoming recognised. However, most current CAPT systems are still limited to segments teaching, rather than suprasegmentals (Derwing and Rossiter, 2002; Seferoğlu, 2003). According to the study conducted by Hincks (2003), this shortage was because automatic speech recognition technologies could only give feedback on the segmental level not on the suprasegmental level, and speech processing technologies were still not ready to handle the prosodic information contained in a speech signal.

Although there are a few CAPT systems that can support suprasegmentals teaching, they also have some shortcomings. For example, some CAPT systems (e.g. WinPitchLTL, VisiPitch) only provide graphic display feedbacks, such as energy and pitch contours of teacher's (or learner's) utterances. A criticism of this type of feedback is that it is a technology push, not a demand pull. This type of simple graphic display feedback is the representations of raw data that requires learners themselves to interpret, even if they may lack the necessary knowledge to perform this challenging task (Neri, et al., 2002).

Some CAPT systems (e.g. *Pro-Nunciation*, *Eyespeak*) show two comparable waveform displays: one according to a teacher's utterance and one according to a learner's utterance. This incorrectly suggests that the goal of pronunciation training is to produce an utterance, the waveform of which closely corresponds to that of the teacher's utterance. In fact, this is not necessary, since two utterances with the same content may both be pronounced very well and still have quite different waveforms (Neri, et al., 2002).

Some CAPT systems (e.g. WinPitchLTL) contain a synthesis feature that allows a teacher to manually modify the prosodic features of a learner's utterance. Then, the learner can hear his/her own voice with correct prosodic contour. However, the teacher needs to have been trained in phonetics and acoustics. Moreover, modifying the prosodic features manually is a tedious and time-consuming task. Once again, it would be dependent upon a relatively low teacher-to-student ratio.

Apart from graphic displays of prosodic contours (e.g. energy contours or pitch contours), some CAPT systems (e.g. *FluSpeak*, *Accent Coach*) also provide a score (e.g. in the form of percentage or degree) that describes the similarity between teacher's speech and learner's speech. However, these systems do not provide learners with any other information about how to correct their pronunciation errors.

Also, most of these CAPT systems, which support suprasegmentals learning, are focussed on intonation and rhythm. Very few CAPT systems teach English stress. Thus, in order to narrow the gap between the demands in second language learning and the supply of current CAPT systems, we propose and develop a computer-assisted sentence

stress teaching system. For a more detailed review of current CAPT systems, refer to our technical report (Lu, et al., 2007).

2.2 Computer-Assisted sentence Stress Teaching and Learning Environment (CASTLE)

Considering the importance and difficulties of learning English stress, the intention of this thesis is to develop a Computer-Assisted sentence-Stress Teaching and Learning Environment (CASTLE) that aims to help ESL learners correctly use sentence stress. In order to develop CASTLE, three research issues are defined, and proposed solutions are presented. In the process of designing CASTLE, we also induce a framework for sentence stress teaching systems.

2.2.1 Research issues and proposed solutions

Perception assistance

The precondition of properly using sentence stress is to identify stress correctly. However, for some ESL learners, correctly perceiving stress is not an easy task. Some learners may not be sensitive to stress (Dupoux, et al., 1997; Peperkamp and Dupoux, 2002). Some learners use the acoustic cues to perceive stress, which are different from the acoustic cues used by native-English speakers (Wang, 2008). These all result in ESL learners not being able to correctly perceive English stress.

Thus, in order to help ESL learners to correctly perceive stress, we equip our CASTLE system with a perception assistance module. In the perception assistance module, we propose a set of stress exaggeration methods which can automatically resynthesise speech learning material with exaggerated prosodic features (i.e. pitch, duration and energy). The stress exaggeration techniques enlarge the differences between stressed and unstressed syllables.

Production assistance

Due to the influence of their first language, some ESL learners tend to stress every word equally or misplace stress (Hahn, 2004). These inappropriate stress and stress errors may make the learners' speech difficult to be understood. Moreover, prompt feedback for learners' stress errors is necessary for production-orientated pronunciation practice. Otherwise, stress errors, which are not rectified promptly, are at risk of becoming fossilised (Hincks, 2003).

In order to help ESL learners to become familiar with the stress patterns of English speech and produce stress correctly, we develop a production assistance module in CASTLE. In this production assistance module, we present a clapping-based pronunciation practice assistance model that is used to help learners become aware of English stress patterns and train them to stress those syllables which are supposed to be stressed. Instead of the conventional categorical representation of stress, we propose a fuzzy representation which is intended to better represent the subjective nature of stress. Based on this fuzzy representation of stress, we propose three feedback models which are intended to help the learners correct their stress errors.

Individualised Speech Material

Apart from the perception and production assistance, another important issue of CAPT systems is what voices are suitable as a teacher's voice for language learners to imitate. Language educators and teachers advocate that CAPT systems should have a number of teacher's voices for learners to select, listen to and imitate, which cover different genders and a wide range of pitch and speech rates (Dyck, 2002; Lee, 2008; Probst, et al., 2002). By listening to and imitating their favourite voices, learners might have a better perception of pronunciation, and their learning interests might also be improved. Some studies have also suggested that language learners may 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 (Bissiri and Pfitzinger, 2009; Sundström, 1998).

Thus, the third research issue is to investigate what makes a teacher's voice preferred by language learners. Based on this information, we equip our CASTLE system with an individualised speech learning material providing module which can provide learning material with learner's preferred voice features (i.e. gender, pitch and speech rate). In CASTLE, the individualised speech learning material can be automatically resynthesised from one original teacher's voice, according to each learner's imitation preference.

2.2.2 A framework for sentence stress teaching systems

In the process of designing the CASTLE system, based on the review of the needs and difficulties faced by ESL learners (refer to Section 2.1), we induce a framework for sentence stress teaching systems.

Firstly, a sentence stress teaching system, as well as other CAPT systems, should be able to provide speech learning material with learners' preferred voices features (e.g. gender, pitch and speech rate), since this may be helpful to create a friendly learning environment and increase learners' learning interests.

Secondly, since correct perception of sentence stress is difficult for some ESL learners, a sentence stress teaching system should be equipped with a perception assistance module which can help learners acoustically distinguish stressed words (or syllables) from unstressed ones.

Thirdly, it is essential that a sentence stress teaching system includes a production assistance module that is to help ESL learners correctly use prosodic features (e.g. pitch, duration and intensity), in order to correctly emphasis the words (or syllables) needed to be stressed.

Therefore, a sentence stress teaching system should be able to provide ESL learners with (i) individualised speech learning material, (ii) perception assistance and (iii) production assistance.

2.2.3 Flowchart of CASTLE system

The flowchart of CASTLE system is illustrated in Figure 2.1.

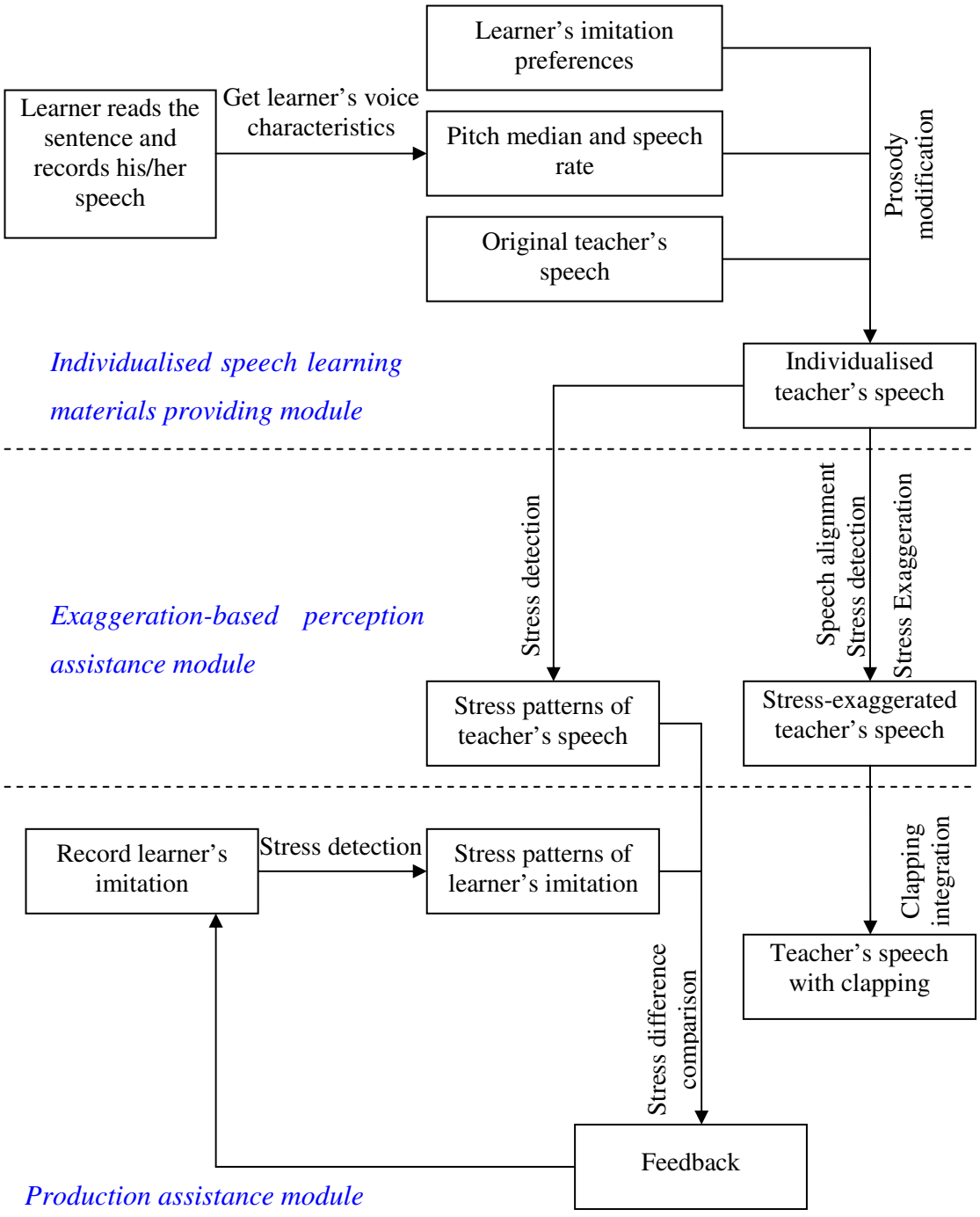


Figure 2.1 Flowchart of CASTLE system

The speech learning material in CASTLE can be any utterances, together with their word transcriptions. For each sentence in the learning material, in order to listen to an

individualised teacher's voice which 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, CASTLE gathers the learner's voice characteristics (i.e. pitch median and speech rate). Then by prosody modification, CASTLE 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, CASTLE 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, CASTLE can provide stress-exaggerated teacher's utterances which enlarge the differences between stressed and unstressed syllables in the teacher's utterances. In order to resynthesise the stress-exaggerated utterances, the syllable-level time alignments and stress labels of the original teacher's utterances are needed. If the human-labeled time alignments and stress labels of the original teacher's utterances are not available, automatic time alignment and stress detection techniques are employed to automatically generate them, respectively.

When learners are able to correctly perceive sentence stress, the task of CASTLE is to help them produce this sentence with correct stress patterns. In order to help learners to become familiar with the rhythms of English and to train them to get used to placing more emphasis on syllables that are supposed to be stressed, CASTLE 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 utterance. Also, when they practice, they can clap their hands simultaneously with the clapping in the teacher's utterance. The stress detector in CASTLE can automatically obtain the stress patterns of the learner's imitation. By comparing the stress pattern differences between the teacher's utterance and the learner's imitation, CASTLE can then indicate specific words or syllables for learners to work on, in order to help the learners correct their stress errors.

In CASTLE, the individualised speech learning material aims to provide learners with speech material possessing their preferred voices features, which may be easier for the learners to imitate and may also help to maintain the learners' learning interest. The

stress-exaggerated learning material is intended to help learners perceive stress correctly. Through practice using the clapping-based learning material, learners are expected to become more aware of the English stress patterns, and they can form the habit of placing more emphasis on the syllables that are supposed to be stressed. Furthermore, the stress-error feedback provided by CASTLE is intended to help learners correct their stress errors. Thus, CASTLE is an English stress learning system that provides learning assistances from perception to production, in addition to offering individualised speech learning material possessing learners' preferred voices features. With the assistance of CASTLE system, learners would be expected to see an improvement on their stress perception and production ability.

2.3 Summary

In this chapter, we have demonstrated the significance of developing a computer-assisted sentence stress learning system. From an ESL teaching and learning perspective, we have presented the importance and difficulties of learning English stress, which has been neglected in traditional English language teaching classes and current CAPT systems.

We have also proposed a framework for sentence stress teaching systems. Based on this framework, we have designed a computer-assisted sentence stress learning system, CASTLE, which is intended to help ESL learners correctly use sentence stress. In order to develop CASTLE, three research issues have been defined and proposed solutions have also been presented. The flowchart of CASTLE is demonstrated.

Chapter 3.

Speech Processing Techniques for CASTLE

— Automatic Phoneme Alignment and Stress Detection

In this chapter, we present the automatic phoneme alignment and stress detection techniques employed in CASTLE. Section 3.1 briefly reviews previous automatic phoneme alignment techniques. In Section 3.2, we propose a linear-regression-based flexible boundary phoneme alignment algorithm for CASTLE, which is to minimise both the phoneme boundary errors and phoneme duration errors. A review of automatic stress detection is given in Section 3.3. In Section 3.4, we present the stress detection techniques in CASTLE, which only use acoustic features.

3.1 Literature review of automatic phoneme alignment

Segmenting continuous speech into syllables or phonemes is a crucial process in CASTLE system. In order to provide stress-exaggerated speech learning material (refer to Chapter 5), syllable-level time alignments of teachers' speech are necessary. Moreover, time alignment is also important for automatic sentence stress detection since syllable duration and syllable nucleus duration (i.e. the vowel duration of a syllable) are important acoustic cues to indicate stress. The accuracy of automatic time alignment has a strong impact on the performance of CASTLE system.

We take the phoneme-level time alignment as an example to demonstrate the segmentation techniques, which is to segment continuous speech into phonemes, since most literature related to speech automatic alignment is to segment continuous speech on the phoneme-level. Moreover, given the phoneme-level time alignments of continuous speech, the syllable-level time alignments can be achieved directly by grouping phonemes into syllables.

3.1.1 Previous work on automatic phoneme alignment

Automatic phoneme alignment has been studied for decades since a number of speech applications (e.g. speech recognition, concatenative speech synthesis) may benefit from the development of automatic phoneme alignment techniques (Hosom, 2009). Depending on the availability of phoneme transcriptions, alignment techniques can be categorised into two groups (Kamakshi Prasad, et al., 2004; Mporas, et al., 2010; Toledano, et al., 2003): implicit and explicit alignments. When phoneme transcriptions are unknown, implicit alignment algorithms (Gholampour and Nayebi, 1998; Kabré, et al., 1991; Salvi, 2006) can be used to detect the phoneme boundaries. When phoneme transcriptions are known, explicit alignment algorithms (Hosom, 2009; Kuo and Wang, 2006; Lo and Wang, 2007; Mporas, et al., 2010) can be used to align speech according to its phoneme transcription.

Implicit alignment

Without using any explicit phoneme knowledge of speech signals, implicit alignment algorithms consider the points that have significant changes of acoustic properties as potential phoneme boundaries such as points with significant changes of amplitude, energies in different frequency bands (Kabré, et al., 1991), short time energy contours (Gholampour and Nayebi, 1998) and class entropy (Salvi, 2006). Compared with explicit phoneme alignment algorithms, implicit alignment algorithms add a degree of freedom to phoneme segmentation because no explicit phoneme knowledge of speech signals is known. Thus, normally the segmentation accuracy of implicit alignment algorithms is lower than explicit alignment algorithms.

Explicit alignment

For explicit alignment techniques, there are speech synthesis based methods and speech recognition based methods. When a speech synthesiser and a phoneme transcription are available, phoneme alignment can be performed by speech synthesis based algorithms. A typical procedure of a speech synthesis based method can be described as follows (Malfrère, et al., 2003). It starts by synthesizing a reference speech according to the

desired phoneme transcription. In the reference speech, the exact locations of phoneme boundaries are known. Then, the acoustic feature vectors are computed for every few milliseconds for both the synthesised reference speech and the original natural speech to be segmented. Finally, according to the acoustic feature vectors, the original natural speech is time aligned with the synthesised reference speech by the Dynamic Time Wrapping (DTW) algorithm (Rabiner and Juang, 1993).

The advantage of speech synthesis based algorithms is that no training stage is needed. Hence, no large dataset with phoneme time alignments is needed. As a result, the speech synthesis based algorithms can be easily adapted to align different languages given a speech synthesiser (Malfrère, et al., 2003). One of the drawbacks of the speech synthesis based algorithms is that the results depend on the similarity between a synthesised reference voice and an original natural speech to be segmented. The same synthesised reference speech signal is used to align all natural speech of this sentence, no matter who is the speaker (Malfrère, et al., 2003).

When a speech recogniser and a phoneme transcription are available, phoneme alignment can be performed by speech recognition based algorithms which are stemmed from the Viterbi-based forced alignment algorithm (Rabiner and Juang, 1993). The forced alignment, which is based on Hidden Markov Models (HMMs), can be considered as simplified Automatic Speech Recognition (ASR). For ASR, given an utterance, we are interested in the word sequence that has the maximum likelihood. To decide the word sequence, segmenting speech in the time domain is necessary. However, when the phoneme transcription of the utterance is known, the only thing of interest is the start-end points of each phoneme. The process of building a conventional Viterbi-based forced alignment phoneme aligner is illustrated in Figure 3.1. A detailed description of the Viterbi-based forced alignment can be found in Rabiner and Juang (1993).

The training stage is illustrated above the horizontal dashed line in Figure 3.1, which is to build a phoneme aligner. There are two steps to build a phoneme aligner: data preparation and HMMs parameters estimation.

- *Data preparation*

The first step of building an automatic phoneme aligner is data preparation. Speech waveforms are coded into sequences of feature vectors. Word transcriptions of speech waveforms are transferred into phoneme transcriptions by looking up a word-to-phoneme dictionary.

- *HMMs parameters estimation*

The second step of building an automatic phoneme aligner is to estimate the parameters of HMMs. There are two different training strategies to achieve the initialisation of a set of phoneme HMMs, depending whether the hand-labeled segmentations of training speech data (*bootstrap data*) are available (Young, et al., 2006). When bootstrap data are available, *isolated-unit training* can be applied to initialise and re-estimate the HMM parameters of each phoneme by Viterbi and Baum-Welch algorithms. When bootstrap data are unavailable, *embedded training* can be performed, in which all HMMs are initialized with the same global mean and variance, and re-estimated simultaneously by Baum-Welch algorithm. For a detailed description of HMMs parameters estimation, refer to Young, et al. (2006, pp. 118).

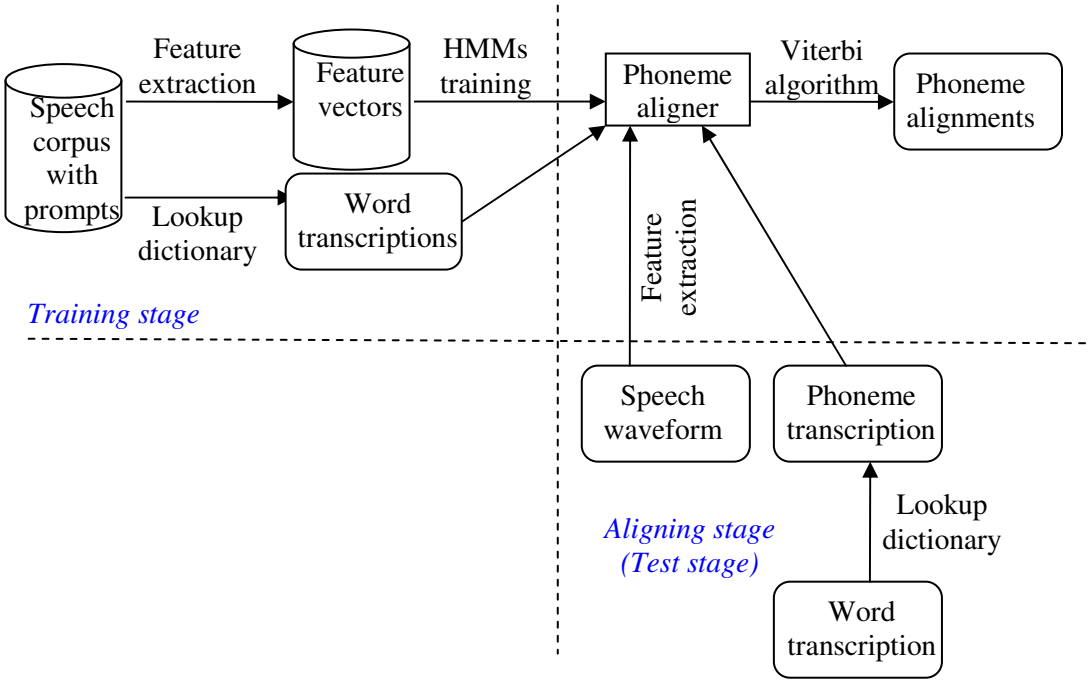


Figure 3.1 Viterbi-based forced alignment

Aligning (test) stage is illustrated on the right hand side of the vertical dashed line. Given a speech waveform and its word transcription, the phoneme aligner is to find the best phoneme-level time alignment by the Viterbi algorithm.

The main advantage of using the forced alignment for phoneme segmentation is that it is developed from the ASR field and has been well studied. However, the goals of phoneme segmentation and ASR are different. In ASR, HMMs are built to identify phonemes, not to detect the phoneme boundaries. Thus, they may have a certain amount of knowledge about phonemes, but the information that they provide about phonemes transition is poor (Lo and Wang, 2007). Therefore, a number of studies have been conducted to improve the segmentation accuracy of the conventional forced alignment algorithm.

Instead of using the maximum likelihood criterion that was used in the conventional forced alignment, the minimum boundary error criterion was used by Kuo and Wang (2006), which aimed to minimise the boundary errors of phoneme alignments represented as a phonetic lattice. Hosom (2009) proposed several modifications to the conventional forced alignment, which included using additional energy-based features, transition-dependent states and probabilities of distinctive phonetic features instead of phoneme-level probabilities. Various refining techniques were used to adjust the raw segmentations obtained by a phoneme alignment algorithm, such as support vector machine (Lo and Wang, 2007), multilayer perceptron (Lee, 2006), statistical correction to compensate for the systematic error (Toledano, et al., 2003), and the context-dependent boundary model (Wang, et al., 2004). Given a number of phoneme segmentations produced by different algorithms, linear and non-linear regression methods were used to get the final phoneme boundaries (Jarifi, et al., 2008; Kominek and Black, 2004; Mporas, et al., 2010; Park and Kim, 2007).

3.1.2 Performance comparison

Numerous automatic phoneme alignment systems have been reported to improve the segmentation accuracy of the conventional forced alignment. However, direct comparing the performances of these systems is not possible (Hosom, 2009), since

different performance criteria and corpora have been applied in the performance evaluations reported by different researchers.

The most frequently used criterion to evaluate the performance of automatic phoneme alignment algorithms is the percentage of agreement between the automatic segmentation and its manual counterpart within a given time tolerance (Hosom, 2009). The time tolerance is usually 20ms, although some systems also measure their accuracy in a tolerance of 10ms or 15ms (Hosom, 2009; Kuo and Wang, 2006; Lo and Wang, 2007; Mporas, et al., 2010). Some systems are evaluated by other criteria such as mean boundary distance and root mean square error (Lee, 2006).

Some researchers evaluate their phoneme alignment systems on speech datasets collected by themselves, which may not be publicly available (Jarifi, et al., 2008; Xie, et al., 2004b). Even for the systems that are evaluated on the same corpus and same evaluation criteria, implementation differences may also prevent a direct comparison. For example, the TIMIT corpus (Garofolo, et al., 1990), an acoustic-phonetic corpus of English speech, is a benchmark dataset for automatic phoneme segmentation. A detailed description of the TIMIT corpus can be found in Section 3.2.2. There are 61 phonemes in the original TIMIT corpus. Different segmentation systems may merge the phonemes of the TIMIT corpus into different phoneme sets. The 61 phonemes in the TIMIT corpus were merged into 50 phonemes in Hosom (2009), Kuo and Wang (2006), Lo and Wang (2007) and Mporas, et al. (2010), while these 61 phonemes were merged into 48 phonemes in Mporas, et al. (2010), 46 phonemes in Pellom and Hansen (1998), and 35 phonemes in Wightman and Talkin (1997). Moreover, some systems are evaluated on the test set of the TIMIT corpus excluding dialect sentences (Hosom, 2009; Kuo and Wang, 2006; Lo and Wang, 2007; Mporas, et al., 2010), while some systems also discard utterances with phones shorter than 10 ms (Hosom, 2009; Kuo and Wang, 2006; Lo and Wang, 2007; Mporas, et al., 2010).

For a rough comparison, Table 3.1 lists the performances of forced alignment systems on the TIMIT corpus. The evaluation criterion used in Table 3.1 is the percentage of agreement between automatic segmentation and its manual counterpart within 20ms. As indicated by (Hosom, 2009), there is no accurate answers to the phoneme alignment problem since the alignments between phonemes are inherently subjective. The inter-

labeler agreement on the phoneme locations of the TIMIT corpus is reported as 93.49% in Hosom (2009).

Table 3.1 Accuracies of different phoneme aligners on the TIMIT corpus.

System	Accuracy (<20ms)
Pellom and Hansen (1998)	85.90%
Hosom (2002)	92.57%
Kuo and Wang (2006)	92.11%
Lo and Wang (2007)	92.47%
Malfrère et al. (2003)	78.18%
Hosom (2009)	93.36%
Mporas et al. (2010)	88.18%
Human agreement (Hosom, 2009)	93.49%

Considering the high agreement between forced-alignment-based automatic segmentation and human segmentation, the automatic phoneme aligner in our CASTLE system is developed based on forced alignment.

3.2 Automatic phoneme alignment in CASTLE

In this section, we analysis previous phoneme alignment algorithms and indicate the deficiency of applying those algorithms directly to CASTLE. In order to overcome the deficiency, we propose a Linear-Regression-based Flexible Boundary phoneme alignment algorithm (LR-FB phoneme aligner) that can minimise the phoneme boundary errors and phoneme duration errors. We also describe the development of the LR-FB phoneme aligner of CASTLE in detail.

3.2.1 Deficiency of previous phoneme alignment algorithms

The phoneme aligner in CASTLE needs to estimate the phoneme boundaries as accurately as possible. At the same time, the estimated phoneme durations also needs to be as accurate as possible. This is because syllable nucleus duration is a very important feature for stress detection. The importance of syllable nucleus duration for stress

detection is demonstrated in Section 3.4.3, which is also confirmed by Ananthakrishnan and Narayanan (2008) and Tamburini and Caini (2005).

The previous phoneme alignment algorithms are suitable to obtain phoneme boundaries, whereas it is problematic to obtain phoneme durations by these previous alignment algorithms. The previous alignment algorithms try to minimise the differences between the estimated phoneme boundaries and their reference counterparts without considering the phoneme duration differences between the estimated and the reference ones. However, acceptable discrepancies between the estimated and reference phoneme boundary time-marks do not always lead to acceptable differences between the estimated and reference phoneme durations.

This problem is illustrated by two examples shown in Figure 3.2. The horizontal line indicates the time axis of an utterance. The two solid dots are the reference boundary time-marks of a syllable nucleus, between which is reference segment t_r. The segments in brackets and between the two triangles are the estimated speech segments, t_{S1} and t_{S2}, respectively.

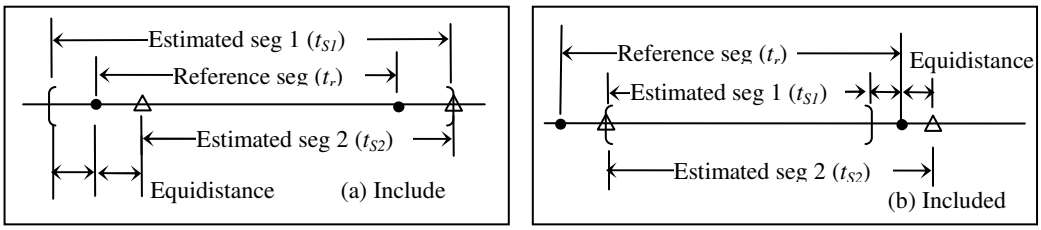


Figure 3.2 Comparison between estimated duration and its reference counterpart

In the example shown in Figure 3.2 (a), the left bracket and left triangle are equidistance from the left dot, and the right bracket and right triangle are overlapped. From this example, we can see that the boundary time-marks of brackets and triangles have the same distances with their reference counterparts. However, reference duration t_r is closer to t_{S2} (duration between the two triangles) than to t_{S1} (duration in the two brackets), since t_{S1} contains the length of its reference counterpart t_r (in other words t_{S1} includes t_r). This makes t_{S1} always greater than t_r.

Similarly, in Figure 3.2 (b), t_{S1} lies inside its reference counterpart t_r (in other words t_{S1} is included in t_r). This makes t_{S1} always less than t_r. t_{S2} is overlapped with t_r. This makes the difference between t_{S2} and t_r smaller than the difference between t_{S1} and t_r.

A good duration estimation model should avoid the situations that the estimated durations include (or are included in) their reference counterparts, otherwise the estimated durations would always be greater than (or less than) their reference counterparts. The four possible types of relationship between estimated and reference durations are illustrated in Figure 3.3, where the dots are the reference boundaries and the brackets are the estimated boundaries. In Figure 3.3 (a) and (b), the estimated duration includes and is included in their reference counterparts, respectively. There is 50% chance that the situations in Figure 3.3 (a) and (b) occur if the automatic speech segmentation machine only tries to minimise the difference between estimated and reference phoneme boundary time-marks.

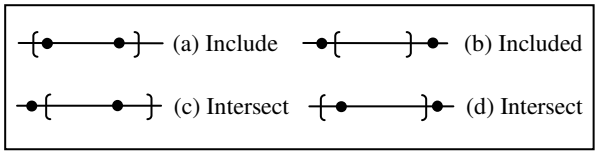


Figure 3.3 Relationships between estimated and reference syllable durations

3.2.2 Linear-regression-based flexible boundary phoneme alignment

In order to reduce the chance that an estimated duration includes (or is included in) its reference counterpart in automatic speech segmentation, we propose an ensemble phoneme alignment algorithm that minimises the discrepancy of both the estimated phoneme boundaries and durations with their reference counterparts. It is a Linear-Regression-based Flexible Boundary phoneme aligner (LR-FB phoneme aligner) which combines the outcomes of multiple base phoneme aligners. LR-FB phoneme aligner minimises the average distance between the predicted and reference boundary-pairs of phonemes.

In conventional speech segmentation systems, the starting point of a phoneme is identical to the ending point of its previous phoneme. Then, the duration of a phoneme is the segment between its starting point and that of its subsequent one's. This duration

dependency expands to a whole utterance, which makes solving the boundaries and durations optimization problem computationally expensive. To solve this problem, in our LR-FB phoneme aligner, the relationship of two conjunctive phoneme segments is flexible, which can be overlapped, connected or unconnected, as shown in Figure 3.4.

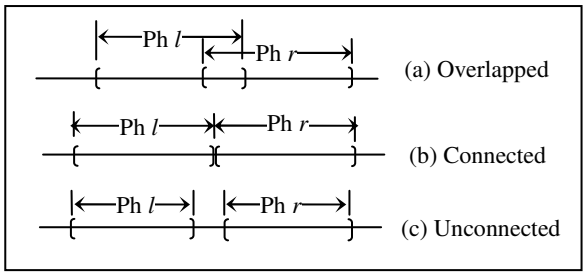


Figure 3.4 Possible boundary relationships of two conjunctive phonemes

The flexible phoneme boundary relationships have the ability to better model the relations between the durations of two conjunctive phonemes than the simply connected phoneme boundary relationship. It has been noticed that it is not easy to segment speech into small units consistently (Hosom, 2009), even for experienced labelers. For example, they have difficulties in marking the boundaries of vowel-to-vowel (Park and Kim, 2007). Moreover, the segmentations of different labelers may be inconsistent, and the segmentations of the same labeler in different times may also be inconsistent. This is mainly because some phoneme transitions are ambiguous. For some speech segment around a boundary, it is difficult to tell which phoneme it belongs to. It may sound like both of the two phonemes or none of them. The connected boundary relationship alone cannot model the ambiguity in phoneme transitions. Thus, in our LR-FB phoneme aligner, it is reasonable to assume that the boundary relation of two conjunctive phonemes can be overlapped, connected or unconnected.

LR-FB phoneme aligner is a weighted sum of segmentation results of different base phoneme aligners, which minimises the average distance between the estimated and reference speech segments. To reduce the possibility that an estimated duration includes (or is included in) its reference counterpart, discrepancies of phoneme boundaries and phoneme durations are taken into consideration in LR-FB phoneme aligner. Since the sum of discrepancies of phoneme durations is a criterion to be minimised in LR-FB phoneme aligner, the phonemes boundaries have to be considered in pairs (i.e. their starting and ending points).

Given utterance u with its phoneme transcription and the outputs of K base phoneme aligners, for phoneme p in u, whose previous and subsequent phonemes are l and r, its starting and ending points generated by the K base phoneme aligners are $t_j(l)$ and $t_j(p)$, where j=1,…, K. $t_j(l)$ is also the ending points of phoneme l and $t_j(p)$is also the starting point of phoneme r in the K base phoneme aligners, since in the K base phoneme aligners, the ending point and starting point of two conjunctive phonemes are identical. The time-mark interval of p achieved by LR-FB phoneme aligner is as follows,

$$[\hat{s}(p), \hat{e}(p)] = \sum_{j=1}^{K} w_{j,l-p+r}[t_j(l), t_j(p)] + [s, e]_{l-p+r} \tag{3.1}$$

where $w_{j,l-p+r}$ is the weight of the j^{th} base phoneme aligner; given the triphone type l-p+r, $[s, e]_{l-p+r}$ is a system error. Figure 3.5 gives an overview of our LR-FB phoneme aligner.

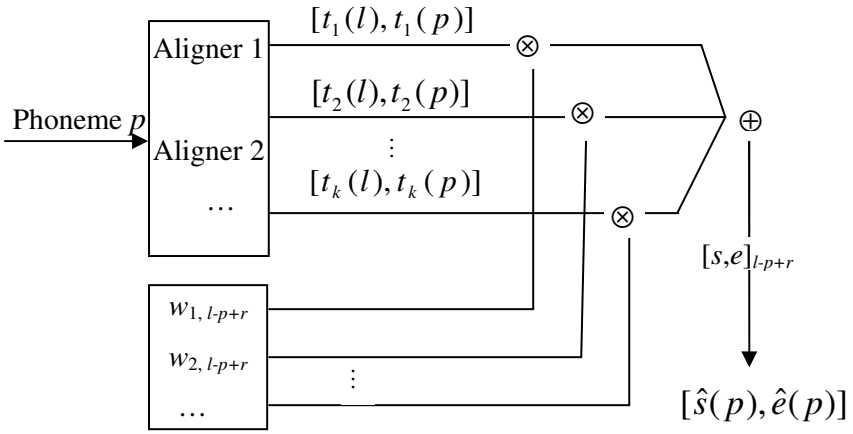


Figure 3.5 Overview of the LR-FB phoneme aligner

To estimate the weights of the base phoneme aligners and system error for each triphone model, we define the *distance* of two segments $[s_i, e_i]$ and $[s_j, e_j]$ by Eq.(3.2), which takes both the boundary difference and duration difference into consideration:

$$\begin{aligned} &Dis([s_i, e_i], [s_j, e_j]) \\ &= ((e_i - s_i) - (e_j - s_j))^2 + (s_i - s_j)^2 + (e_i - e_j)^2 \end{aligned} \tag{3.2}$$

The criterion to estimate the weights of the base phoneme aligners and system error for each triphone model is to minimise the total *distance* between the predicted and

reference boundary-pairs of the corresponding triphone instances. For a triphone model *l-p+r*, the objective function is:

$$\text{Min: } \sum_{p \in A} Dis([\hat{s}(p), \hat{e}(p)], [t_r(l), t_r(p)]) \tag{3.3}$$

$$\hat{s}(p) = \sum_{j=1}^{K} w_{j,l-p+r} t_j(l) + s_{l-p+r},$$

$$\hat{e}(p) = \sum_{j=1}^{K} w_{j,l-p+r} t_j(p) + e_{l-p+r}$$

$$\text{s.t.} \quad \sum_{j=1}^{K} w_{j,l-p+r} = 1$$

where $t_r(l)$ and $t_r(p)$ are the reference ending points of phonemes l and p. Note that for the optimization problem given in Eq.(3.3), the feasible region of solutions may not be continuous, and therefore there is no guarantee to find the global minimum. The solution searching might be ended at a local minimum.

3.2.2 TIMIT speech corpus

The automatic phoneme aligner of CASTLE is developed based on the TIMIT corpus because the TIMIT speech corpus (Garofolo, et al., 1990) is a benchmark dataset for automatic phoneme alignment, which is an acoustic-phonetic corpus of English speech. The TIMIT corpus consists of 6300 utterances, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. Each speech has a time-aligned word transcription and a time-aligned phoneme transcription. These time-aligned word transcriptions and phoneme transcriptions are gained from manually checked automatic segmentation.

In order to train the HMMs for each phoneme, the speech of the TIMIT corpus is coded into feature vectors. The parameters used in coding of feature vectors are summarised in Table 3.2.

Table 3.2 Parameters used to train the LR-FB phoneme aligner in CASTLE

Parameter Name	Value
Sampling frequency	16,000Hz
Window size, type	10ms, Hamming
Frame size	5ms
Pre-emphasis factor	0.97
Feature vector	12 MFCC with energy, and their corresponding delta and acceleration coefficients
Phoneme set	52 phonemes
Filterbank channels	26

3.2.3 Experiments

Our experiments are to investigate the effectiveness of the LR-FB phoneme aligner. We conducted our experiments on the TIMIT corpus. The TIMIT handbook suggests the training set and the test set. In our experiments, we used the training set to train HMMs and the core test set to test different base phoneme aligners and our LR-FB phoneme aligner. According to the suggestions in the TIMIT handbook, we discarded SA1 and SA2 sentences in both training set and test set because they were uttered by both training and test speakers. Then, in our experiments, the training set and test set contained 3696 and 192 sentences, respectively.

The base phoneme aligners were trained by using the HTK toolkit (Young, et al., 2006). There were 33 base phoneme aligners that were 11 different numbers of mixtures (from 1, 2, 4, … , up to 20, step size is 2) for 3-state, 5-state and 7-state HMMs, respectively.

Evaluation criteria

Seven criteria were used in the evaluation, which are listed in Table 3.3.

Mean boundary error of phonemes (*MBdyErr_ph*) is defined as the average difference between the estimated and reference phoneme boundaries, as shown in Eq. (3.4).

$$MBdyErr\_ph = \frac{1}{N_P}\sum_{i\in P}(|s_{r,i} - s_{e,i}| + |e_{r,i} - e_{e,i}|)/2 \quad (3.4)$$

where P is the phoneme set; N_P is the cardinality of P; and $s_{r,i}$, $e_{r,i}$ ($s_{e,i}$ and $e_{e,i}$) are the reference (estimated) starting and ending points of the i^{th} phone.

Mean boundary error of syllable nucleus (*MBdyErr_v*) is defined in a similar manner as the mean boundary error of phonemes, which was the average difference between the estimated and reference syllable nucleus boundaries.

Mean duration error of phonemes (*MDurErr_ph*) is defined as the average difference between the estimated and reference phoneme durations.

$$MDurErr\_ph = \frac{1}{N_P}\sum_{i\in P}|(e_{r,i} - s_{r,i}) - (e_{e,i} - s_{e,i})| \quad (3.5)$$

Similarly, mean duration error of syllable nuclei (*MDurErr_v*) is the average difference between the estimated and reference syllable nucleus durations.

Segmentation accuracy of boundaries (*Acc_bdy*), segmentation accuracy of phonemes (*Acc_ph*) and segmentation accuracy of syllable nuclei (*Acc_v*) are defined as the percentage of correctly segmented boundaries, phonemes and syllable nuclei. A phoneme or syllable nucleus is correctly segmented if both its starting and ending points have a deviation within a tolerance with respect to its reference counterpart.

Table 3.3 Performances of base phoneme aligners and the LR-FB phoneme aligner

Measuring Quantity	Abbreviation	Best single base phoneme aligner	LR-FB aligner	Improvement
Mean boundary error of phonemes	*MBdyErr_ph*	9.28ms (5-st. 16-mix.)	7.58ms	1.23ms
Mean boundary error of syllable nuclei	*MBdyErr_v*	8.81ms (5-st. 4-mix.)	8.09ms	1.19ms
Mean duration error of phonemes	*MDurErr_ph*	12.58ms (5-st. 16-mix.)	10.70ms	1.88ms
Mean duration error of syllable nuclei	*MDurErr_v*	13.64ms (7-st. 6-mix.)	11.81ms	1.83ms
Segmentation accuracy of boundaries	*Acc_bdy* (<20ms)	90.56% (5-st. 16-mix.)	92.70%	2.14%
Segmentation accuracy of phonemes	*Acc_ph* (<20ms)	83.35% (5-st. 16-mix.)	86.79%	3.44%
Segmentation accuracy of syllable nuclei	*Acc_v* (<20ms)	81.59% (7-st. 4-mix.)	83.98%	2.39%

Experimental results

Since the training data is limited and some triphones may meet sparse data problems, to achieve a robust estimation of the weights in the LR-FB phoneme aligner, we clustered triphones into groups. The clustering method we used was the same as the method of creating tied-state triphones (Young, et al., 2006).

Table 3.3 shows the performances of base phoneme aligners and the LR-FB phoneme aligner for each evaluation criterion within a tolerance of 20ms. The LR-FB aligner increases phoneme segmentation accuracy *Acc_ph* to 86.79% from 83.35%. Syllable nucleus segmentation accuracy *Acc_v* is increased to 83.98% by the LR-FB phoneme aligner from 81.59%. In the LR-FB aligner, the *mean boundary errors* of phonemes and syllable nuclei decrease to 7.58ms and 8.09ms, respectively; also, the *mean duration errors* of phonemes and syllable nuclei decrease to 10.70ms and 11.81ms.

LR-FB phoneme aligner minimises both phoneme boundary differences and phoneme duration differences between the predicted and reference time alignments of phonemes. Thus our LR-FB phoneme aligner is more suitable for the application of CASTLE. The effectiveness of LR-FB phoneme aligner has also been demonstrated in our experiments.

3.3 Literature review of automatic stress detection

In this section, we briefly review the literature related to automatic stress detection. Some systems that only rely on syntactic and lexical features to predict stress labels are not included here, since CASTLE is developed to help ESL learners to perceive and produce sentence stress physically. A syllable or word that is supposed to be stressed from a syntactic or lexical perspective is not always stressed in learners' speech.

3.3.1 Previous work on automatic stress detection

Since Wightman and Ostendorf (1994) first presented their work in automatic prosodic events detection, detection of stress, as well as other prosodic features such as boundary

tones and break indices has attracted researchers' attention for over a decade. The stress detector presented by Wightman and Ostendorf (1994), which combines a decision tree and a Markov model, maps a sequence of acoustic features to a sequence of stress labels. The detection accuracy reaches 84% at the syllable level on a subset of the Boston University Radio News corpus (BU-RN). For a description of the BU-RN corpus, refer to Section 3.4.1.

Imoto et al. (2000) identified sentence stress by acoustic features (i.e. vowel duration, pitch and intensity), which were integrated into a linear discriminant function. The agreement between automatic labeled and hand-labeled sentence stress patterns was 90% on a subset of the TIMIT dataset (New England dialect). In Imoto et al. (2002), a two-stage sentence stress classifier based on hidden Markov model (HMM) was used to detect the degree of syllable stress. In the first stage, the classifier classified syllables into stressed and unstressed. In the second stage, the stressed syllables were classified into primary stressed and secondary stressed. Two stress detectors were presented in Tamburini and Caini (2005). One was a supervised system that was based on multivariate Gaussian discriminators, and the other one was an unsupervised system that was based on a continuous prominence function. In Li, et al. (2007), acoustic features, Mel-scale frequency cepstral coefficients (MFCC), energy and pitch were fed into to a HMM-based detector to predict the best stress labels. The rate of correctly detected stressed syllables was 80.6% on the BU-RN corpus.

More recent studies on stress detection were reported by Ananthakrishnan and Narayanan (2008), Rangarajan Sridhar et al. (2008), and Deshmukh and Verma (2009). The stress detection model presented in Ananthakrishnan and Narayanan (2008) built on HMMs, which employed a maximum *a-posterior* framework, instead of a maximum likelihood framework. Also, an *n*-gram prosodic language model was integrated into the detector to predict the most likely stress label sequences. Acoustic features, as well as syntactic and lexical features, were exploited in this model. An accuracy of 86.75% was achieved at the syllable level on the BU-RN corpus. A discriminative classification framework based on maximum entropy was proposed by Rangarajan Sridhar et al. (2008). Acoustic, lexical and syntactic features were used to predict the best prosodic label sequence. The proposed model obtained an accuracy of 86% at the word level on the BU-RN corpus. An acoustic knowledge-based syllable grouping technique for

classifying stress was proposed by Deshmukh, and Verma (2009). Three grouping techniques were presented, which were based on the energy of syllable nuclei, both energy and duration of syllable nuclei, and syllable identity. The best performances of the proposed group-specific classifiers was achieved by grouping syllables according to their identities, which were 93.1% at the syllable level and 85% at the word level on a word corpus.

Syntactic and lexical information cannot be used in CASTLE system to detect the stress labels of learners' utterances because ESL learners do not always follow pronunciation rules. However, from the above review, we can see that in order to achieve higher detection accuracy, most stress detection systems exploited acoustic features, as well as syntactic and lexical information. For example, canonical stress information, which can be obtained from an electronic pronunciation dictionary, was used in Ananthakrishnan and Narayanan (2008) to build probabilistic models of prosodic event sequences, since syllables associated with lexical stress were more likely to be stressed in utterances. In order to improve classification accuracy, the phonological rule, which is that one and only one syllable can be primarily stressed in a word, was employed by Deshmukh and Verma (2009) to post-process the results of automatic stress classification. Part-of-speech (POS) tags and supertags were used in the models proposed in Ananthakrishnan and Narayanan (2008) and Rangarajan Sridhar, et al. (2008), respectively, which represented the syntactic information of a sentence.

3.3.2 Performance comparison

Direct performance comparison of different stress detection approaches is almost impossible. Different approaches may be evaluated on different corpora, which may not be publicly available. Even if some stress detection approaches perform on the same corpus, they may use different subsets of the corpus as a training set and a test set. Also, different approaches may detect stress on different levels, either on a syllable level or on a word level. Moreover, whether the manually labeled word or syllable boundaries are available, and whether the lexical and syntactic information is being used, all these implementation details have a significant impact on the performance of different stress detection approaches. For a rough comparison, the reported results of recent work on

automatic stress detection are listed in Table 3.4. In the last column of Table 3.4, *A* means the detection results are obtained only by exploiting acoustic features; *L* means the detection results are obtained only by using lexical features; *S* means the detection results are obtained only by using syntactic features; and *A+S* means the detection results are obtained by using both acoustic and syntactic features. Rangarajan Sridhar et al. (2008) also gave a comparison of previous work on automatic stress detection. On the BU-RN corpus, the inter-labeler agreement on the presence versus absence of stress is reported as 91% (Ostendorf, et al. 1995).

Table 3.4 Performance comparison of previous stress detectors.

Authors	Algorithm	Features	Corpus	Level	Accuracy (%)
Hasegawa-Johnson et al. (2005)	neutral network	acoustic, syntactic	A subset of the BU-RN corpus	word	76.58 (A) 82.67 (S) 83.91 (A+S)
Li et al. (2007)	HMM	acoustic	BU-RN	syllable	80.6
Ananthakrishnan and Narayanan (2008)	HMM	acoustic, lexical, syntactic	BU-RN	syllable	86.75 (A+L+S)
Rangarajan Sridhar et al. (2008)	maximum entropy discriminative model	acoustic, lexical, syntactic	BU-RN	word	80.09 (A) 84.60 (L+S) 85.13 (A+L+S)
			Boston Directions	word	74.51 (A) 79.81 (L+S) 80.01(A+L+S)
Deshmukh, and Verma (2009)	acoustic knowledge-based syllable grouping	acoustic, lexical	words produced by Indian	syllable	93.1(A+L)
				word	85(A+L)

3.4 Automatic stress detection in CASTLE

In this section, we describe the development of automatic stress detector(s) of CASTLE in detail. In order to evaluate the effectiveness of the automatic stress detector(s) of CASTLE, experiments are conducted on the Boston University Radio News Corpus.

3.4.1 Boston University Radio News speech corpus

Boston University Radio News corpus (BU-RN) (Ostendorf, et al., 1995) consists of continuous speech produced by seven FM radio news announcers (i.e. F1A, F2B, F3A, M1B, M2B, M3B and M4B) associated with WBUR, a public radio station.

There are two types of speech in the BU-RN corpus: *radio news* and *lab news*. The *radio news* type of speech consists of news stories recorded in the WBUR radio studio during actual news broadcasts. The *lab news* type of speech consists of six announcers' recording which reads the same four news stories in a laboratory environment.

Most speech in the BU-RN corpus is annotated with the word transcription, phoneme-level time alignments and prosodic markers (e.g. stress label). The word transcripts are generated by hand. The phoneme-level time alignments are generated automatically, and some of them are hand corrected. The prosodic labels are marked by hand and are available only for a subset of the corpus.

Our experiment data are the speech of the BU-RN corpus which has prosodic labels. The prosodic labels of the BU-RN corpus are in the *Tones and Break Indices* (ToBI) format (Silverman, et al., 1992) that is widely used in intonation symbolic transcription, phonetics and speech synthesis. According to the ToBI transcription system, there are eight types of tonal events associated with stressed syllables, which are listed in Table 3.5. In Table 3.5, multiplication symbol (*) indicates that the tone is associated with a stressed syllables, and plus symbol (+) indicates that the two tones are associated. Other types of ToBI labels are considered as being associated with unstressed syllables. For a more detailed description of the ToBI transcription, refer to MIT courseware (2006).

Table 3.5 ToBI labels associated with stressed syllables.

ToBI labels	Description
H*	High pitch accent
L*	Low pitch accent
L+H*	Bitonal low tone with high tone on stressed syllable
L*+H	Bitonal high tone with low tone on stressed syllable
!H*	Downsteped high pitch accent
L+!H*	Bitonal pitch accent with low tone followed by a downstepped high tone prominence
L*+!H	Bitonal pitch accent with low tone prominence followed by downstepped high tone
H+!H*	Bitonal pitch accent with high tone followed by downstepped high prominence

3.4.2 Feature extraction

Considering that CASTLE aims to assist ESL learners to perceive stress acoustically not to predict stress from word transcription, the stress detection models in CASTLE only employs acoustic features (i.e. without syntactic and lexical information). Pitch, duration and intensity are the acoustic cues to indicate English stress. Stressed syllables tend to be pronounced longer, louder, and with significant pitch movement. The input feature vectors of the stress detector(s) in CASTLE are listed in Table 3.6.

Table 3.6 Input features of the stress detector(s) in CASTLE

	Acoustic features	Abbreviation
Pitch-based (27 features)	Average of F0	F0_avg
	Maximum of F0	F0_max
	Minimum of F0	F0_min
	Range of F0	F0_range
	Difference between F0_avg and F0_max	F0_maxavg_diff
	Difference between F0_avg and F0_min	F0_avgmin_diff
	Difference between F0_avg and the utterance average F0	F0_avgutt_diff
	Ratio of F0_avg of current syllable to F0_avg of previous syllable	curPreF0Ratio
	Ratio of F0_avg of current syllable to F0_avg of next syllable	curNxtF0Ratio
Duration-based (6 features)	Normalized syllable duration	dur
	Normalized syllable nucleus duration	durNucleus
	Ratio of dur of current syllable to dur of previous syllable	curPreDurRatio
	Ratio of dur of current syllable to dur of next syllable	curNxtDurRatio
	Ratio of dur_v of current syllable to dur_v of previous syllable	curPreDurVRatio
	Ratio of dur_v of current syllable to dur_v of next syllable	curNxtDurVRatio
Intensity-based (9 features)	Average of intensity	in_avg
	Maximum of intensity	in_max
	Minimum of intensity	in_min
	Range of intensity	in_range
	Difference between in_avg and in_max	in_maxavg_diff
	Difference between in_avg and in_min	in_avgmin_diff
	Difference between in_avg and the utterance average intensity	in_avgutt_diff
	Ratio of in_avg of current syllable to in_avg of previous syllable	curPreInRatio
	Ratio of in_avg of current syllable to in_avg of next syllable	curNxtInRatio
Vowel	Identity of the syllable nucleus	sylNucleusID

Following Jande (2008), in our experiments, pitch-based features are measured in the Hertz scale, as well as the semitone scale and the Equivalent Rectangular Bandwidth (ERB) scale. As indicated by Jande (2008), the relationship between human auditory perception and pitch measured in Hertz is not linear. Equal intervals of pitch measured in Hertz are judged by listeners to produce different pitch increments. Thus, psycho-acoustic scales, such as semitone, ERB and Mel, are constructed to measure the response of the human auditory perception system. According to the study carried out by Nolan (2003), the semitone scale is logarithmic of the Hertz scale; the ERB scale is between linear and logarithmic of the Hertz scale below 500 Hz, and logarithmic above 500Hz; the Mel scale is a linear transformation of the Hertz scale below 500Hz, and a logarithmic transformation above 500Hz. In our stress detection model, pitch-based features are not measured in Mel scale, since below about 500Hz, the Mel of the Hertz scale and Hertz scale are approximately linear, and the pitch of human speech is normally below 500Hz. The conversion formula from Hertz to semitone used in our research is given in Eq. (3.6),

$$F0_{Semitone} = 12 * \log^{*}(F0_{Hz} / F0_{HzR}) \tag{3.6}$$

where $F0_{HzR}$ is the fundamental frequency of a reference sound. Also, the conversion formula from Hertz to ERB used in our research is given in Eq. (3.7).

$$F0_{ERB} = 11.17 * \ln((F0_{Hz} + 312)/(F0_{Hz} + 14680)) + 43 \tag{3.7}$$

The intensity-based features are measured in a high frequency band that is above 500Hz, since studies (Sluijter, et al., 1997; Tamburini and Caini, 2005) found that the intensity of stressed and unstressed syllables had a remarkable difference in the high frequency band (i.e. above 500Hz), and the intensity difference between stressed and unstressed syllables in the low frequency band (i.e. below 500Hz)was negligible.

Identity of the syllable nucleus is also included in the input feature vectors. If the syllable time alignment of an utterance is available, deriving the identities of the syllable nuclei is straight forward. If the syllable time alignment of an utterance is unavailable, an automatic phoneme segmentation processing can be employed to generate the syllable-level time alignment.

Thus, the input feature vectors of our stress detection models are 43-dimensional vectors that are composed of 27 pitch-based features, 6 duration-based features, 9 intensity-based features, and the identity of the syllable nucleus, as shown in Table 3.6.

3.4.3 Experiments

The experiments are designed to investigate the effectiveness of the stress detectors that are based only on acoustic features. Given the training data, CASTLE can employ any classification technologies to build a stress detector. We used four classifiers in the experiments: logistic regression (LR), multilayer perception (MLP), sequential minimal optimization (SMO) and decision tree (DT). We also tested a meta classifier (i.e. vote) that combines the four single classifiers by using unweighted average of probability estimates. The classification experiments were performed on WEKA (Waikato Environment for Knowledge Analysis, http://www.cs.waikato.ac.nz/ml/weka/), which is an open source toolkit containing many machine learning algorithms for data mining tasks (Witten and Frank, 2005).

10-fold stratified cross-validation was employed to test the automatic stress detectors. In other words, all experimental data were pooled and then partitioned into ten equal-sized folds. Each time, nine folds were used as the training set by a classifier and the remaining fold was used as the test set. This process was repeated until each fold had been used as the test set once. Classification accuracies were averaged over all folds.

The importance of acoustic features for stress detection was ranked by the *information gain* criterion on the WEKA platform. The three most important features are, in a decreasing order: sylNucleusID, durNucleus and curPreF0Ratio. The high correlation between the identity of the syllable nucleus and stress label can be explained by the fact that some syllable nuclei normally appear in unstressed syllables, such as /ən/ (button), /əm/ (bottom), /əl/ (apple) and schwa /ə/ (about). The importance of syllable nucleus duration is consistent with the study results of Ananthakrishnan and Narayanan (2008), and (Tamburini and Caini, 2005).

The performances of the stress detectors are shown in Table 3.7. The best performance is achieved by MLP, 84.488%. The automatic detection accuracy of MLP in our experiments is still lower than the inter-transcriber agreement (91%), and the best reported result 86.75% (Ananthakrishnan and Narayanan, 2008) that employed acoustic, syntactic and lexical features. However, MLP performs better than those stress detectors that only exploit acoustic features, which are listed in Table 3.4. As we indicated in Section 3.3, it is almost impossible to compare the performances of different stress detection approaches directly, due to the implementation differences. The comparison here is just a rough assessment.

Our automatic stress detectors are more suitable to detect stress labels of ESL learners' speech since our stress detectors only use acoustic features. Also, the automatic stress detectors perform better than the previously reported acoustic-feature-based stress detectors that as listed in Table 3.4.

Table 3.7 Performances of different stress detectors

Stress detectors	Accuracy (%)	Precision		Recall	
		stressed	unstressed	stressed	unstressed
LR	84.1382	0.763	0.874	0.717	0.898
MLP	84.488	0.759	0.883	0.742	0.892
SMO	83.8798	0.753	0.876	0.724	0.892
DT	82.6268	0.722	0.874	0.725	0.873
Vote	84.3375	0.763	0.877	0.726	0.897

3.5 Summary

In this chapter, we have reviewed two foundational speech processing techniques for developing CASTLE, i.e. automatic phoneme alignment and stress detection. The automatic phoneme alignment and stress detection techniques will be used in Chapter 5 to achieve the time alignments and stress labels of teachers' utterances, respectively, when the hand-labeled time alignments and stress labels of teachers' utterances are not available. The automatic phoneme alignment and stress detection techniques will also be used in Chapter 6 to obtain the time alignments of ESL learners' speech production and detect learners' stress errors.

Considering that phoneme duration is an important cue for the stress detection, in this chapter, we have proposed a Linear-Regression-based Flexible Boundary (LR-FB) phoneme aligner for CASTLE, which minimises both phoneme boundary differences and phoneme duration differences between the predicted and reference time alignments of phonemes. The effectiveness of our LR-FB phoneme aligner has been demonstrated on the TIMIT corpus. LR-FB phoneme aligner decreases the mean duration errors of phonemes from 12.58ms (the best result of a single base phoneme aligner) to 10.70ms.

It is not reasonable to employ syntactic and lexical features to predict stress labels of ESL learners' speech in CASTLE, since a syllable or word, which is supposed to be stressed from a syntactic or lexical perspective, is not always stressed by ESL learners. Thus, we have developed stress detectors for CASTLE, which are only based on acoustic features. The effectiveness of our stress detectors has been evaluated on the Boston University Radio News corpus. Although the performance of our best stress detector (MLP-based) is still lower than inter-labeler agreement (91%) and the performance of the stress detectors that employ both acoustic features, and syntactic and lexical features, our best stress detector (MLP-based) outperforms previously reported stress detectors that only exploit acoustic features.

Chapter 4.

Individualised Speech Material Module

In order to provide individualised speech learning material that is preferred by language learners to listen to and imitate, we investigate which voice features (i.e. gender, pitch and speech rate) make a teacher's voice preferable for language learners to imitate. Our investigation of the voice features is from learner's imitation preference perspective. Thus, a "golden speaker" in our research refers to a speaker whose voice is preferred by language learners to imitate.

This chapter is organised as follows: In Section 4.1, we review previous work that is intended to find what voices are suitable for language learners to imitate. Section 4.2 presents the differences between our research and previous research. In Section 4.3, we introduce the prosody modification techniques that are employed in our study to resynthesise sample voices with different voice features. Section 4.4 describes the setup of the experiments that we conducted to explore language learners' imitation preferences. Experimental results and discussions are provided in Section 4.5. Section 4.6 concludes our investigation of learners' preferred voice to be imitated. We finally summarise our individualised speech material providing module in Section 4.7.

4.1 Previous research on voices suitable for learners to imitate

Imitation is the most commonly used method to improve pronunciation, and also considered as one of the most effective methods (Ding, 2007). Then the question is whether different voices uttering the same learning material make a difference to pronunciation learning. If so, what voices are suitable for language learners to imitate? Some research has been carried out to try to find that what voices are suitable for language learners to imitate.

4.1.1 The learner's 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 (Bissiri and Pfitzinger, 2009; Sundström, 1998). Also, speech synthesis technologies have been developed to synthesise native-like utterances with learners' voice characteristics (Bissiri and Pfitzinger, 2009; Felps, et al., 2009; Hirose, 2004; Nagano and Ozawa, 1990; Sundström, 1998).

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 modification 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 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 resynthesised utterances by mapping the prosody of a teacher's voice onto a learner's voice are not suitable for learners to imitate.

Felps et al. (2009) proposed a foreign accent conversion method which was claimed to be able to correct both prosody errors and segmental errors. The foreign accent conversion had two steps. The first step was a prosodic conversion that transformed the prosody of teacher's utterances into learner's utterances. In order to correct segmental errors, the second step was a segmental conversion that was based on the *source-filter* theory. The *source-filter* theory (Fant, 1960) of speech production hypothesizes that a speech signal can be seen as a glottal source excitation, passing through a filter that spectrally shapes the sounds of speech. The segmental conversion resynthesised the learner's utterances by combining the teacher's spectral envelop with the learner's glottal excitation. Felps et al. (2009) claimed that the resynthesised utterances by the foreign accent conversion sounded like the learner's own voice with a native accent.

However, the voice qualities after the prosodic conversion and segmental conversion (Felps, et al., 2009) were reported to be lowered to 2.96 and 2.67, respectively, on a 5-point scale, where a score of 1 meant bad voice quality and a score of 5 meant excellent voice quality. Thus, the voice quality of this approach needs to be improved before it can be used in 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. 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-quality 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, for some learners, hearing their own voices could be distracting, and could hinder them from improving their pronunciation.

4.1.2 Voices of multiple speakers

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, 2007). By listening to and imitating their favourite 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 (2007) 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, 2007).

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 speaker and a female speaker, 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 effort 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 suitable 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 for 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 a suitable speaker for different learners. The study

conducted by Probst et al. (2002) investigated "golden speakers" from the pronunciation improvement perspective. In other words, they studied which voice can help learners more effectively improve their pronunciation. In their study, the criteria for evaluating the effectiveness of different teachers' voices were the reductions of phone error and duration error. The subjects were randomly divided into three groups. Given six native speakers' voices, Group 1 subjects were allowed to choose one speaker's voice by themselves to imitate. Group 2 subjects imitated the voices that were the most similar to their own voices in term of pitch and speech rate, 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 practiced each sentence (3.5 times) fewer times than Group 2 (4.5 times) and group 3 (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 suitable teacher's voices for language learners to imitate 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.

4.2 In search of golden speaker from imitation preference perspective

We study the "golden speaker" from learners' imitation preference perspective. We investigate which 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.

We focus on three voice features: gender, pitch and speech rate. 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 resynthesise several sample voices with different prosodies based on the prosodic features (e.g. pitch and speech rate) of a learner's voice and the learners' imitation preferences.

Our prosody conversion transfers the prosodic features of a *learner's* voice to a *teacher's* voice, unlike previous reported prosody conversions that transfer the prosodic features of a *teacher's* voice to a *learner's* voice. Because our prosody conversion is based on a teacher's voice, the resynthesised utterances can be free from segmental error. Previous prosody conversions are normally based on a learner's voice (Nagano and Ozawa, 1990; Sundström, 1998; Hirose, 2004; Bissiri and Pfitzinger, 2009), which causes the resynthesised utterances to inevitably inherit the segmental errors (e.g. mispronounced phonemes) from the learner's utterances. Compared with teacher's speech, 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 resynthesise multiple sample voices with different prosodies by prosody modification. Compared with recording multiple teachers' voices, providing multiple sample voices based on the prosody modification reduces the effort of producing speech learning material and saves storage space in a computer. Also, the prosody modification can resynthesise voices with any prosodic features that language learners may prefer. By investigating learners' imitation

preferences, CAPT systems can then be developed to provide learners' favorite voices that may please them and promote their learning interests.

Our investigation of the golden speaker is different from previous research. We explore the golden speaker of a language learner from the imitation preference perspective. We investigate which 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. In our study, we focus on three voice features: gender, pitch and speech rate. To reduce the influence of other characteristics of teachers' voices (e.g. voice quality and clarity), CASTLE uses a single teacher's voice as the source to automatically resynthesise several sample voices with different prosodies based on the prosodic features of a learner's voice.

4.3 Prosody modification techniques

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

4.3.1 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 maintain its pitch values. PSOLA algorithms can be implemented both in the time domain and in the frequency domain. Considering the low computational complexity of the time-domain PSOLA, the duration modification in CASTLE is implemented in the time domain.

4.3.2 Pitch modification

Pitch modification changes pitch median $fMedian_{Ori}$ of an original teacher's utterance to new pitch median $fMedian_{Ind}$ according to a learner's preference. The new pitch values of the individualised 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. (4.1).

$$\begin{aligned} newPitch &= oldPitch * \Delta f \\ \Delta f &= fMedian_{Ind} / fMedian_{Ori} \end{aligned} \tag{4.1}$$

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 new one is to maintain the shape of a pitch contour in a perception scale (i.e. log scale),

$$Log(newPitch) = Log(oldPitch) + Log(fMedian_{Ind}) - Log(fMedian_{Ori}) \tag{4.2}$$

Thus, the relationship between the old and new pitch values in Eq (4.2) can be expressed as follows:

$$newPitch = oldPitch * \frac{fMedian_{Ind}}{fMedian_{Ori}} \tag{4.3}$$

The pitch modification, which is 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 140Hz 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 that 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 resynthesised 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[2]) 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 the formants of U_0. In the second step, the duration of U_1 is lengthened to the duration of U_0. It 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. It can be implemented by the pitch modification (refer to Section 2.2). In the last step, U_4 is

[2] U_i refers to the utterance after the i^{th} step modification. For example, U_0 is an original teacher's utterance, U_1 is a transit utterance after the first step of formants modification, and U_4 is the individualised teacher's utterance after the formants modification.

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 the duration and pitch values of U_0, but the formants of U_4 are 1.2 times of the formants of U_0.

Thus, a pitch modification is performed with a formants modification if the pitch median of an utterance 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.

4.4 Experimental setup

The experiments are to investigate how the voice features (i.e. gender, pitch and speech rate) 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, pitch and speech rate.

4.4.1 Speech material

The learning material was selected from the BU-RN corpus (Ostendorf, et al., 1995). For a detailed description of the BU-RN corpus, refer to Section 3.4.1. Two paragraphs PRLP2, PRLP4 uttered by female native speaker F1A were selected as the learning material. These two utterances were segmented into short portions. The duration of each segmented short portions ranged between 2s and 3s. 10 utterances, 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 experimental procedure to participants, and the other nine utterances were used as test utterances.

4.4.2 Participants

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; another seven of the subjects were aged between 30 and 39, and the other one subject was older than 49. 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 to 30 years. The duration of their living in English speaking countries ranged between 4 months to 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, where 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.

4.4.3 Procedures

The experiments were conducted on the CASTLE system. Figure 4.1 is a screenshot of CASTLE system providing individualised teacher's utterances with different prosodic features.

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 resynthesised individualised teacher's utterances based on the pitch median and speech rate of the subject's voice and the original teacher's utterance that was produced by native speaker F1A. There were three types of individualised teacher's utterances: speed similar and gender different utterances (SpS_GeD), speed different and pitch similar utterances (SpD_PiS), and prosody similar utterances (SpS_PiS). For each subject, the SpS_GeD 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_PiS 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_PiS 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_PiS utterances had similar pitch medians and speech rates to the subject's utterances. Since the pitch medians of SpS_PiS utterances and the subject's utterances were same, SpS_PiS utterances sounded like being produced by a speaker whose gender is the same as the subject.

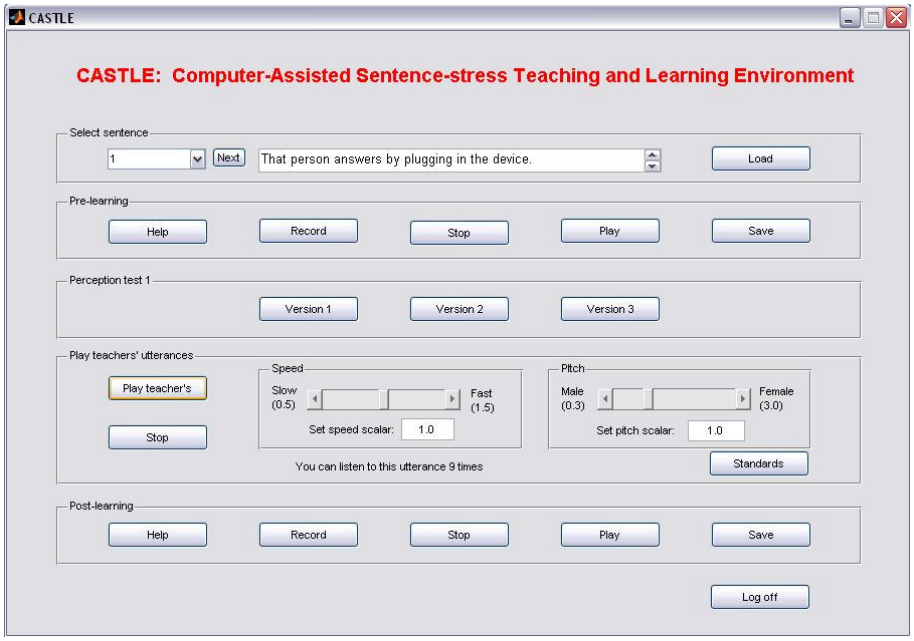


Figure 4.1 Screenshot of CASTLE system.

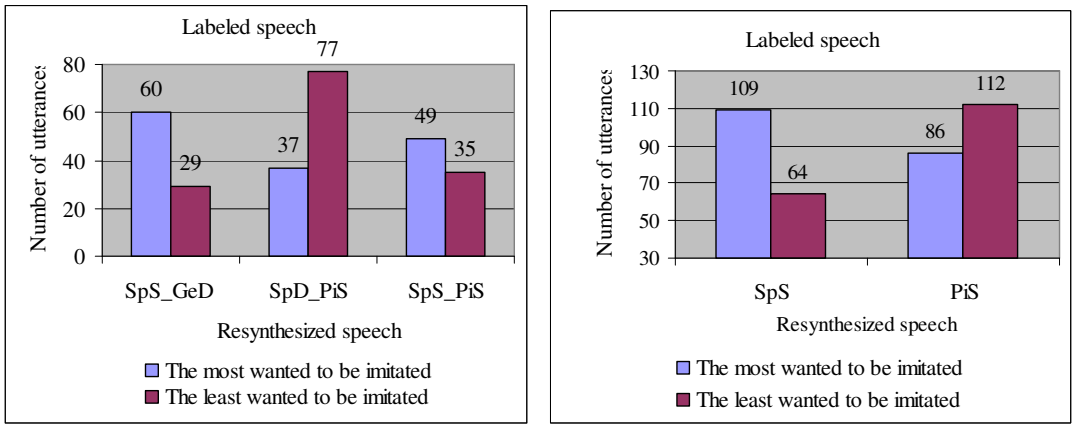
The orders of presenting the three individualised teacher's utterances are different for each learning sentence. The presenting order is in a loop format. For the first sentence of the learning material, the order was SpS_GeD, SpD_PiS and SpS_PiS. For the second sentence, the order was SpS_PiS, SpS_GeD and SpD_PiS. For the third sentence, the order was SpD_PiS, SpS_PiS and SpS_GeD. 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 individualised 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 individualised 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) individualised teacher's utterances favored by them to imitate. It is same for the "least wanted" labels. For each sentence, 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" individualised 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.4.1. When the pitch median was changed from the male (female) pitch range to the female (male) pitch range, the perceived gender of the resynthesised individualised teacher's utterances would change from male (female) to female (male). By allowing subjects to change pitch median and speed of the individualised teacher's utterances, we could observe whether their imitation preferences of a sentence change along with their practices.

4.5 Experimental results and discussions

The distributions of the most and least wanted to be imitated utterances labeled by the subjects are given in Fig. 4.2(a). Since for the three types of resynthesised individualised 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 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_GeD utterances which are more than the numbers of SpD_PiS utterances (37) and SpS_PiS utterances (49) labeled as "the most wanted utterances". In the utterances being labeled as the least wanted to be imitated, 29 are SpS_GeD utterances, 77 are SpD_PiS utterances, and 35 are SpS_PiS utterances.



(a) (b)

Figure 4.2 Distributions of the most and the least wanted to be imitated speech

From Fig.4.2(a), we can see that in the SpS_GeD utterances, the number of utterances labeled as "the most wanted" is as twice 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 preferred by learners to imitate. In the SpD_PiS 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_PiS 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 similar pitches and speech rates to their own voices.

We also noticed that all the SpD_PiS utterances resynthesised in our experiments were faster than the subjects' voices, since all the recorded subjects' voices 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. In the following we identify SpD_PiS utterances as having similar pitches to the subjects' voices but higher speeds than the subjects' voices, since all the SpD_PiS utterances resynthesised in our experiments were faster than the subjects' voices,.

The influences of speech rate and pitch of speech learning material on learners' imitation preferences are illustrated in Fig. 4.2(b). SpS refers to the resynthesised individualised teacher's utterances possessing similar speech rates to learners' utterances, which include both SpS_GeD and SpS_PiS utterances. Also, PiS denotes the resynthesised individualised teacher's utterances having similar pitch to learners' utterances, which include both SpD_PiS and SpS_PiS utterances. From Fig. 4.2(b), we can see that in the SpS utterances, the number of utterances with "the most wanted" label is nearly as twice 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 PiS 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 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. The male subject labeled more than half of the opposite gender voices as the "least wanted" voices. The other *eight* subjects 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 that 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 their home countries. They did not have any problem understanding radio news or TV programs. In contrast, seven subjects who labeled more than 6 fast teacher's utterances as "the least wanted" had a moderate 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 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_GeD) of the opposite gender to themselves. Two of the ten subjects showed preferences to fast speed voices (SpD_PiS). Another two of the ten subjects showed preferences to the voices (SpS_PiS) that 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_GeD) of the opposite gender to themselves. One of the five subjects preferred voices (SpD_PiS) having a same gender and a faster speed to their own voices. One of the five subjects preferred voices (SpS_PiS) 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 as their familiarity with the utterance increases. In the experiments, for each sentence, after labeling their imitation preferences, the subjects were asked to imitate a resynthesised 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, they then 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 with speech material may influence learners' imitation preferences.

In order to analyze each subject as an individual, Table 4.1 lists the average of the absolute deviations from the mean of the number of utterances that were labeled as "the most wanted to be imitated", and the average of the absolute deviations from the mean of the number of utterances that were labeled as "the least wanted to be imitated". The average of the absolute deviations from the mean is calculated by Eq (4.4)

$$AveDeviation = \sum_{i=1}^{n} | x_i - \bar{x} | \tag{4.4}$$

where $\bar{x}$ is the mean of x_1, x_2, … x_n. For example, if a subject chooses 3 SpS_GeD utterances, 6 SpD_PiS utterances and 0 SpS_PiS 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 4.1, 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.

Table 4.1 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

Note that the quality of the resynthesised individualised teacher's utterances in our study is 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 resynthesising one female native speaker's voices, they were amazed. Few subjects realized that there were some minor distortions in the individualised teacher's utterances, but they felt that the individualised teacher's utterances were all still on an acceptable level. Their imitation preferences 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 pitches and speech rates to learners' own voices are not always what they prefer to imitate.

4.6 Conclusions

Our experimental results show that a teacher's voice, which has similar pitch and speech rate 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. 4 out of 15 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. 7 out of 15 subjects in our experiments preferred a slow version of the speech material to catch their unfamiliar pronunciation features (e.g. linking, assimilation, elision). These seven subjects had a

moderate 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 as their familiarity with the utterance increases.

In order to meet learners' different imitation needs, we advocate an automatic prosody modification function in CAPT systems to provide speech material with a wide variety of prosodic features. For a CAPT system, automatic prosody modification can be used to resynthesise voices with learners' preferred prosodic features. Learners then can have an opportunity of listening to voices with more variations.

Thus, in the individualised speech material module of CASTLE, we employ the automatic prosody modification functions to provide speech material with learners' preferred voice features. In CASTLE, we also allow learners to control the prosody modification. In order to listen to speech material with different voice features, learners can drag the sliders or adjust the pitch value and speech rate changing factors. This prosody modification control gives learners more autonomy in the process of learning.

4.7 Summary

In order to provide learners with their preferred voice for them to imitate in language learning sessions, in this chapter, we have investigated which voice features (i.e. gender,

pitch and speech rate) make a teacher's voice preferable for a language learner to listen to and to imitate.

Our investigation of the "golden voice" is different from the study conducted by Probst et al. (2002). Probst et al. (2002) takes multiple native speakers' voices as teacher's voices provided to ESL learners for pronunciation learning. In our study, based on a single teacher's voice and the prosodic features of a learner's voice, several individualised teacher's voices with different prosodies are automatically resynthesised. Since the individualised 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.

Also, the criterion for evaluating the effectiveness of different teacher's voices in our study is different from that of the study conducted by Probst et al. (2002). The evaluation criteria used by Probst et al. (2002) are the improvements of learners' pronunciation. However, learners' pronunciation improvements may also be influenced by other factors such as learning motivation and learning ability. Considering of this, in our study, we assessed different teacher's voices from learners' imitation preference point for view. Moreover, providing learners' favourite voices is importance for CAPT systems since it may help to develop a pleasant learning environment and increase learners' learning interests.

Our experimental results show that learners' imitation preferences are influenced by many factors such as English background, English language proficiency and learning stages. We found that voices with similar pitch values and speech rates to learners' own voices are not always what they prefer to imitate. Our investigation concludes that different people may have different imitation preferences.

Thus, we advocate that CAPT systems should have an automatic prosody modification function that can automatically resynthesise speech learning material with learners preferred voice features, given teacher's utterances. In the individualised speech material module of CASTLE, we employ the automatic prosody modification functions to resynthesise individualised speech material. The module also allows learners to control the prosody modification to generate their preferred voices for them to mimic.

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 three voice features (i.e. gender, pitch and speed). However, there is a lack of clear evidences of how providing learners' preferred voices could 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.

Chapter 5.

Exaggeration-based Perception Assistance Module

Since it is challenging for some ESL learners to correctly perceive sentence stress as discussed in Section 2.1.2, we propose an exaggeration-based perception assistance module that is intended to help ESL learners to correctly perceive sentence stress. In the perception assistance module, we present a set of automatic stress exaggeration methods to enlarge the differences between stressed and unstressed syllables in teachers' speech.

This chapter is organised as follows: In Section 5.1, we introduce the *hyper-pronunciation training* method, inspired by which we propose to automatically resynthesise stress-exaggerated speech learning material by prosody modification techniques. Section 5.2 reviews the literature related to prosody modification for pronunciation training. In Section 5.3, we present our stress exaggeration methods that exaggerate the prosodic differences between stressed and unstressed syllables. Section 5.4 describes the perceptual experiments that we conducted to evaluate the effectiveness of our stress exaggeration methods. Section 5.5 summarises our exaggeration-based perception assistance module.

5.1. Hyper-pronunciation training

Exaggeration has been advocated by Todaka (1995) to help ESL learners increase their awareness of English-specific acoustic features and effectively apply these features into their spoken English. This exaggeration-based pronunciation training method is called *hyper-pronunciation training*. The effectiveness of the hyper-pronunciation training has been reported in empirical studies (Bissiri and Pfitzinger, 2009; Nagamine, 2002).

Nagamine (2002) employed the hyper-pronunciation training method to teach Japanese speakers English intonation. All students in Nagamine's study, who received hyper-pronunciation training, showed dramatic improvement in terms of producing utterances with pitch ranges and pitch contours closer to native English speakers, although the

improvement in pitch did not lead to an improvement in native speakers' comprehensibility. Nagamine (2002) claimed that phrase boundaries or pauses were more influential factors than an appropriate use of F0 shape to affect perceived comprehensibility.

In the study conducted by Bissiri and Pfitzinger (2009), stimuli with normal stress and emphasized stress were used to teach Italian speakers lexical stress of German. The pronunciation improvements of learners who were trained through stimuli with emphasis on stressed syllables were more significant than the pronunciation improvements of learners who were trained through normally stressed stimuli.

In most studies that employed the hyper-pronunciation training method (Bissiri and Pfitzinger, 2009; Nagamine, 2002), teachers' utterances with exaggerated pronunciation are necessary. Moreover, in the studies in Todaka (1995) and Nagamine (2002), the speech learning material with exaggerated prosodic features was directly provided by English teachers pronouncing learning material exaggeratedly in English language teaching classes. However, this classroom-based teaching model cannot satisfy learners' needs of learning English. Nowadays, learners prefer to be able to study and practise their pronunciation whenever and wherever they want to, rather than just in classroom settings.

Yoon (2008) introduced techniques of automatically resynthesising utterances with exaggerated prosody by manipulating either the fundamental frequency (F0) contour, the segmental durations, or the intensity contour of an utterance. The automatic prosody exaggeration techniques presented in Yoon (2008) increased the pitch, duration and intensity of all F0 peak areas, and decreased these of all F0 valley areas. Yoon indicated that these techniques could be employed to develop a tool for prosody teaching. However, the automatic prosody exaggeration techniques introduced by Yoon (2008) are not suitable to teach English stress. Firstly, English stress is not always associated with F0 peaks. Some syllables can be stressed by a lower pitch and longer duration such as L* type pitch accent (refer to the ToBI transcription in Section 3.4.1). Secondly, the intensity exaggeration in Yoon (2008) dealt with overall intensity that was claimed to be not a reliable indicator of stress (van Katwijk, 1974). High frequency intensity, instead of overall intensity, is a more reliable cue for stress perception (Sluijter, et al., 1997).

Therefore, in this chapter, we present a set of automatic stress exaggeration methods that are more suitable to teach English stress. Our automatic stress exaggeration methods can be employed in CAPT systems to generate the stress exaggerated teaching material automatically and immediately by a computer whenever learners need them.

5.2. Pronunciation training based on prosody modification

With the development of speech synthesis, numerous studies have been reported to use prosody modification for second language pronunciation training. Most of them transform the prosody of teacher's utterances into learners' utterances, in order to resynthesise the learners' utterances with native-like prosody while maintaining the learners' voice characteristics (Hirose, 2004; Nagano and Ozawa, 1990; Sundström, 1998). The notion is that the resynthesised utterances give learners an opportunity to listen to their own voices with correct prosody. However, only modifying the prosody of learners' voices can also make the resynthesised utterances inherit segmental errors (e.g. mispronounced phonemes) from the learners' original utterances, which are unavoidable in language learner' speech. Thus, these prosody-modified learners' voices are not suitable for learners to imitate.

Our work differs from these previous studies described above as our stress exaggeration methods are based on teachers' utterances that are unlikely, or much less likely to have segmental errors than learners' speech. This makes our resynthesised prosody-exaggerated utterances more likely segment-error free.

The foreign accent conversion proposed in Felps et al. (2009) was 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, where a score of 1 meant bad voice quality and a score of 5 meant excellent voice quality. Thus, the voice quality of the foreign accent conversion needs to be improved before it can be used in CAPT systems. (For a more detailed description of the foreign accent conversion method proposed by Felps et al., refer to Section 4.1.1.)

Hincks (2002) used WaveSurfer (Beskow and Sjölander, 2000), an open source tool for sound visualization and manipulation (www.speech.kth.se/wavesurfer), to teach Swedish speakers the English lexical stress of individual technical *words*. The learners could use WaveSurfer to modify the duration and pitch of reference utterances, and listen to the resynthesised utterances. Experimental results showed that the learners' production had been significantly improved after the perceptual training. Our study is also different from the work carried out by Hincks (2002). Hincks only manipulated the duration and pitch of individual words, and Hincks's method was to teach lexical stress in individual words. In contrast, we manipulate the duration, pitch and intensity in the context of sentences. In Hincks (2002), learners were asked to lengthen or shorten the duration of sound by themselves, while in our system the prosodic parameters are manipulated automatically.

Bissiri and Pfitzinger (2009) taught Italians the lexical stress of German morphologically complex words by prosody modification. They dealt with both normal speech and exaggerated speech. The exaggerated speech, which emphasized stressed syllables, was resynthesised by copying the prosodic parameters (i.e. local speech rates, intonation contours and intensity contours) of German teachers' utterances with emphasized stress to the utterances of Italian learners. In the study carried out by Bissiri and Pfitzinger, German teachers' utterances produced with exaggerated prosodic features were necessary. Also, our stress exaggeration methods are different from Bissiri and Pfitzinger's (2009) emphasis method that needs native speakers' utterances produced in an exaggerated way. Our stress exaggeration methods generate exaggerated utterances from normal speech (i.e. unexaggerated utterances). Thus, our methods have wide applications, as native speakers' deliberately exaggerated utterances are not necessary in our approaches and exaggerated utterances are normally unavailable. Moreover, our stress exaggeration methods can exaggerate the differences between stressed and unstressed syllables on different levels (e.g. to elongate a stressed syllable to 1.5 or 2 times of its normal duration) to meet the different needs of different learners.

As indicated in Section 5.1, although the automatic prosody exaggeration techniques introduced by Yoon (2008) can exaggerate the prosodic features of an utterance, these techniques are not suitable to teach English stress, because stressed syllables are not always associated with high pitch values and overall intensity is not a reliable indicator

of stress (van Katwijk, 1974). Our stress exaggeration methods differ from Yoon's prosody exaggeration techniques. The aim of our stress exaggeration methods is to make stress patterns of utterances more perceivable. Thus, our system enlarges the prosodic differences between stressed and unstressed syllables. For pitch, we manipulate both pitch changes and pitch levels since stressed syllables can be associated with significant pitch movements or high pitch levels (refer to Section 5.3.1), while Yoon (2008) only considered pitch changes (i.e. pitch movements). For duration, we elongate the durations of stressed syllables (refer to Section 5.3.2) that are not necessarily associated with F0 peaks. For intensity, we manipulate the high frequency band, in which stressed and unstressed syllables have considerable differences (refer to Section 5.3.3), while Yoon (2008) dealt with overall intensity that was claimed not to be a good indicator of stress (van Katwijk, 1974).

In the above, we reviewed literature focusing on prosody modification for foreign language learning, which are related to our present research. A comprehensive review of speech technology for education can be found in Eskenazi (2009).

5.3. Automatic stress exaggeration

Inspired by the hyper-pronunciation training method, we present a set of stress exaggeration methods that can automatically resynthesise exaggerated stimuli, in order to help ESL learners perceive English stress correctly. The resynthesised stimuli exaggerate the differences between stressed and unstressed syllables, given the stress patterns of an utterance. In this chapter, we assume that the stress patterns of an utterance are available, since they can be obtained by the automatic stress detection techniques introduced in Section 3.4.

Prosodic features (i.e. pitch, duration and intensity) and vowel quality (Sluijter, et al., 1997; Xie, et al., 2004a) are the acoustic cues to indicate English stress. Stressed syllables tend to be pronounced longer, louder, and with significant pitch movements (Dalton and Seidlhofer, 1994). Pitch and duration are generally claimed to be important for stress perception. Overall intensity is claimed to be a less important parameter in determining stressed syllables (van Katwijk, 1974). However, studies (Sluijter and van

Heuven, 1996; Sluijter, et al., 1997; Tamburini and Caini, 2005) found that the intensity of stressed and unstressed syllables had a considerable difference in high frequency band, which made high frequency intensity a reliable cue for stress perception. Sluijter et al. (1997) argued that the high intensity in the high frequency band of stressed syllables was due to an increase in vocal effort when producing stressed syllables, which could be perceived as a greater loudness. The vowels of *stressed* syllables tend to be full, while the vowels of *unstressed* syllables tend to be reduced (Solé Sabater, 1991; Xie, et al., 2004a). However, full vowels can also appear in *unstressed* syllables. Thus vowel quality is not a reliable indicator of stress (Xie, et al., 2004a).

We propose a set of stress exaggeration methods to enlarge the differences between stressed and unstressed syllables. Vowel quality features are not manipulated in our present research, considering its low reliability to indicate stress and its high complexity of resynthesise. There are four stress exaggeration methods discussed in this section: (i) pitch-based exaggeration, (ii) duration-based exaggeration, (iii) intensity-based exaggeration, and along with (iv) a combined exaggeration method that combines the previous three stress exaggeration methods. In the pitch-based stress exaggeration, pitch level and pitch movement are taken into consideration. Our method increases (or decreases) the pitch level of a stressed syllable if the stressed syllable has a high pitch level, or expand its pitch range if the stressed syllable captures a significant pitch movement. In the duration-based stress exaggeration, our proposed method smoothly lengthens the duration of stressed syllables and shortens the duration of unstressed syllables. In the intensity-based stress exaggeration, the high frequency intensity of stressed syllables is increased. The combined stress exaggeration is to apply all the three single-prosody-feature-based stress exaggeration operations (i.e. exaggerating pitch-based, duration-based and intensity-based features) to stressed and unstressed syllables.

5.3.1 Pitch-based stress exaggeration

Both pitch levels and pitch movements can contribute to stress syllables. Stressed syllables may relate to higher (or lower) pitch levels, or significant pitch movements. According to the ToBI (Tones and Break Indices) transcription system (Silverman, et al., 1992), there are eight types of tonal events associated with stressed syllables, which are

listed in the first two columns of Table 5.1. For a description of the ToBI transcription system, refer to Section 3.4.1.

We have developed two types of pitch-based stress exaggeration techniques: (i) exaggeration based on pitch level, and (ii) exaggeration based on pitch movement. In the 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. In the exaggeration based on pitch movement, the pitch range of a stressed syllable is expanded if the syllable is stressed by a significant pitch movement. In our CASTLE system, the pitch-based stress exaggeration is either based on pitch level or based on pitch movement, depending on the ToBI labels of the stressed syllable. The exaggeration operation for each ToBI label is listed in the third column of Table 5.1.

Table 5.1 ToBI labels and their corresponding exaggeration operations.

ToBI labels	Description	Exaggeration
H*	High pitch accent	Based on pitch level
L*	Low pitch accent	
L+H*	Bitonal low tone with high tone on stressed syllable	Based on pitch movement
L*+H	Bitonal high tone with low tone on stressed syllable	
!H*	Downsteped high pitch accent	
L+!H*	Bitonal pitch accent with low tone followed by a downstepped high tone prominence	
L*+!H	Bitonal pitch accent with low tone prominence followed by downstepped high tone	
H+!H*	Bitonal pitch accent with high tone followed by downstepped high prominence	

Exaggeration based on pitch level

This exaggeration technique is used to exaggerate the effect of pitch level on a stressed syllable by multiplying its pitch contour with a positive pitch changing factor, Δf , as it is shown in Eq. (5.1),

$$\begin{aligned} newPitch_{Level} &= oldPitch * \Delta f \\ \Delta f &= newMedian / oldMedian \end{aligned} \tag{5.1}$$

where Δf is the ratio of the new pitch median to the old pitch median. A detailed description of pitch level modification can be found in Section 4.3.2. 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 accent, Δf is set to be less than 1, in order to make its pitch level even lower.

In our experiment, Δf is set to 1.2 for stressed syllables associated with an H* pitch accent. This is equivalent to adding a constant of 0.263 on an octave scale or 3.156 on a semitone scale, which is shown as follows:

$$Octave_{Increase} = \frac{\log(newPitch_{Level}/100)}{\log 2} - \frac{\log(oldPitch/100)}{\log 2} \tag{5.2}$$

$$= \log(newPitch_{Level}/oldPitch) = \log(1.2) = 0.263$$

$$Semitones_{Increase} = 12 * Octave_{Increase} = 3.156 \tag{5.3}$$

Δf is set to 0.8 for stressed syllables associated with an L* pitch accent. It is equivalent to adding a constant of -0.3219 on an octave scale or -3.8631 on a semitone scale.

Exaggeration based on pitch movement

The exaggeration based on pitch movement is to expand the pitch movement of a stressed syllable by multiplying its pitch range with a scale factor, Δp, as it is shown in Eq (5.4).

$$newPitch_{Movement} = PitchMedian + (oldPitch - PitchMedian) * \Delta p \tag{5.4}$$

To expand the pitch movement, Δp has to be greater than 1. The exaggeration based on pitch movement vertically stretches the pitch contour of a stressed syllable, and in the meantime, remains its pitch median unchanged. A pitch value closer to the pitch median has a smaller change than a pitch value diverging from the pitch median. Figure 5.1 illustrates the pitch contour difference between an originally stressed syllable and its corresponding resynthesised stressed syllables with the exaggeration based on pitch movement. As shown in Figure 5.1, the exaggeration based on pitch movement makes the stressed syllable have a wider pitch range. In our experiment, Δp is set to 1.5 for stressed syllables and 0.5 for unstressed syllables.

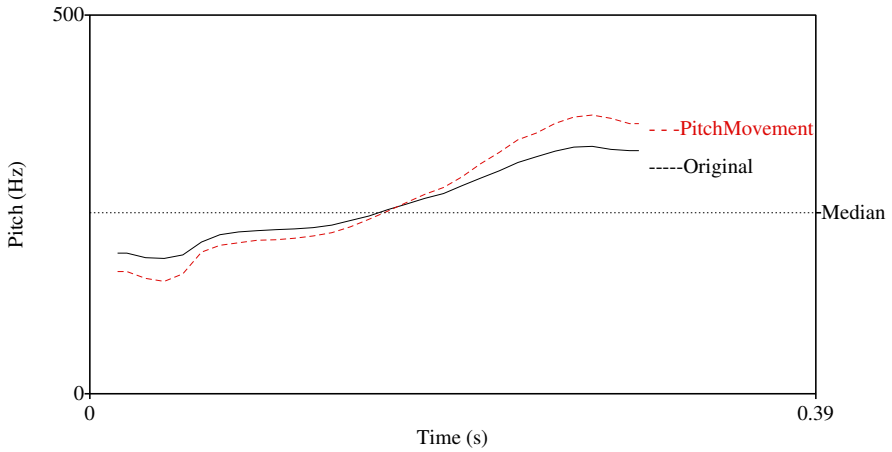


Figure 5.1 Pitch contour comparison.

5.3.2 Duration-based stress exaggeration

The duration-based stress exaggeration is to elongate the durations of stressed syllables, and shorten the durations of unstressed syllables to make the differences between stressed and unstressed syllables more noticeable. A duration exaggeration contour is generated according to the stress patterns of an utterance. In a duration exaggeration contour, a peak means to increase the duration of its corresponding stressed syllable, while a valley means to decrease the duration of its corresponding unstressed syllable. The duration exaggerated stimuli are resynthesised by multiplying the original utterances with its duration exaggeration contour.

Stress has more influence on the duration of a syllable nucleus than the durations of a syllable onset or a syllable coda (i.e. the part of a syllable that precedes or follows a syllable nucleus). The study carried out by Ananthakrishnan and Narayanan (2008) found that syllable nucleus duration is the most important acoustic cue for stress. In our experiment, the duration exaggeration of a syllable nucleus is set to be greater than the duration exaggeration of syllable onsets and syllable codas. For stressed syllables, we set the duration enhancing factor of syllable nuclei to 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 to 0.8, and it linearly increases to 1 for the syllable onsets and the syllable codas. An example is illustrated in Figure 5.2. The utterance in this example is “The device is attached to a plastic wristband.”. In Figure 5.2, from the

first panel to the last panel, they are waveform, pitch contour, ToBI label of the stressed syllables, phonetic transcription, word transcription and duration exaggeration contour. In the duration exaggeration contour, the centre dash line corresponds to a factor of 1, above which corresponds to an elongation of the syllable durations and below which corresponds to a shortening of the syllable durations.

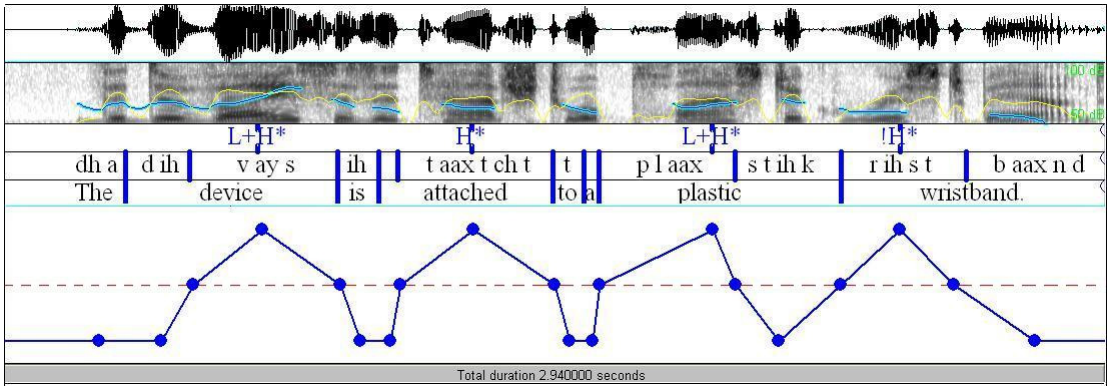


Figure 5.2 Duration-based stress exaggeration

5.3.3 Intensity-based stress exaggeration

The intensity differences between stressed and unstressed syllables lie in the high frequency band (i.e. above 500Hz) (Sluijter and van Heuven, 1996; Sluijter, et al., 1997; Tamburini and Caini, 2005). Thus, 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 reducing its high frequency band.

For stressed syllables, the intensity-based stress exaggeration is implemented by increasing 9dB of the frequencies above 500Hz. Increasing 9dB of the high frequency band that is greater than 500Hz, is equivalent to multiplying the values in that region by a factor of 2.8184 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 2.8184 linearly within a frequency band of 100Hz. Figure 5.3 illustrates the spectrums of an originally stressed syllable and its corresponding intensity exaggerated equivalent. For unstressed syllables, the intensity-based stress exaggeration is implemented by decreasing 5dB of the frequencies above 500Hz.

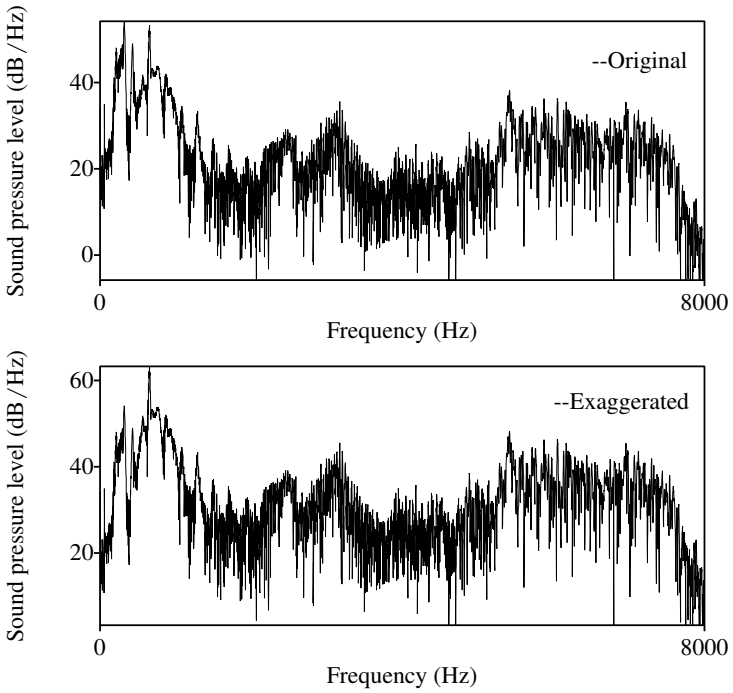


Figure 5.3 Spectrum comparison.

5.3.4 Combined stress exaggeration

A combined stress exaggeration method is used to investigate the interactions among the three previously described single-prosody-feature-based exaggeration methods, which integrates the pitch-based, duration-based and intensity-based stress exaggeration methods. In the combined stress exaggeration, pitch-based exaggeration and intensity-based exaggeration are manipulated first on original stimuli in a cascaded manner, then a duration-based exaggeration is implemented on the pitch and intensity exaggerated stimuli.

Note that the parameters used in this section (e.g. pitch changing factor Δf , pitch range scale factor Δp) are for demonstration purpose, and were also used in our perceptual experiments. In the CASTLE system, learners can adjust these parameters freely according to their imitation preferences.

5.4 Perceptual experiments

Perceptual experiments were conducted to evaluate the effectiveness of the four stress exaggeration methods described in Section 5.3. The experiments were to test the following hypotheses: (i) whether the resynthesised stimuli can help ESL learners to perceive English stress more accurately; (ii) which one of the three single-prosody-feature-based stress exaggeration methods is more effective to help English stress perception; (iii) whether the combined stress exaggeration is more effective than the three single-prosody-feature-based exaggeration methods.

5.4.1 Participants

Fifteen non-native English speakers voluntarily participated in the perceptual test. Seven were male and eight were female. Fourteen of the participants were aged between 20 and 49, and the other one participant was older than 49. Their first languages were: Japanese, Spanish, Urdu, Malay (N=2), Persian (N=2), and Chinese (N=8). They had a history of learning English for between 5 to 30 years. They had been living in an English speaking country for a duration ranging from 4 month to 9 years.

5.4.2 Speech material

The learning material was selected from the BU-RN corpus (Ostendorf, et al., 1995). For a detailed description of BU-RN corpus, refer to Section 3.4.1. Five paragraphs, PRLP2, PRLP4, RRLP2, RRLP3 and RRLP4 uttered by female native speaker F1A were selected as learning material. The duration of the five utterances ranged between 11s and 55s. There were hand-labeled ToBI format stress patterns and hand-corrected automatic phone alignments associated with each utterance. These five utterances were segmented into short portions with a duration ranging from 2s to 3s. 16 utterances were selected from the segmented short speech portions as the learning stimuli.

These 16 utterances were distributed as follows. One of the 16 utterances was selected as an “example utterance” to demonstrate the test procedure for participants. The other

15 utterances were divided into 5 clusters. Each of them had 3 sentences. In each cluster, the number of syllables ranged from 39 to 49, and the number of stressed syllables ranged from 14 to 15. The distribution of syllables and stressed syllables in each cluster is given in Table 5.2.

Table 5.2 Distribution of syllables and stressed syllables in sentences and clusters.

Sentence	Stressed syllables / Syllables	Stressed syllables / Syllables	Cluster
1	3/15	14/43	1
2	7/15		
3	4/13		
4	3/12	15/39	2
5	6/15		
6	6/12		
7	5/13	15/39	3
8	4/12		
9	6/14		
10	5/18	14/49	4
11	5/17		
12	4/14		
13	5/14	14/39	5
14	4/13		
15	5/12		
16	4/12	/	Example utterance

Different utterance clusters might have slightly different stress labeling difficulties. In order to reduce the influence of different stress labeling difficulties of different utterance clusters in the perceptual test, we reordered the utterance clusters. There were five types of rearrangements of the clustered listening material as shown in Table 5.3. Each utterance cluster was considered as a group. For each type of listening material, the utterances in the first group were not modified, which were just original native speakers' utterances; the utterances in the other four groups were resynthesised by different stress exaggeration methods, which are given in the last column of Table 5.3. The utterances in the second group were resynthesised with the pitch-based stress exaggeration. The duration-based stress exaggeration was operated on the utterances in the third group. The utterances in the fourth group were resynthesised with the intensity-based stress exaggeration. And the utterances in the fifth group were resynthesised by the combined stress exaggeration.

In the perceptual test, each subject was randomly assigned one type of listening material to label their stress patterns.

Table 5.3 Distribution of utterance clusters in each type of listening material.

Group	Type 1	Type 2	Type 3	Type 4	Type 5	Exaggeration
1	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Original
2	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 1	Pitch-based
3	Cluster 3	Cluster 4	Cluster 5	Cluster 1	Cluster 2	Duration-based
4	Cluster 4	Cluster 5	Cluster 1	Cluster 2	Cluster 3	Intensity-based
5	Cluster 5	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Combined

5.4.3 Results and discussion

Recall, *precision* and *F-measure* (Engelbrecht, et al., 2009) are employed to evaluate the accuracy of the listeners' stress pattern labeling, which are widely used classification accuracy criteria. *Recall* is a measure of completeness. In this chapter, *recall* is defined as the number of syllables correctly labeled as stressed divided by the total number of stressed syllables. *Precision* is a measure of exactness, which is defined as the number of syllables correctly labeled as stressed divided by the total number of syllables labeled as stressed. The higher a *recall* score is, the more stressed syllables are correctly labeled by the listener. Also, the higher a *precision* score is, the more the syllables that are labeled as stressed by a listener are truly stressed. *F-measure* is the weighted harmonic mean of *precision* and *recall*, as it is shown in Eq.(5.5),

$$F\text{-}Measure = \frac{(1+\alpha) * precision * recall}{\alpha * precision + recall} \tag{5.5}$$

where α is a positive weight of *precision*. If α is set to 1, the *F-measure* weights *recall* and *precision* equally, which is called balanced *F-measure*. Our experiments treat *recall* and *precision* equally. Then we employ the balanced *F-measure* in our experiments. The average *recall*, *precision* and balanced *F-measure* of each listening material group are listed on Table 5.4.

Table 5.4 Comparison of listeners' stress pattern labeling accuracy

Measures (Average)	Listening material groups				
	Original	Pitch-based	Duration-based	Intensity-based	Combined
Recall	0.504	0.608	0.676	0.620	0.707
Precision	0.742	0.791	0.822	0.794	0.859
F-measure	0.623	0.700	0.749	0.707	0.783

The experimental results show that all the three single-prosody-feature-based stress exaggeration methods (pitch-based, duration-based and intensity-based), along with the combined stress exaggeration, have improved the accuracy of the listeners' stress pattern labeling. The exaggerated utterances of the combined method are more helpful to improve the listeners' stress perception than the exaggerated utterances of all the three single-prosody-feature-based exaggeration methods. The highest scores of *recall*, *precision* and *F-measure* are achieved by participants labeling the resynthesised utterances that are manipulated by the combined stress exaggeration method. Among the three single-prosody-feature-based exaggeration methods, the exaggeration of duration is more effective than the other two methods. Moreover the resynthesised utterances improved both *recall* and *precision*. Since balanced *F-measure* is a combination of *recall* and *precision*, in the following analysis, we only use balanced *F-measure* as an integrated criterion. Figure 5.4 shows the boxplots (i.e. minimums, first quartiles, medians, third quartiles and maximums) of the balanced *F-Measures* of the listeners' stress pattern labeling on the original and the resynthesised utterance groups.

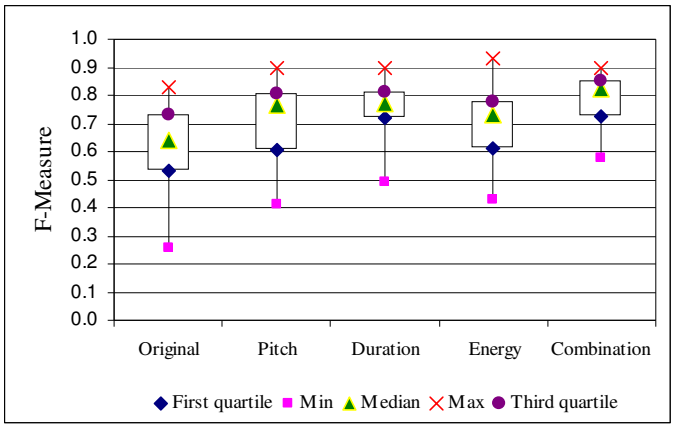


Figure 5.4 Boxplot of the *F-Measures* of listeners' stress pattern labeling.

The Student's t-Test is used to investigate the stress perception differences between the original utterances and the resynthesised utterances. We found that (i) the *F-measure* difference between the original utterances and the pitch-based exaggerated utterances, and the *F-measure* difference between the original utterances and the intensity-based exaggerated utterances are significant at the significance level of 0.05; and (ii) the duration-based exaggerated utterances and the utterances exaggerated by the combined method significantly improved the subjects' stress labeling accuracy at the significance level of 0.01.

As we expected, the combined stress exaggeration method is more effective than every single-prosody-feature-based exaggeration method since different prosodic parameters may interact with each other and the interactions make stress more outstanding. Among the three single-prosody-feature-based stress exaggeration methods, duration-based exaggeration turned out to be the most effective resynthesis method to increase listeners' awareness of stress. This is consistent with the studies conducted by Xie et al. (2004a), and Ananthakrishnan and Narayanan (2008) which show that duration is the most powerful indicator of English stress.

Another interesting finding of the experiments is that the relatively low effectiveness of the pitch-based and intensity-based stress exaggerations is partly caused by assigning stress labels to unstressed syllables that are in the same words with the stressed syllables. An explanation of these stress-labeling shifts may be that these syllables were uttered so fast that the listeners could not locate the prosodic features to the right positions. Some listeners complained that the utterances were too fast, and they wished to slow them down. However, the slowdown function was not available in the perception test. Thus, the resynthesis that elongates stressed syllables makes the stressed syllables more outstanding and easier to be located. Also, when resynthesis combines the pitch-based and intensity-based stress exaggerations along with duration-based exaggeration, it is much easier for the listeners to locate the positions of stressed syllables accurately.

5.5 Summary

In order to help ESL learners perceive English stress correctly, we have presented a set of stress exaggeration methods in this chapter, which include exaggerations of pitch-based, duration-based, intensity-based prosodic features, and a combined stress exaggeration method integrating the three single-prosody-feature-based stress exaggeration methods. The effectiveness of the four stress exaggeration methods has been investigated in our perceptual experiments. The results of our perceptual experiments showed that: (i) all the three single-prosody-feature-based stress exaggeration methods alone improved ESL listeners' English stress identification accuracy; (ii) among the three single-prosody-feature-based stress exaggeration methods, the duration-based exaggeration method was the most effective one; (iii) and the

combined stress exaggeration method improved the listeners' English stress perception accuracy more significantly than every single-prosody-feature-based method; (iv) the combined method and the duration-based exaggeration method helped the listeners to improve their stress labeling accuracy at the significance level of 0.01 for a Student's t-Test, and the pitch-based and intensity-based methods helped the listeners to improve their stress labeling accuracy at the significance level of 0.05.

Our stress exaggeration methods can be employed in CAPT systems to support hyper-pronunciation training that has been showing its effectiveness. Currently, the exaggerated speech materials of hyper-pronunciation training are produced by teachers in traditional classroom-based pronunciation teaching. This cannot satisfy learners' needs of learning pronunciation anytime and anywhere. Our stress exaggeration methods can resynthesise prosody-exaggerated utterances without human interaction. Thus, the automatic stress exaggeration methods presented in this chapter can be used to support CAPT-based hyper-pronunciation training.

We also notice that there are some research issues in the exaggeration-based perception assistance module worth to be further investigated. One of them is whether stress exaggeration levels specified by learners could be more effective to help the learners detect English stress than fixed exaggeration levels. In order to simplify the description of the exaggeration methods and the perceptual experiment procedure, we used fixed exaggeration levels in the present research. However, learners specified exaggeration levels might be more effective since the learners can adjust the exaggeration levels according to their perception preferences. Learners specified exaggeration levels might also lower the voice quality of resynthesised stimuli since we cannot expect that all learners possess the knowledge needed in stress exaggeration.

Another issue worth investigating is whether the hyper-pronunciation-based English stress perception training can help learners to increase their ability to perceive stress in normal speech (unexaggerated speech). A positive effect is expected since once learners become aware of how acoustic features to be used to stress syllables, they should be more sensitive to the differences between stressed and unstressed syllables. However, raising the awareness of the functions of prosodic features in stress may take an extended period of time.

Chapter 6.

Production Assistance Module

In order to help ESL learners produce sentence stress correctly, we design a production assistance module, in which we propose and develop a clapping-based pronunciation practice assistance model and three stress-error feedback models. The clapping-based pronunciation practice model is to help ESL learners to be aware of the rhythm of English language and to train them to get used to place more emphasis on syllables that are supposed to be stressed. By analysing the limitation of the conventional categorical representation of stress, we propose to use fuzzy sets to represent stress. Based on the fuzzy representation of stress, three stress-error feedback models are presented, which are intended to provide learners with valuable feedback in order to help them realise and correct their stress errors.

This chapter is organised as follows: In Section 6.1, a clapping-based pronunciation practice assistance model is proposed. In Section 6.2, we present a limitation of the conventional categorical representation of stress, and advocate a fuzzy representation. Based on the fuzzy representation of stress, three stress-error feedback models are presented in Section 6.3. The flowchart of the production assistance module is illustrated in Section 6.4. Section 6.5 summarises our production assistance module.

6.1 Clapping-based pronunciation practice assistance model

6.1.1 Clapping in pronunciation learning

Sentence stress is closely related to other aspects of suprasegmental features such as rhythm. The rhythm of English language is composed of regular occurrences of stressed syllables (Dalton and Seidlhofer, 1994). Therefore, sentence stress learning can benefit from rhythm learning.

Clapping, as well as other body languages such as snapping figures and stepping, has been used in classrooms, in order to help ESL learners to recognise the rhythm of English language (Fan, et al., 1998; Florez, 1998). English is a rhythmic language, in which some syllables are given more emphasis than others. However, some ESL learners tend to incorrectly stress words (or syllables), or they stress every word in an utterance more or less equally (Hahn, 2004). This incorrect stress and rhythms of ESL learners' speech have a negative impact on the intelligibility of their speech (Cutler and Clifton, 1984; Field, 2005), or it can even make the learners difficult to be understood. In order to raise the ESL learners' awareness of the stress and rhythm of English language, a popular approach is to ask ESL learners to clap hands at stressed syllables.

However, some non-native English-speaking teachers are not used to, or they are reluctant to teach rhythm, due to their lack of professional knowledge and confidence in teaching in this way (Ilčiukienė, 2005). In order to teach rhythm through body languages such as clapping hands and snapping fingers, teachers need to be able to correctly perceive and extract rhythms from utterances themselves. Extracting rhythms from utterances is a challenging task for non-native English-speaking teachers, although they may be able to speak English fluently (Gong, 2002). Without practice or training, rhythm extraction from utterances is also challenging for some native speakers (Orton, 2000). As indicated by Gong (2002), non-native English teachers in China were more willing to teach rhythm using CAPT systems rather than by themselves since computer software required less professional knowledge from them.

Thus, in order to help ESL learners to be aware of the rhythm of English language and to produce sentence stress correctly, we present a Clapping-based Pronunciation Practice Assistance (CPPA) model.

6.1.2 Description of the CPPA model

Given a teacher's utterance and its word transcription, the 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. We assume that the syllable-level time alignments and stress labels of the original teacher's utterance are available

since they can be obtained by the automatic phoneme alignment and stress detection techniques presented in Chapter 3.

The clapping-based teacher's utterance can be achieved in two steps, given the following information: a pre-recorded clapping sound and an original teacher's utterance, along with its corresponding word transcription, syllable-level time alignments and stress labels. The first step is to generate a *clapping-pattern sound* that has the same duration of the original teacher's utterance. Also, the *clapping-pattern sound* has a clapping at the time mark of every stressed syllable and keeps silence anywhere else. The second step is to resynthesise the *clapping-based teacher's utterance* by adding the teacher's utterance and its corresponding clapping-pattern sound in time domain.

An example of resynthesising a clapping-based teacher's utterance is given as follows: The utterance in this example is "The de*vice* is at*tached* to a *pla*stic *wrist*band." which is illustrated in Figure 6.1. From the first panel to the last panel of Figure 6.1, they are the waveform, pitch contour, ToBI label of the stressed syllables, phonetic transcription and word transcription. The stress labels are in ToBI (*Tones and Break Indices*) format (Silverman, et al., 1992). Refer to Section 3.4.1, for a detailed description of ToBI transcription. Figure 6.2 illustrates the process of resynthesising the clapping-based teacher's utterance of the original speech in Figure 6.1. Figure 6.2 (a) redisplays the waveform of the original teacher's utterance. Figure 6.2 (b) shows the waveform of the clapping-pattern sound of the original teacher's utterance. From Figure 6.2 (b), we can see that the time intervals of stressed syllables in this utterance are approximately equal (Yavas, 2006, p. 21). Thus, this utterance is a typical example to show that English is a stress-timed language. Model CPPA then resynthesises the clapping-based teacher's utterance by adding the waveforms of Figure 6.2 (a) and Figure 6.2 (b). The waveform of the clapping-based teacher's utterance is shown in Figure 6.2 (c).

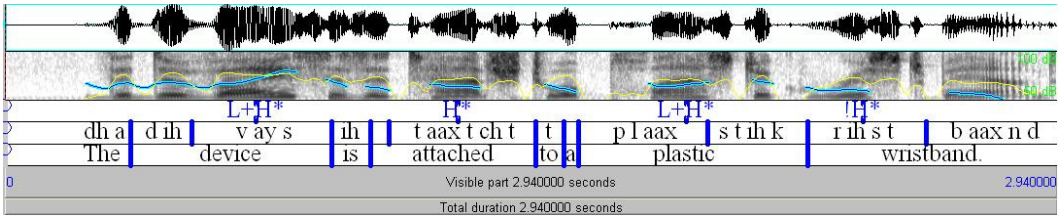


Figure 6.1 Illustration of the utterance
The utterance is "The de*vice* is at*tached* to a *pla*stic *wrist*band."

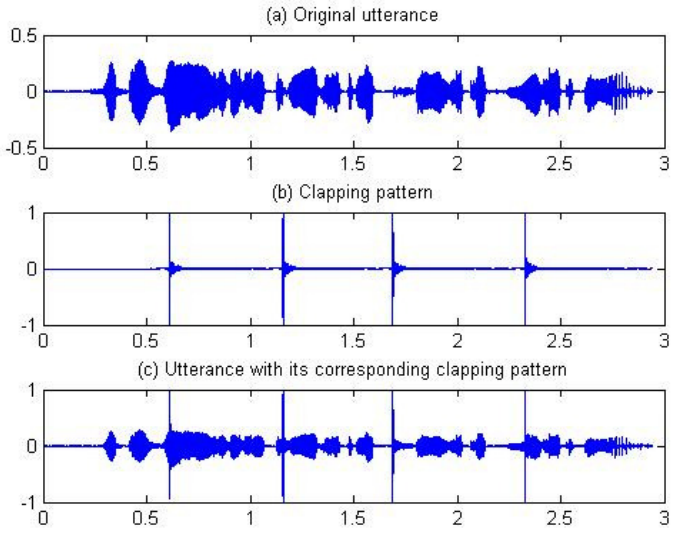


Figure 6.2 Resynthesis of clapping-based teacher's utterance.
(a) Waveform of the original teacher's utterance. (b) Waveform of the clapping-pattern sound of the original teacher's utterance. (c) Waveform of the resynthesised clapping-based teacher's utterance

In our CASTLE system, learners can practise uttering a sentence, while playing a teacher's clapping-based production of this sentence. By clapping at stressed syllables, CASTLE can help learners to become familiar with and get used to the rhythm of English language. By reinforcing the learning, learners will form a habit of putting more emphasis on stressed syllables and produce this sentence with correct stress patterns.

6.2 Representation of stress

6.2.1 A limitation of the categorical representation of stress

In linguistics, stress is conventionally treated as a categorical concept, in which a syllable is either stressed or unstressed. There is nothing in between. However, stress is a subjective concept. Given an utterance and asked 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 (Imoto, et al., 2000). Two reasons may contribute to the uncertainty of stress. On one hand, for some syllables, a labeler cannot be 100% sure their stress patterns. This kind of uncertainty can be seen as *intra-labeler*

uncertainty. On the other hand, some syllables, which are perceived as stressed by a labeler, can also be perceived as unstressed by another labeler. This kind of uncertainty can be seen as *inter-labeler* uncertainty.

There are empirical evidences of disagreement on stress labeling. For a variety of speaking styles, the reported pairwise inter-labeler agreements ranged between 81% and 91% (Wightman, 2002). The inter-labeler agreement on the presence versus absence of stress on a subset of the BU-RN corpus was reported to be 91% (Ostendorf, et al., 1995). As indicated by Deshmukh and Verma (2009), one of the main reasons for the inter-labeler disagreement was that the inter-labeler calibration was not very rigorous, in which some labelers were stricter with the stress errors than others.

The categorical representation of stress is insufficient to represent the subjective nature of stress. Taylor (2000) argues that it is pointless to define and try to find the strict boundaries of suprasegmentals (e.g. stress) that are underlying continuous phenomena. A vivid analogous given by Taylor (2000) is how people describe the temperature of an object. It is just a convenient way to describe the temperature of a certain object as *hot* or *cold*. It does not mean that people really think the temperature is categorical. It is worthless to find the boundary between hot and cold. Analogously, it does not make much sense to try to find the exact boundary between stressed and unstressed syllables.

Since the categorical representation of stress is incapable to describe the uncertainty or ambiguity of stress, we propose to extend the categorical representation of stress to a fuzzy representation.

6.2.2 A fuzzy representation of stress

As defined by Zadeh (1965), a Fuzzy Set A is a collection of objects that satisfy a certain (or several) property; each object x has a membership value $\mu_A(x) \in [0,1]$ of A, which demonstrates the probability of x belonging to A. μ_A is the membership function of A.

A fuzzy representation of stress is composed of a fuzzy set of stressed syllables and a fuzzy set of unstressed syllables. In the following, we use STRESSED to represent the fuzzy set of stressed syllables and UNSTRESSED to represent the fuzzy set of unstressed syllables. STRESSED 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 STRESSED is given in Eq. (6.1)

$$\mu_{stressed}(s) = f(s) \tag{6.1}$$

where $f(s) \in [0,1]$. The higher the membership value of a syllable belonging to STRESSED is, the more likely the syllable is stressed. The concept of unstressed is complementary to the concept of stressed. Thus, the fuzzy set UNSTRESSED can be defined as.

$$\mu_{unstressed}(s) = 1 - f(s) \tag{6.2}$$

The less the difference between the membership values $\mu_{stressed}(s)$ and $\mu_{unstressed}(s)$, the closer syllable *s* is to the boundary between stressed and unstressed syllables. The fuzzy representation of stress can be reduced to the categorical representation by setting *s* as stressed (or unstressed), when the membership value of $\mu_{stressed}(s)$ is greater (or less) than $\mu_{unstressed}(s)$.

The fuzzy representation of stress provides a new perspective of how to express the uncertainty of stress. The application of the fuzzy representations of stressed and unstressed syllables can be seen from the three stress-error feedback models described in the next section.

6.3 Fuzzy representation based stress-error feedback models

Based on the fuzzy representation of stress, three stress-error feedback models are presented in this section, which can be alternatively employed in the production assistance module of CASTLE to provide feedback for learners' stress errors.

The sentence stress differences between a teacher's utterance and learners' corresponding imitations are useful feedback to help learners realise and correct their stress errors. However, it would be possibly destructive to provide feedback for every

minor deviation since learners may be confused and quickly become discouraged (Menzel, et al., 2001). For example, a syllable in a teacher's utterance is slightly stressed, while in a learner's imitation, the corresponding syllable is, to some extent, unstressed, as it is illustrated in Figure 6.3, where the square and triangle indicate the membership values of the teacher's syllable and the learner's corresponding imitation belonging to the fuzzy set STRESSED. The stress difference between the teacher's production and the learner's production on this syllable is less significant since the membership values of the teacher's syllable and the learner's syllable are very close. If a computer-assisted stress learning system keeps on indicating these insignificant errors, this may decrease learners' learning interests and have a negative effect on learners' self-confidence (Menzel, et al., 2001), especially for novices.

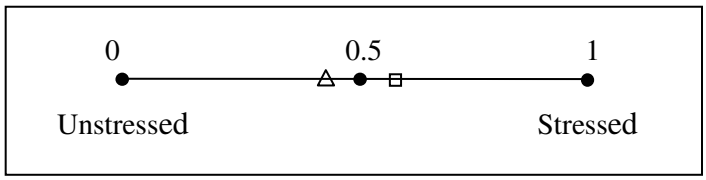


Figure 6.3 Stress difference between a teacher's syllable and a learner's imitation.

In order to protect and promote learners' learning interests and provide more useful feedback, CASTLE is intended only to provide feedback for learners' stress errors that have a more noticeable stress divergence from teachers' utterances. In order to avoid incorrect feedback, following the pronunciation training system Fluency (Eskenazi et al., 2000), CASTLE does not pass judgment on learners, and only indicates specific words or syllables for learners to work on.

A stress-error feedback model is to provide feedback for the stress errors in learners' imitations by post-processing the results of automatic stress detector(s) used in the stress-error feedback model. Based on the fuzzy representation of stress, a prototype of stress-error feedback model can be described by the input and output variables expressed in Table 6.1. Given the inputs U_t, $\mu_{stressed}(s_i)_t$, U_s, σ and D, a stress-error feedback model is to provide information for the incorrectly stressed syllables W.

Based on this prototype, three stress-error feedback models are presented in the sub-sections 6.3.1, 6.3.2 and 6.3.3, respectively, i.e. feedback model based on prediction

confidence (Feedback$_{PC}$), feedback model based on multiple stress detectors (Feedback$_{MC}$), and feedback model based on learners' more than one imitation (Feedback$_{DI}$). Feedback$_{PC}$ is intended to deal with the *intra-labeler* uncertainty. Feedback$_{MC}$ is developed to handle the *inter-labeler* uncertainty. Feedback$_{DO}$ aims to indicate learners' repeated stress errors.

Table 6.1 Inputs and output of the prototype of stress-error feedback model

	Variable Description
Inputs	U_t: a teacher's utterance
	$\mu_{stressed}(s_i)_t$: the membership value of syllable s_i in U_t belonging to the fuzzy set STRESSED, where s_i is the ith syllable of U_t; i=1, 2, …, n; and n is the number of syllables in U_t
	U_s: a learner's corresponding imitation of U_t
	σ : divergence threshold of membership values
	D: automatic stress detector(s)
Output	$W = \{s_i \mid \mid \mu_{stressed}(s_i)_t - \mu_{stressed}(s_i)_s \mid \geq \sigma\}$: syllables with stress errors, where $\mu_{stressed}(s_i)_s$ is the membership value of the ith syllable in U_s belonging to the fuzzy set STRESSED. $\mu_{stressed}(s_i)_s$ is automatically detected by D.

6.3.1 Model Feedback$_{PC}$

Feedback model Feedback$_{PC}$ is developed to provide feedback for stress errors, based on the prediction confidence of the automatic stress detector employed in model Feedback$_{PC}$. The automatic stress detection techniques in CASTLE have been introduced in Section 3.4. For each syllable, a stress detector predicts its stress label along with a prediction confidence. A prediction confidence is a numeric value ranging between 0.5 and 1. The higher the prediction confidence is, the more likely the syllable belongs to the predicted stress class.

Model Feedback$_{PC}$ is to handle the intra-labeler uncertainty that is caused by the uncertainty of an individual human labeler's labeling decision. The stress detector employed in model Feedback$_{PC}$ can be treated as a human labeler. And the prediction

confidence of a stress label obtained by the stress detector can be used to describe the degree of this syllable belonging to the fuzzy set STRESSED, which is equivalent to the confidence of the human labeler about the stress label labeled by himself/herself.

Based on the prediction confidence, a fuzzy representation of stress can be developed. Given syllable s_i in utterance U and automatic stress detector D, the stress label of s_i is predicted as L_{auto_i} by D at confidence level c_i that is automatically obtained by D. Then the membership value of s_i belonging to the fuzzy set STRESSED can be defined as:

$$\mu_{stressed}(s_i) = \begin{cases} c_i, & if\ L_{auto\_i}\ is\ stressed \\ 1 - c_i, & if\ L_{auto\_i}\ is\ unstressed \end{cases} \tag{6.3}$$

Since the concepts of stressed and unstressed are complementary to each other, the membership value of s_i belonging to the fuzzy set UNSTRESSED can be defined as:

$$\mu_{unstressed}(s_i) = \begin{cases} 1 - c_i, & if\ L_{auto\_i}\ is\ stressed \\ c_i, & if\ L_{auto\_i}\ is\ unstressed \end{cases} \tag{6.4}$$

Note that the prediction confidence of an automatic stress detector and the fuzzy representation of stress are different concepts although they may correlate with each other. A stress detector is a binary classifier that treats each syllable as stressed or unstressed. The prediction confidence of the stress detector is used to describe the degree of confidence about the predicted class. However, the fuzzy representation of stress considers that each syllable has a degree of being stressed. The fuzziness of stress contributes to the prediction confidence of a stress detector. Then prediction confidence can be used to approximately describe the fuzziness of stress.

Given two syllables s_i and s_j, according to the definitions of $\mu_{stressed}$ and $\mu_{unstressed}$ in Eqs (6.3) and (6.4), the membership value difference between $\mu_{stressed}(s_i)$ and $\mu_{stressed}(s_j)$ is equal to the difference between $\mu_{unstressed}(s_i)$ and $\mu_{unstressed}(s_j)$, as shown in (6.5).

$$\begin{aligned} &|\mu_{stressed}(s_i) - \mu_{stressed}(s_j)| = |\mu_{unstressed}(s_i) - \mu_{unstressed}(s_j)| \\ &\begin{cases} |c_i - c_j| & if\ L_{auto\_i}\ and\ L_{auto\_j}\ are\ same. \\ |1 - c_i - c_j| & if\ L_{auto\_i}\ and\ L_{auto\_j}\ are\ different. \end{cases} \end{aligned} \tag{6.5}$$

Thus, in the following we only consider the membership value difference of two syllables belonging to the fuzzy set STRESSED.

In order to provide more reliable and useful feedback information to ESL learners, Feedback$_{PC}$ only reports learners' the incorrectly stressed syllables that satisfies the criterion that divergence $d(s_i)$ between $\mu_{stressed}(s_i)_t$ and $\mu_{stressed}(s_i)_s$ is not less than threshold σ, i.e.

$$d(s_i) = \mid \mu_{stressed}(s_i)_t - \mu_{stressed}(s_i)_s \mid \geq \sigma \tag{6.6}$$

where $\mu_{stressed}(s_i)_t$ is the membership value of s_i in a teacher's utterance belonging to the fuzzy set STRESSED; and $\mu_{stressed}(s_i)_s$ is the membership value of s_i in a learner's imitation belonging to STRESSED.

Model Feedback$_{PC}$ is suitable to be employed to *provide a various levels* of feedback to learners. For example, a lower threshold σ can be used to provide strict feedback for advanced learners, while a higher threshold σ is more suitable for novices to get feedback for their serious stress errors.

Considering the high detection accuracy of Multiple Layer Perception (MLP) based stress detector (refer to Section 3.4), in our CASTLE system, model Feedback$_{PC}$ employs a MLP-based stress detector to automatically detect the stress labels of teachers' utterances and learners' imitations. Model Feedback$_{PC}$ provides feedback for learners' stress errors based on the criterion given in Eq.(6.6). Learners are also allowed to control divergence threshold σ, in order to provide feedback to different levels of stress errors (e.g. significant errors or less obvious errors).

6.3.2 Model Feedback$_{MC}$

Model Feedback$_{MC}$ is developed to deal with the inter-labeler uncertainty that is caused by disagreement among human labelers. A common way to deal with inter-labeler disagreement is majority voting. In the experiment conducted by Imoto (2000), sentence stress of English utterances was labeled by eleven native speakers, and the syllables were labeled as stressed if over eight labelers perceived them as stressed.

To simulate the majority voting strategy, model Feedback$_{MC}$ employs a number of stress detectors. The stress detectors vote to decide the stress labels for each syllable. The membership value of syllable s_i belonging to the fuzzy set STRESSED can be written as:

$$\mu_{stressed}(s_i) = p / m \tag{6.7}$$

where m is the number of stress detectors employed in Feedback$_{MC}$; and p is the number of stress detectors that detect s_i as a stressed syllable. Then the membership value of syllable s_i belonging to the fuzzy set UNSTRESSED is,

$$\mu_{unstressed}(s_i) = q / m \tag{6.8}$$

where q is the number of stress detectors that detect s_i as an unstressed syllable.

The majority voting strategy makes model Feedback$_{MC}$ insignificantly affected by a specific stress detector employed in Feedback$_{MC}$. Thus, model Feedback$_{MC}$ can be more reliable than feedback models that only depend on a single stress detector.

In our CASTLE system, model Feedback$_{MC}$ employs three stress detectors: an MLP-based stress detector, an LR-based (Linear Regression) stress detector, and an SMO-based (Sequential Minimal Optimization) stress detector. The membership value of a syllable belonging to the fuzzy set STRESSED is obtained by Eq. (6.7). Then, model Feedback$_{MC}$ can provide feedback for learners' incorrectly stressed syllables that are detected by the criterion given in Eq. (6.6).

The difference between model Feedback$_{MC}$ and a traditional voting method (e.g. equally weighted) can be described using the following example. Assume that three stress detectors are employed in both model Feedback$_{MC}$ and the traditional voting method, divergence threshold σ of model Feedback$_{MC}$ is set to 0.5. Given a teacher's utterance and a learner's corresponding imitation, a syllable with a stress error will be informed to learners by the traditional voting method, if the syllable (s_t) in the teacher's utterance is detected as stressed by two stress detectors and the corresponding syllable (s_s) produced in the learner's imitation is detected as stressed by only one stress detector, since s_t is classified as stressed by the traditional voting method and s_s is classified as unstressed. However, this syllable is not reported by model Feedback$_{MC}$ as a stress error, since the membership value difference between s_t and s_s is 1/3 which is less than σ.

$$|\mu_{stressed}(s_t) - \mu_{stressed}(s_s)| = 2/3 - 1/3 = 1/3 < \sigma \tag{6.9}$$

6.3.3 Model Feedback$_{DI}$

Given a learner's dual imitations of a teacher's utterance, feedback model Feedback$_{DI}$ is developed to identify learners' repeated stress errors. It is important that a CAPT system indicates learners' repeated stress errors. If an error repeatedly occurs in learner's multiple productions, then eigher (i) the learner fails to realise this stress error, or (ii) the learner realises this stress error, but have difficulties producing this syllable with the correct stress. Also, in a CAPT system, learner's multiple productions of a sentence are easy to obtain since the learner is practising his/her pronunciation. Thus, based on learner's a sequential imitations of a teacher's utterance, model Feedback$_{DI}$ is intended to indicate learner's recurring stress errors. We choose *two* as the number of learner's sequential imitations, because the more imitations in the sequence, the more the learner's stress errors are at risk of becoming fossilised, and also errors occurring in two productions are unlikely due to the learner's carelessness.

In our CASTLE system, model Feedback$_{DI}$ works as follows: given learner's dual imitations of a teacher's utterance, a stress detector (e.g. model Feedback$_{MC}$, model Feedback$_{PC}$) is used to find the stress errors in learner's dual imitations. If a syllable is detected as with stress error in the learner's both imitations, model Feedback$_{DI}$ then indicate this syllable for the learner to have more practice.

6.4 Flowchart of the production assistance module

The production assistance module of CASTLE system is composed of a clapping-based pronunciation practice assistance (CPPA) model and three alternative stress-error feedback models, which are presented in sections 6.1 and 6.4, respectively. Figure 6.4 illustrates the flowchart of the production assistance module.

Firstly, the production assistance module helps learners to become familiar with the rhythm of English language. Given a teacher's utterance and its corresponding word transcription, the CPPA model is used to generate the clapping-based teacher's utterance that is resynthesised by adding a clap to every stressed syllable of the original teacher's utterance.

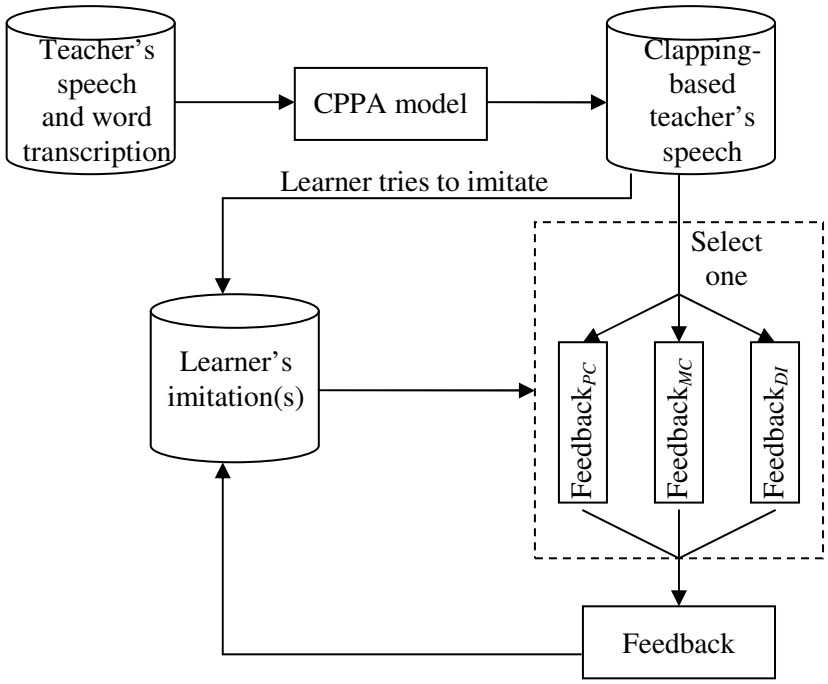


Figure 6.4 Flowchart of the production assistance module

Then the learners can listen to and imitate the clapping-based teacher's utterance. When they practice, they can also clap their hands with the clapping in the teacher's utterance. This aims to help the learners get used to the stress patterns of this utterance.

In order to correct learners' stress errors and improve their stress-using ability, the feedback model employed in the production assistance module can be used to provide feedback for the learners' inappropriately stressed or unstressed syllables.

In CASTLE, there are three alternative feedback models (i.e. Feedback$_{PC}$, Feedback$_{MC}$ and Feedback$_{DI}$) that can be selected and employed in the production assistance module. Learners can select one of the feedback models according to their own preference. In models Feedback$_{P}$ and Feedback$_{P}$, by setting up thresholds, they can also get feedback for stress errors that are on different levels. For example, models Feedback$_{PC}$ and Feedback$_{MC}$ with a high threshold will only provide learners with feedback for their sever stress errors, and a low threshold will make models Feedback$_{PC}$ and Feedback$_{MC}$ provide learners with feedback for their less significant stress errors. If model Feedback$_{DI}$ is selected, learners can get feedback for their repeated stress errors. The alternative feedback models are intended to satisfy learners' different needs. If learners

keep on practising their pronunciation with the feedback indicating how to correct their stress errors, their stress-using ability is expected to be improved gradually.

6.5 Summary

In this chapter, we have presented the production assistance module of CASTLE and how it works, which is developed to assist ESL learners to produce sentence stress correctly. The production assistance module includes a clapping-based pronunciation practice assistance model and three alternative stress-error feedback models (i.e. Feedback_{PC}, Feedback_{MC} and Feedback_{DI}).

The clapping-based pronunciation practice assistance model can be used to help ESL learners get a sense of the rhythm in English and guide them to produce utterances with appropriate English rhythm.

We have also proposed to use a fuzzy representation to describe stress, instead of a categorical representation that is conventionally used in linguistics. Compared with the categorical representation, the fuzzy representation of stress can better describe the subjective nature of stress (i.e. *intra-labeler* and *intra-labeler* uncertainties).

Based on the fuzzy representation of stress, three alternative stress-error feedback models have been presented, which are intended to help ESL learners to rectify their inappropriately stressed or unstressed words in their pronunciation. ESL learners can select one of these feedback models according to their preferences. They can also adjust the threshold parameter of their selected feedback models to get feedback for their significant or less significant stress errors.

The production assistance module is intended to help learners become familiar with the rhythm of English language and to establish a habit of stressing appropriate words or syllables in their speech. Then the learners' sentence stress production ability is expected to be improved.

Chapter 7.

An Enhanced Fuzzy Linear Regression Model

In addition to the development of CASTLE system, we have also conducted research in the field of fuzzy linear regression (Lu and Wang, 2009). One of the deficiencies of previous fuzzy linear regression models is that with the increase of the magnitudes of independent variables, the spreads of estimated fuzzy dependent variables are increasing, even though the spreads of observed dependent variables actually decrease or remain unchanged. In this chapter, we propose an enhanced fuzzy linear regression model (Lu and Wang, 2009) which can overcome the spreads increasing problem encountered by previous fuzzy linear regression models.

This chapter is organised as follows: Section 7.1 introduces Fuzzy Linear Regression (FLR). In Section 7.2, we provide a brief introduction to fuzzy numbers, and then describe the *spreads increasing problem* of previous fuzzy linear regression models in more detail. Related literature to solve the *spreads increasing problem* is reviewed in Section 7.3. In Section 7.4, an enhanced FLR model, FLR_{FS}, is proposed, which is able to model the linear relationship between the dependent and independent variables better than the previous models. Four numerical experiments are used to demonstrate the effectiveness of model FLR_{FS} in Section 7.5. Section 7.6 summarises this chapter by giving our conclusions and future work of fuzzy linear regression.

7.1 Fuzzy linear regression

Fuzzy linear regression (FLR) was first proposed by Tanaka et al. (1982) as an extension of the classical regression analysis, which is becoming a powerful tool to explore the vague relationship between dependent and independent variables (Coppi, 2008). In fuzzy regression, some elements of the regression models are represented by imprecise data.

General FLR models for crisp input-fuzzy output data (Tanaka, et al., 1982) and fuzzy input-fuzzy output data (Sakawa and Yano, 1992a) can be represented as follows, respectively:

$$\hat{\tilde{Y}}_i = \tilde{A}_0 + \tilde{A}_1 x_{i1} + \cdots + \tilde{A}_j x_{ij} + \cdots + \tilde{A}_m x_{im} \qquad (\text{FLR}_{CF})$$

$$\hat{\tilde{Y}}_i = \tilde{A}_0 + \tilde{A}_1 \tilde{X}_{i1} + \cdots + \tilde{A}_j \tilde{X}_{ij} + \cdots + \tilde{A}_m \tilde{X}_{im} \qquad (\text{FLR}_{FF})$$

where $\tilde{A}_j$ is the *j*th fuzzy regression coefficients; x_{ij} or $\tilde{X}_{ij}$ is the *j*th independent variable of the *i*th instance; x_{i0} ($\tilde{X}_{i0}$) is 1; $\hat{\tilde{Y}}_i$ is the *i*th estimated dependent variable; i=1, 2, …, n; and j=0, 1, …, m. A tilde character (~) is placed above the name of a fuzzy variable to distinguish a fuzzy variable from a crisp variable. As crisp numbers can be seen as special fuzzy numbers, model FLR_{CF} can be treated as a special case of model FLR_{FF}.

The methods to estimate the fuzzy regression coefficients can be roughly categorized into two groups. One is the linear programming (LP) methods (Nasrabadi, et al., 2005; Shakouri, et al., 2007; Tanaka, et al., 1982); and the other is the least-squares (LS) methods (Bargiela, et al., 2007; Coppi and D'Urso, 2003; Coppi, et al., 2006; Diamond, 1988; Modarres, et al., 2005; Yang and Lin, 2002). LP methods minimise the total spread of the estimated dependent variables or that of the fuzzy regression coefficients, subject to the constraint that the estimated dependent variables include the observed dependent variables within a certain *h*-level. The advantage of the LP methods is low computational complexity. However, the LP methods have been criticized by Redden and Woodall (1994) as (i) they are extremely sensitive to outliers (Hung and Yang, 2006); (ii) they do not allow all observations to contribute to the estimation; and (iii) the estimated intervals become wider as more data are collected. Multi-objective fuzzy regression techniques are developed to overcome the deficiencies of the LP methods (Nasrabadi, et al., 2005; Özelkan and Duckstein, 2000; Sakawa and Yano, 1992b; Tran and Duckstein, 2002). LS methods minimise the total difference between the estimated dependent variables and their observed counterparts. Thus, compared with the estimations of the LP methods, the estimations of the LS methods have relatively small differences between the estimated dependent variables and the observed ones. However, the LS methods have relatively higher computational complexity. A comprehensive literature review of fuzzy regression can be found in Kahraman, et al. (2006).

A problem of model FLR_{FF} is that with the increase of the magnitudes of independent variables, the spreads of estimated dependent variables are increasing, even though the spreads of observed dependent variables are roughly constant or decreasing (Chen and Dang, 2008; Kao and Chyu, 2002, 2003; Kao and Lin, 2005; Nasrabadi and Nasrabadi, 2004). We call it *spreads increasing problem* (refer to Section 7.2.3) in this chapter. Some models (Chen and Dang, 2008; Coppi, et al., 2006; D'Urso, 2003; Kao and Chyu, 2002, 2003; Kao and Lin, 2005; Nasrabadi and Nasrabadi, 2004) addressed this problem, and their deficiencies are briefly discussed below. More details are given in section 7.3.

FLR models presented by Chen and Dang (2008), Coppi, et al. (2006) and D'Urso (2003) can avoid the *spreads increasing problem* by modelling centres and spreads of dependent variables separately. However, the number of parameters to be estimated in model FLR_{CD08} (Chen and Dang, 2008) proportionally increases with the increase of the number of instances. Although more parameters involved in a regression model increase the model fitness, these also decrease the model generality (Kao and Lin, 2005). Therefore, model FLR_{CD08} is unsuitable for large dataset regression (refer to Section 7.3.4). In models $FLR_{D'Urso03}$ (D'Urso, 2003) and $FLR_{Coppi06}$ (Coppi, et al., 2006), the spreads of estimated dependent variables are only determined by the centres of the estimated dependent variables. This limits the ability of $FLR_{D'Urso03}$ and $FLR_{Coppi06}$ to model the spreads of the dependent variables by independent variable (refer to Section 7.3.3).

Although solutions proposed by Kao and Chyu (2002, 2003) and Nasrabadi and Nasrabadi (2004) also alleviate the *spreads increasing problem,* these solutions still cannot model a *decreasing* trend in the spreads of the observed dependent variables, as the magnitudes of the independent variables increase. For example, in these models proposed by Kao and Chyu (2002, 2003) and Nasrabadi and Nasrabadi (2004), if the independent variables are crisp, the spreads of the estimated dependent variables can only be a constant (refer to Section 7.3), even though the spreads of the observed dependent variables are decreasing with the increase of the magnitudes of the independent variables, as shown in Example 1.

Example 1. In Table 7.1, the independent variable is the height of the male candidates; and the fuzzy dependent variable measures how a candidate's height belongs to the concept *tall*. *L*-type fuzzy numbers in the form of (m_y, α_y) are used to describe *tall* (for a detailed description of *L*-type fuzzy number, refer to Section 7.2). m_y is the centre of a fuzzy number, which measures the possibility of a given candidate's height belonging to *tall*. In this example, m_y is not greater than 1. α_y is the spread of a fuzzy number, which describes the vagueness of m_y. The taller a candidate's height is, closer the possibility of the candidate's height is to 1, and lessens the vagueness of the candidate's height belonging to *tall*. However, it is difficult to model this relationship between the candidates' heights and *tall* by model FLR_{FF}, because of the *spreads increasing problem* in model FLR_{FF}, which is that the estimated dependent variables can only increase with the magnitudes of the independent variables increasing. Moreover, neither the models proposed by Kao and Chyu (2002, 2003) nor the model proposed by Nasrabadi and Nasrabadi (2004) can capture the relationship between height and *tall*, because in these models, when the independent variables are crisp, the spread of the estimated dependent variable can only be a constant (refer to Section 7.3), which is not true for dataset1 in Table 7.1.

Note that another problem of modeling the relationship between the candidates' heights and *tall* by FLR_{FF} is that the estimated spreads of *tall* may be negative, since the relationship between the candidates' heights and *tall* is not strictly linear. When the heights are greater than 2.1, the spreads of observed *tall* stop decreasing and the spreads of estimated *tall* are negative. Following the arguments of D'Urso (2003) and Coppi et al. (2006), negative predicted spreads can be interpreted as a lack of uncertainty and set to 0.

Table 7.1 Dataset1

i	Height (m)	Tall $(m_y, \alpha_y)_L$
1	1.7	$(0.60, 0.30)_L$
2	1.8	$(0.70, 0.25)_L$
3	1.9	$(0.80, 0.10)_L$
4	2.0	$(0.90, 0.05)_L$
5	2.1	$(1.00, 0.00)_L$

From the above, we can see that in the previous FLR models the increasing trend in the spreads of the estimated dependent variable limits the ability of FLR to model the

relationship between the dependent and independent variables. To alleviate this problem, in this chapter, we propose a flexible spreads FLR model (FLR_{FS}). In our model FLR_{FS}, the spreads of estimated dependent variables are able to fit the spreads of observed dependent variables, no matter if the spreads of the observed dependent variables are increased, decreased or unchanged, as the magnitudes and the spreads of the independent variables change.

7.2 Fuzzy number and the spreads increasing problem

In this section, we briefly introduce fuzzy numbers and the arithmetic rules of fuzzy numbers; then describe the *spreads increasing problem*.

7.2.1 Fuzzy number

The definition of fuzzy numbers given by Dubois and Prade (1980) is as follows.

Definition 7.2.1. A fuzzy number $\tilde{A}$ is a convex normalized fuzzy set of the real line **R**; its membership function $\mu_{\tilde{A}}(x)$ satisfies the following criteria:

i) α-cut set of $\tilde{A}$, $\mu_\alpha = \{x \mid \mu_{\tilde{A}}(x) \geq \alpha\}$, is a closed interval;

ii) $\mu_1 = \{x \mid \mu_{\tilde{A}}(x) = 1\}$ is nonempty;

iii) convexity: for $\lambda \in [0,1]$, $\mu_{\tilde{A}}(\lambda x_1 + (1-\lambda)x_2) \geq \min(\mu_{\tilde{A}}(x_1), \mu_{\tilde{A}}(x_2))$.

Definition 7.2.2. A *LR*-type fuzzy number $\tilde{A}$ is defined as follows (Coppi, et al., 2006; D'Urso, 2003; Zimmermann, 1991):

$$\mu_{\tilde{A}}(x) = \begin{cases} L(\dfrac{m_a - x}{\alpha_a}) & \text{for } x \leq m_a \\ R(\dfrac{x - m_a}{\beta_a}) & \text{for } x > m_a \end{cases}$$

where m_a is called centre or mean value; α_a and β_a are called left and right spreads respectively, $\alpha_a, \beta_a > 0$; L(z) and $R(z)$ are reference functions that map $\Re^+ \to [0,1]$, and strictly decreasing for $z \geq 0$. Also, L (or R) satisfies the following conditions: (i)

$L(0)=1$, $L(x)<1$ for $\forall x > 0$; (ii) $L(x)>0$ for $\forall x < 1$; (iii) $L(1)=0$, or [$L(x)>0$, $\forall x$ and $L(+\infty) = 0$]. $\tilde{A}$ can be denoted as $\tilde{A} = (m_a,\ \alpha_a,\ \beta_a)_{LR}$. If $\alpha_a = \beta_a$, then $\tilde{A}$ is symmetric, $\tilde{A} = (m_a,\ \alpha_a)_L$, which is called *L*-type fuzzy number.

Definition 7.2.3. If $L(x)=R(x)=1-x$, $\tilde{A}$ is a triangular fuzzy number [3]. Furthermore, if $\alpha_a = \beta_a$, then $\tilde{A}$ is a symmetric triangular fuzzy number.

7.2.2 Arithmetic operations on fuzzy numbers

By applying Zadeh's extension principle (Zadeh, 1965), the arithmetic operations of fuzzy numbers can be expressed as follows:

$$(\tilde{A} + \tilde{B})(z) = \sup_{x+y=z} T(\tilde{A}(x) + \tilde{B}(y))$$

$$(\tilde{A} * \tilde{B})(z) = \sup_{x*y=z} T(\tilde{A}(x) * \tilde{B}(y))$$

where $T(\cdot)$ is a triangular norm. The *T*-norm based *LR*-type fuzzy number addition preserves the shape. However, multiplication is not shape preserving, namely, the product of two *LR*-type fuzzy numbers may not be *LR*-type.

Dubois and Prade (1980) provided an approximation form for *LR*-type fuzzy number multiplication. According to their approximation formulas, the multiplication of two *LR*-type fuzzy numbers can be presented as follows:

i) if $\tilde{A} > 0$ and $\tilde{B} > 0$,

$$(m_a,\ \alpha_a,\ \beta_a)_{LR} \cdot (m_b,\ \alpha_b,\ \beta_b)_{LR} \approx (m_a m_b,\ m_a\alpha_b + m_b\alpha_a,\ m_a\beta_b + m_b\beta_a)_{LR}$$

ii) if $\tilde{A} < 0$ and $\tilde{B} > 0$,

$$(m_a,\ \alpha_a,\ \beta_a)_{LR} \cdot (m_b,\ \alpha_b,\ \beta_b)_{LR} \approx (m_a m_b,\ -m_a\beta_b + m_b\alpha_a,\ -m_a\alpha_b + m_b\beta_a)_{LR}$$

iii) if $\tilde{A} < 0$ and $\tilde{B} < 0$,

$$(m_a,\ \alpha_a,\ \beta_a)_{LR} \cdot (m_b,\ \alpha_b,\ \beta_b)_{LR} \approx (m_a m_b,\ -m_a\beta_b - m_b\beta_a,\ -m_a\alpha_b - m_b\alpha_a)_{LR}$$

[3] For easy explanation, we assume that all *LR*-type fuzzy numbers in this chapter are triangular fuzzy numbers.

7.2.3 Spreads increasing problem

For simplicity, most research considers $\tilde{A}_j$, $\tilde{X}_{ij}$ and $\hat{\tilde{Y}}_i$ in model FLR_{FF} as *LR*-type fuzzy numbers or triangular fuzzy numbers. By using the approximation formulas of Dubois and Prade (1980), Yang and Lin (2002) described model FLR_{FF} as:

$$\hat{\tilde{Y}}_i = \tilde{A}_0 + \tilde{A}_1\tilde{X}_{i1} + \cdots + \tilde{A}_m\tilde{X}_{im} \approx (m_{\hat{y}_i},\ \alpha_{\hat{y}_i},\ \beta_{\hat{y}_i})_{LR}$$

$$m_{\hat{y}_i} = m_{a_0} + \sum_{j=1}^{m} m_{a_j} m_{x_{ij}}$$

$$\alpha_{\hat{y}_i} = \alpha_{a_0} + \sum_{\tilde{A}_j>0,\, j=1}^{m} [s_{ij}(m_{a_j}\alpha_{x_{ij}} + m_{x_{ij}}\alpha_{a_j}) + (1-s_{ij})(m_{a_j}\alpha_{x_{ij}} - m_{x_{ij}}\beta_{a_j})] \tag{7.1}$$

$$+ \sum_{\tilde{A}_j<0,\, j=1}^{m} [s_{ij}(-m_{a_j}\beta_{x_{ij}} + m_{x_{ij}}\alpha_{a_j}) + (1-s_{ij})(-m_{a_j}\beta_{x_{ij}} - m_{x_{ij}}\beta_{a_j})]$$

$$\beta_{\hat{y}_i} = \beta_{a_0} + \sum_{\tilde{A}_j>0,\, j=1}^{m} [s_{ij}(m_{a_j}\beta_{x_{ij}} + m_{x_{ij}}\beta_{a_j}) + (1-s_{ij})(m_{a_j}\beta_{x_{ij}} - m_{x_{ij}}\alpha_{a_j})] \tag{7.2}$$

$$+ \sum_{\tilde{A}_j<0,\, j=1}^{m} [s_{ij}(-m_{a_j}\alpha_{x_{ij}} + m_{x_{ij}}\beta_{a_j}) + (1-s_{ij})(-m_{a_j}\alpha_{x_{ij}} - m_{x_{ij}}\alpha_{a_j})]$$

s_{ij}=1, if $\tilde{X}_{ij} \geq 0$; s_{ij}=0, if $\tilde{X}_{ij} < 0$.

From Eqs. (7.1) and (7.2), we can see that as the magnitudes of the independent variables (i.e. $|m_{x_{ij}}|$) increase, the spreads of the estimated dependent variables (i.e. $\alpha_{\hat{y}_i}$ and $\beta_{\hat{y}_i}$) increase. For example, when $\tilde{A}_j > 0$ and $\tilde{X}_{ij} > 0$, the left and right spreads of $\hat{\tilde{Y}}_i$ are

$$\alpha_{\hat{y}_i} = \alpha_{a_0} + \sum_{\tilde{A}_j>0,\, j=1}^{m} (m_{a_j}\alpha_{x_{ij}} + m_{x_{ij}}\alpha_{a_j}),$$

$$\beta_{\hat{y}_i} = \beta_{a_0} + \sum_{\tilde{A}_j>0,\, j=1}^{m} [s_{ij}(m_{a_j}\beta_{x_{ij}} + m_{x_{ij}}\beta_{a_j})$$

which increase with the increase of $|m_{x_{ij}}|$. Similarly, we can deduce that the spreads of $\hat{\tilde{Y}}_i$ increase as $|m_{x_{ij}}|$ increase, when $\tilde{A}_j > 0$ and $\tilde{X}_{ij} < 0$ ($\tilde{A}_j < 0$ and $\tilde{X}_{ij} < 0$; or $\tilde{A}_j < 0$ and $\tilde{X}_{ij} > 0$).

It is the inherent property of model FLR_{FF} that determines the spreads of $\tilde{\hat{Y}}_i$ increasing with the increase of $|m_{x_{ij}}|$. This property will affect the regression performance of model FLR_{FF}, when the spreads of the observed dependent variables are not increasing as the magnitudes of $\tilde{X}_{ij}$ increase. We name this property as *spreads increasing problem* of model FLR_{FF} in this chapter.

7.3 Review on related literature

The *spreads increasing problem* has been addressed in several papers (Chen and Dang, 2008; Coppi, et al., 2006; D'Urso, 2003; Kao and Chyu, 2002, 2003; Nasrabadi and Nasrabadi, 2004), and some solutions have been proposed. However, the previous solutions still have some deficiencies.

7.3.1 Model FLR_{KC02} and model FLR_{KC03}

Kao and Chyu (2002) proposed a crisp coefficients FLR model (FLR_{KC02}) to tackle the *spreads increasing problem*, which can be expressed as follows:

$$\tilde{\hat{Y}}_i = a_0 + a_1\tilde{X}_{i1} + \cdots + a_j\tilde{X}_{ij} + \cdots + a_m\tilde{X}_{im} + \tilde{\varepsilon} \qquad (FLR_{KC02})$$

$$\tilde{\varepsilon} = (0,\ l,\ r)_{LR}$$

where each coefficient a_j is a crisp number; j=0, 1, …, m; $\tilde{\varepsilon}$ is a triangular fuzzy error term; and $\tilde{X}_{ij} = (m_{x_{ij}},\ \alpha_{x_{ij}},\ \beta_{x_{ij}})_{LR}$. A two-stage methodology is proposed to obtain the crisp coefficients and the fuzzy error term. The first stage is to estimate crisp coefficients a_j by applying the classical least-squares method to the defuzzified (such as centroids) independent and dependent variables. In the second stage, fuzzy error term $\tilde{\varepsilon}$ is determined by minimising the total difference between the membership values of the estimated dependent variables and those of the observed dependent variables. Totally, there are m+3 parameters to be estimated in model FLR_{KC02}, which are l, r, and a_j's.

For crisp independent variables, a deficiency of model FLR_{KC02} is that the spreads of each estimated response variable are the spreads of $\tilde{\varepsilon}$, which are always constants. An example of model FLR_{KC02} for a single crisp independent variable is as follows:

$$\hat{\tilde{Y}}_i = 4.95 + 1.71x_i + (0,\ 3.01,\ 1.80)_{LR} \tag{7.3}$$

In the above model, the left and right spreads of all the estimated response variables are the spreads of $\tilde{\varepsilon}$, which is 3.01 and 1.80, respectively, even though the spreads of the observed response variables change as the independent variables change.

For fuzzy independent variables, a deficiency of model FLR_{KC02} is that the spread of each estimated response variable cannot be less than a constant. For instance, the left spread of the *i*th instance cannot be less than $a_1\alpha_{x_{i1}} + \cdots + a_m\alpha_{x_m}$. However, $a_1\alpha_{x_{i1}} + \cdots + a_m\alpha_{x_m}$ has no relationship with the left spread of the *i*th observed response variable. A numerical example of model FLR_{KC02} for a single fuzzy independent variable is as follows:

$$\hat{\tilde{Y}}_i = 3.5724 + 0.5193\tilde{X}_i + (0,\ 0.24,\ 0.24)_{LR} \tag{7.4}$$

The left spreads of the estimated responses are $0.5193\alpha_{x_i} + 0.24$. However, crisp coefficient 3.5724 and 0.5193 are determined by applying the classical least-squares method to the centroids of the independent and dependent variables, which have no relationship with the spreads of the observed dependent variables.

The model proposed by Kao and Chyu (2003) (FLR_{KC03}) has a similar form with model FLR_{KC02}. Except in model FLR_{KC03}, as shown in the following, centre *c* of error term $\tilde{\varepsilon}$ can be any crisp number, not only the origin.

$$\hat{\tilde{Y}}_i = a_0 + a_1\tilde{X}_{i1} + \cdots + a_j\tilde{X}_{ij} + \cdots + a_m\tilde{X}_{im} + \tilde{\varepsilon} \qquad (FLR_{KC03})$$
$$\tilde{\varepsilon} = (c, l, r)_{LR}$$

Model FLR_{KC03} is not able to avoid the deficiencies of model FLR_{KC02} either, which are described above.

7.3.2 Model FLR_{NN04}

The *spreads increasing problem* in model FLR_{FF} is caused by the fuzzy arithmetic rules. To avoid the *spreads increasing problem*, Nasrabadi and Nasrabadi (2004) defined new arithmetic operations for symmetric fuzzy numbers, and used these operations in fuzzy regression analysis. The arithmetic operations defined by Nasrabadi and Nasrabadi (2004) are as follows:

> For any *L*-type fuzzy numbers $\tilde{A} = (m_a,\ \alpha_a)_L, \tilde{B} = (m_b,\ \alpha_b)_L$, and an algebraic operation $\times$ on $\Re$, $\otimes$ is the corresponding algebraic operation of $\times$ on *L*-type fuzzy numbers. $\otimes$ is defined as: $\tilde{A} \otimes \tilde{B} = (m_a \times m_b,\ \alpha_a \times \alpha_b)_L$.

Based on the above definition of the arithmetic operations, FLR_{FF} can be written as follows:

$$\begin{aligned} \hat{\tilde{Y}}_i &= \tilde{A}_0 + \tilde{A}_1 \tilde{X}_{i1} + \cdots + \tilde{A}_m \tilde{X}_{im} \\ &= (m_{a0} + \sum_{p=1}^{m} m_{a_p} m_{x_{ip}},\ \alpha_{a0} + \sum_{p=1}^{m} \alpha_{a_p} \alpha_{x_{ip}})_L \end{aligned} \qquad (FLR_{NN04})$$

To some extent, model FLR_{NN04} can avoid the *spreads increasing problem*, because in model FLR_{NN04}, the spreads of estimated dependent variables have no relationship with the magnitudes of independent variables. However, a deficiency of model FLR_{NN04} is that the spreads of the estimated dependent variables can only depend on the spreads of the independent variables, because the spreads of the observed dependent variables may also depend on the magnitudes of the independent variables such as the example shown in Example 1.

7.3.3 Model $FLR_{D'Urso03}$ and model $FLR_{Coppi06}$

The models proposed by D'Urso (2003) and Coppi et al. (2006) are able to circumvent the *spreads increasing problem* by modelling the centres of dependent variables by classical regression methods, and meanwhile modelling the spreads of the dependent variables on their estimated centres.

For multiple independent variables $X_i = (\tilde{X}_{i1}, \cdots \tilde{X}_{ij}, \cdots \tilde{X}_{im})$ and single dependent variable $\tilde{Y}_i = (m_{y_i}, \alpha_{y_i}, \beta_{y_i})_{LR}$, estimated response $\hat{\tilde{Y}}_i = (m_{\hat{y}_i}, \alpha_{\hat{y}_i}, \beta_{\hat{y}_i})_{LR}$ obtained by the regression model proposed in D'Urso (2003) is as follows, in which $\tilde{X}_{ij} = (m_{x_{ij}}, \alpha_{x_{ij}}, \beta_{x_{ij}})_{LR}$:

$$m_{y_i} = m_{\hat{y}_i} + \varepsilon_i,\ m_{\hat{y}_i} = \mathrm{M}_{xi}\mathrm{a} + \mathrm{A}_{xi}\mathrm{r} + \mathrm{B}_{xi}\mathrm{s} \qquad (7.5) \qquad (\mathrm{FLR}_{D'Urso03})$$

$$\alpha_{y_i} = \alpha_{\hat{y}_i} + \lambda_i,\ \alpha_{\hat{y}_i} = m_{\hat{y}_i} b + d \qquad (7.6)$$

$$\beta_{y_i} = \beta_{\hat{y}_i} + \rho_i,\ \beta_{\hat{y}_i} = m_{\hat{y}_i} g + h \qquad (7.7)$$

where ε_i, λ_i and ρ_i are residuals; $\mathrm{M}_{xi} = (1, m_{x_{i1}} \cdots, m_{x_{ij}}, \cdots, m_{x_{im}})$; $\mathrm{A}_{xi} = (1, \alpha_{x_{i1}} \cdots, \alpha_{x_{ij}}, \cdots, \alpha_{x_{im}})$; $\mathrm{B}_{xi} = (1, \beta_{x_{i1}}, \cdots, \beta_{x_{ij}}, \cdots, \beta_{x_{im}})$; (m+1) dimension vectors a, r and s are the regression parameters for centres m_{y_i}; $\mathrm{a} = (\mathrm{a}_0, \mathrm{a}_1, \ldots, \mathrm{a}_m)^{\mathrm{T}}$, $\mathrm{r} = (\mathrm{r}_0, \mathrm{r}_1, \ldots \mathrm{r}_m)^{\mathrm{T}}$, $\mathrm{s} = (\mathrm{s}_0, \mathrm{s}_1, \ldots, \mathrm{s}_m)^{\mathrm{T}}$; i=1, 2, …, n; j=1, 2, …, m; and b and d (g and h) are the regression parameters to estimate left (right) spreads α_{y_i} (β_{y_i}). Model $\mathrm{FLR}_{D'Urso03}$ is based on three sub-models. The first one as shown in Eq.(7.5) estimates the centres of the dependent variables. The other two sub-models in Eqs. (7.6) and (7.7) model the left and right spreads of the dependent variables based on the estimated centres that are obtained in Eq. (7.5).

Model $\mathrm{FLR}_{D'Urso03}$ is able to avoid the *spreads increasing problem*, because of Eqs. (7.6) and (7.7), in which $\alpha_{\hat{y}_i}$ and $\beta_{\hat{y}_i}$ only depend on $m_{\hat{y}_i}$ that can increase or decrease with the increase of $\alpha_{x_{ij}}$'s and $\beta_{x_{ij}}$'s. For instance, assume that regression parameter r_k is negative and b is positive. Then, $m_{\hat{y}_i}$ decreases with the increase of $\alpha_{x_{ik}}$ (the left spread of the kth independent variable), and $\alpha_{\hat{y}_i}$ increase with the increase of $m_{\hat{y}_i}$. This makes $\alpha_{\hat{y}_i}$ decrease with the increase of $\alpha_{x_{ik}}$.

However, in model $\mathrm{FLR}_{D'Urso03}$, $\alpha_{\hat{y}_i}$ ($\beta_{\hat{y}_i}$) is determined by $m_{\hat{y}_i}$. That limits the ability of $\mathrm{FLR}_{D'Urso03}$ to model α_{y_i} (β_{y_i}) by independent variable X_i. For single independent variable, the three sub-models of $\mathrm{FLR}_{D'Urso03}$ can be rewritten as follows:

$$m_{\hat{y}_i} = (1, m_{x_i}) * (a_0, a_1)^T + (1, \alpha_{x_i}) * (r_0, r_1)^T + (1, \beta_{x_i}) * (s_0, s_1)^T \quad (7.8)$$

$$= a_1 m_{x_i} + r_1 \alpha_{x_i} + s_1 \beta_{x_i} + k$$

$$\alpha_{\hat{y}_i} = ((1, m_{x_i}) * (a_0, a_1)^T + (1, \alpha_{x_i}) * (r_0, r_1)^T + (1, \beta_{x_i}) * (s_0, s_1)^T) b + d \quad (7.9)$$

$$= a_1 b m_{x_i} + r_1 b \alpha_{x_i} + s_1 b \beta_{x_i} + kb + d$$

$$\beta_{\hat{y}_i} = ((1, m_{x_i}) * (a_0, a_1)^T + (1, \alpha_{x_i}) * (r_0, r_1)^T + (1, \beta_{x_i}) * (s_0, s_1)^T) g + h \quad (7.10)$$

$$= a_1 g m_{x_i} + r_1 g \alpha_{x_i} + s_1 g \beta_{x_i} + kg + h$$

where $k = a_0 + r_0 + s_0$. Assume that in a single independent variable dataset, m_{y_i}'s (the centres of the dependent variables) increase with the increase of α_{x_i}'s and β_{x_i}'s (the spreads of the independent variables); and α_{y_i}'s (the spreads of the dependent variables) increase with the increase α_{x_i}'s, but decrease when β_{x_i}'s increase. According to Eq. (7.8), $m_{\hat{y}_i}$ is able to describe the relationship between $m_{\hat{y}_i}$ and $\tilde{X}_i$ (i.e. $m_{\hat{y}_i}$ increases when α_{x_i} and β_{x_i} increase. This requires r_1 and s_1 to be positive.). However, $\alpha_{\hat{y}_i}$ is not able to describe the relationship between α_{y_i} and $\tilde{X}_i$ properly because when α_{y_i} increases with the increase of α_{x_i}, b needs to be positive, and meanwhile α_{y_i} decreases with the increase of β_{x_i} that requires b to be negative. Thus, in this case, model $FLR_{D'Urso03}$ is not able to properly describe the relationship between $\tilde{X}_i$ and $\tilde{Y}_i$.

Similarly, the model proposed by Coppi et al. (2006) is also composed of three sub-models. For crisp inputs $x_i = (x_{i1}, \cdots, x_{im})$ and *LR*-type fuzzy outputs $\tilde{Y}_i = (m_{y_i}, \alpha_{y_i}, \beta_{y_i})_{LR}$, estimated responses $\hat{\tilde{Y}}_i = (m_{\hat{y}_i}, \alpha_{\hat{y}_i}, \beta_{\hat{y}_i})_{LR}$ obtained by the regression model proposed in Coppi, et al. (2006) is:

$$m_{y_i} = m_{\hat{y}_i} + \varepsilon_i, \; m_{\hat{y}_i} = F(x_i)a \quad (7.11) \qquad (FLR_{Coppi06})$$

$$\alpha_{y_i} = \alpha_{\hat{y}_i} + \lambda_i, \; \alpha_{\hat{y}_i} = m_{\hat{y}_i} b + d \quad (7.12)$$

$$\beta_{y_i} = \beta_{\hat{y}_i} + \rho_i, \; \beta_{\hat{y}_i} = m_{\hat{y}_i} g + h \quad (7.13)$$

where $F(x_i) = [f_1(x_i), \ldots, f_k(x_i), \ldots, f_p(x_i)]$; f_k's are suitably chosen functions. For crisp input-fuzzy output, model $FLR_{D'Urso03}$ is a specification of model $FLR_{Coppi06}$, in which $F(x_i) = [1, x_{i1}, \cdots, x_{im}]$. Similar to model $FLR_{D'Urso03}$, the sub-models of $FLR_{Coppi06}$ shown in Eqs. (7.12) and (7.13) also depend on the sub-model given in Eq. (7.11). Thus,

$FLR_{Coppi06}$ has the same problem as $FLR_{D'Urso03}$, which is that α_{y_i} (β_{y_i}) cannot be linearly modelled by X_i freely.

7.3.4 Model FLR_{CD08}

To address the *spreads increasing problem*, a variable spread fuzzy linear regression model FLR_{CD08} is proposed by Chen and Dang (2008), which is a three-phase method.

In the first phase, regression coefficients are treated as fuzzy numbers and the membership functions of the least-squares estimates of the regression coefficients are constructed, since Chen and Dang argue that the membership functions of fuzzy sets are more capable of capturing the relationship between independent variables and dependent variables than crisp numbers (Chen and Dang, 2008). To avoid the *spreads increasing problem*, in the second phase, the fuzzy regression coefficients are defuzzified by the center of gravity method to crisp regression coefficients. In the third phase, for each instance, fuzzy error term $\tilde{E}_i$ is determined by a mathematical programming method. The objective function of the mathematical programming method is to minimise the total difference between the estimated and observed membership values of response variables, E_{KC} (refer Section 7.4.2), subject to the constraints that the spreads of each estimated response variable are equal to those of the observed response variable. For predicting the response of an unseen instance, a Mamdani fuzzy inference system (Zimmermann, 1991) is applied to the derived regression model.

A generic model of FLR_{CD08} is:

$$\hat{\tilde{Y}}_i = (b_0)_c + (b_1)_c \tilde{X}_{i1} + \cdots + (b_j)_c \tilde{X}_{ij} + \cdots + (b_m)_c \tilde{X}_{im} + \tilde{E}_i \qquad (7.14) \qquad (FLR_{CD08})$$

$$\tilde{E}_i = (0, \alpha_i, \beta_i)_{LR} \qquad (7.15)$$

where $(b_j)_c$'s are the defuzzified crisp regression coefficients; $\tilde{E}_i$ is the estimated error term of the ith instance; i=1, 2, …, n; and j=0, 1, …, m.

The parameters to be estimated in FLR_{CD08} are regression coefficients, $(b_j)_c$'s, and left and right spreads of error terms, α_i's and β_i's respectively. Totally, there are m+2n+1

parameters to be estimated in FLR_{CD08}, which are proportional to the number of instances, *n,* and the dimension of the dataset, *m*. For large datasets, *n* is usually significantly greater than *m*. More parameters involved in a regression model increase the model fitness, but these also decrease the model generality (Kao and Lin, 2005). Thus, FLR_{CD08} is unsuitable to large dataset regression, whose number of instances is large.

7.4 Flexible spreads FLR model FLR_{FS}

From the above sections, we have seen the shortcomings of the previous FLR models. In this section, we describe our flexible spreads FLR model (FLR_{FS}) that is able to overcome the *spreads increasing problem* and the deficiencies of the previous FLR models mentioned above.

7.4.1 Description of model FLR_{FS}

We first describe model FLR_{FS} for single regression, then extend it to multiple regression.

In fuzzy regression analysis, the spreads and magnitudes of independent variables are all the information that can be obtained. A general case is that the spreads of dependent variables may depend on both the spreads and magnitudes of the independent variables. Thus, a general FLR model should be able to allow: (i) the spreads of estimated dependent variables depend on both the spreads and magnitudes of the independent variables; (ii) the spreads of the estimated dependent variables can change freely (increase, decrease or fixed) as the spreads and magnitudes of the independent variables change. FLR_{FS} is a model that possesses these two properties.

Also, considering the relationship between fuzzy numbers and crisp numbers, in FLR_{FS} the centres of estimated dependent variables are modelled by the centres of independent variables, and the spreads of the estimated dependent variables are modelled by both the centres and spreads of the independent variables. Since the centre of a fuzzy number is

the element belonging to a fuzzy concept with 100%, it can be treated as a crisp number. Thus, in FLR_{FS} the estimation of the centres of dependent variables is based on classical linear regression. The spreads of fuzzy numbers can be treated as the vagueness of the fuzzy numbers. The vagueness of dependent variables depends not only on the vagueness of independent variables but also on the centres of the independent variables such as the dataset1 given in Table 7.1. To capture this relationship between the spreads of the dependent variables and the independent variables, in FLR_{FS} the spreads of the dependent variables are estimated by both the centres and spreads of the independent variables.

Model FLR_{FS} for single independent variable $\tilde{X}_i = (m_{x_i},\ \alpha_{x_i},\ \beta_{x_i})_{LR}$, can be described as follows:

$$\hat{\tilde{Y}}_i = k_0 + k_1 m_{x_i} + \tilde{S}_i, \qquad (FLR_{FS} \text{ single})$$

$$\tilde{S}_i = (0,\ \alpha_{s_i},\ \beta_{s_i})_{LR},$$

$$\alpha_{s_i} = k_{ll}\alpha_{x_i} + k_{lm} m_{x_i} + k_{lr}\beta_{x_i} + c_l,$$

$$\beta_{s_i} = k_{rl}\alpha_{x_i} + k_{rm} m_{x_i} + k_{rr}\beta_{x_i} + c_r,$$

$$\alpha_{s_i} \geq 0,\ \beta_{s_i} \geq 0,\ i=1, 2, \ldots, n.$$

where k_0 and k_1 are crisp regression coefficients; $\tilde{S}_i$ is a fuzzy spread term for the ith instance; and k_{ll}, k_{lm}, k_{lr}, k_{rl}, k_{rm}, k_{rr}, c_l and c_r are crisp spread coefficients. To achieve the fuzzy regression model, the parameters (i.e. k_0, k_1, k_{ll}, k_{lm}, k_{lr}, k_{rl}, k_{rm}, k_{rr}, c_l and c_r) need to be determined, subject to the constraints that the spread of $\tilde{S}_i$ should be non-negative. Parameters k_{ll} and k_{rl} reflect the influence of the left spreads of the independent variables on the left and right spreads of the dependent variables respectively. Similarly, parameters k_{lr} and k_{rr} show how the right spreads of the independent variables affect the left and right spreads of the dependent variables. Parameters k_{lm} and k_{rm} give the information of how the spreads of the dependent variables depend on the centres of the independent variables.

All parameters, k_0, k_1, k_{ll}, k_{lm}, k_{lr}, k_{rl}, k_{rm}, k_{rr}, c_l and c_r, can be zero, positive or negative. Thus, the spreads of the estimated dependent variables can increase or decrease freely as

the spreads and magnitudes of the independent variables change. Thus, the model FLR_{FS} is able to avoid the *spreads increasing problem*.

For *L*-type independent variable $\tilde{X}_i = (m_{x_i},\ \alpha_{x_i})_L$ and dependent variable $\tilde{Y}_i = (m_{y_i},\ \alpha_{y_i})_L$, a simplified FLR_{FS} model can be expressed as:

$$\hat{\tilde{Y}}_i = k_0 + k_1 m_{x_i} + \tilde{S}_i,$$

$$\tilde{S}_i = (0,\ \alpha_{s_i})_L,$$

$$\alpha_{s_i} = k_{ll}\alpha_{x_i} + k_{lm} m_{x_i} + c,$$

$$\alpha_{s_i} \geq 0,\ i=1, 2, \ldots, n.$$

A generalized model FLR_{FS} for multiple independent variables $\mathrm{X}_i = (\tilde{X}_{i1}, \cdots \tilde{X}_{ij}, \cdots, \tilde{X}_{im})$, can be described as follows:

$$\hat{\tilde{Y}}_i = k_0 + k_1 m_{x_{i1}} + \cdots + k_j m_{x_{ij}} + \cdots + k_m m_{x_{im}} + \tilde{S}_i, \quad (\text{FLR}_{FS}\ \text{multiple})$$

$$\tilde{S}_i = (0,\ \alpha_{s_i},\ \beta_{s_i})_{LR},$$

$$\alpha_{s_i} = \sum_{t=1}^{m} k_{llt}\alpha_{x_{it}} + \sum_{t=1}^{m} k_{lmt} m_{x_{it}} + \sum_{t=1}^{m} k_{lrt}\beta_{x_{it}} + c_l,$$

$$\beta_{s_i} = \sum_{t=1}^{m} k_{rlt}\alpha_{x_{it}} + \sum_{t=1}^{m} k_{rmt} m_{x_{it}} + \sum_{t=1}^{m} k_{rrt}\beta_{x_{it}} + c_r,$$

$$\alpha_{s_i} \geq 0,\ \beta_{s_i} \geq 0,\ i=1, 2, \ldots, n,\ j=0, 1, \ldots, m.$$

where $\tilde{X}_{ij} = (m_{x_{ij}},\ \alpha_{x_{ij}},\ \beta_{x_{ij}})_{LR}$; k_j's are crisp regression coefficients; $\tilde{S}_i$ is a fuzzy spread term for the *i*th instance; and k_{llt}'s, k_{lmt}'s, k_{lrt}'s, k_{rlt}'s, k_{rmt}'s, k_{rrt}'s, c_l and c_r are crisp spread coefficients. Model FLR_{FS} for multiple regression is a generalization of model FLR_{FS} for single regression.

The parameters to be estimated in model FLR_{FS} are k_j's, k_{llt}'s, k_{lmt}'s, k_{lrt}'s, k_{rlt}'s, k_{rmt}'s, k_{rrt}'s, c_l and c_r. The number of the parameters is $7m+3$, which is only proportional to the dimension of the dataset. Considering that more parameters in a model decrease the model generality, FLR_{FS} is thus more suitable to low dimension dataset regression.

As we can see from above, model FLR_{FS} can be easily extended from single regression to multiple regression. However, the computational complexity of model FLR_{FS} will increase significantly with the increase of data size and the dimension of the independent variables. A gradient-descent optimization strategy proposed by Bargiela, et al. (2007) deals with the high-dimensional data linear regression.

Model FLR_{FS} can be used for descriptive purposes to study the fuzzy relationship between dependent and independent variables. Also, it can be used for prediction purposes. Based on the arguments of D'Urso (2003) and Coppi et al. (2006), if non-positive predicted spreads are interpreted as a lack of uncertainty and can be set to 0 for practical purposes, model FLR_{FS} can then be used for prediction purposes.

Although model FLR_{FS} and model FLR_{KC02} have a similar form, they have several significant differences. (i) In model FLR_{FS}, fuzzy spread variable $\widetilde{S}_i$ is different for each instance, which is determined by the centres and spreads of independent variables $\widetilde{X}_i$. In model FLR_{KC02}, $\tilde{\varepsilon}$ is an error term that is fixed for all instances. (ii) In model FLR_{FS}, regression coefficients, k_1 and k_2, describe the relationship between m_{y_i} (the centres of the dependent variables) and m_{x_i} (the centres of the independent variables). In model FLR_{KC02}, the regression coefficients are obtained from modelling the relationship between the centroids of $\widetilde{Y}_i$ and $\widetilde{X}_i$; but they are used to describe the relationship between $\widetilde{Y}_i$ and $\widetilde{X}_i$.

The similarity between models $FLR_{D'Urso03}$, $FLR_{Coppi06}$ and FLR_{FS} is that all of them model the centres and spreads of dependent variables separately. There are also two differences between models $FLR_{D'Urso03}$, $FLR_{Coppi06}$ and FLR_{FS}. (i) In $FLR_{D'Urso03}$ and $FLR_{Coppi06}$, the centres of dependent variables are modelled by both the centres and spreads of independent variables. As mentioned above, the centre of a fuzzy number can be treated as a crisp number. Therefore, in FLR_{FS} the centres of dependent variables are determined by classical linear regression. (ii) In $FLR_{D'Urso03}$ and $FLR_{Coppi06}$, the spreads of dependent variables are modelled by their corresponding estimated centres. In FLR_{FS}, the spreads of estimated responses variable are able to depend on both the centres and spreads of independent variables linearly.

7.4.2 Property of model FLR_{FS}

Property of FLR_{FS}. The feasible region of model FLR_{FS} contains the feasible regions of models FLR_{KC02}, FLR_{KC03} and FLR_{NN04}. Thus, models FLR_{KC02}, FLR_{KC03} and FLR_{NN04} can be seen as special cases of model FLR_{FS}.

Proof. For simplicity, we only proof the above property of FLR_{FS} for single regression. For multiple regression, the property can be proofed in a similar way.

Model FLR_{FS} for single regression can be rewritten as:

$$\hat{\tilde{Y}}_i = (k_0 + k_1 m_{x_i},\ k_{ll}\alpha_{x_i} + k_{lm} m_{x_i} + k_{lr}\beta_{x_i} + c_l,\ k_{rl}\alpha_{x_i} + k_{rm} m_{x_i} + k_{rr}\beta_{x_i} + c_r)_{LR} \qquad (7.16)$$

(a) For single regression, both model FLR_{KC02} and model FLR_{KC03} can be written as:

$$\hat{\tilde{Y}}_i = a_0 + a_1 \tilde{X}_i + \tilde{\varepsilon} = a_0 + a_1 (m_{x_i},\ \alpha_{x_i},\ \beta_{x_i})_{LR} + (m_\varepsilon,\ \alpha_\varepsilon,\ \beta_\varepsilon)_{LR}$$

$$= \begin{cases} (a_0 + a_1 m_{x_i} + m_\varepsilon,\ a_1\alpha_{x_i} + \alpha_\varepsilon,\ a_1\beta_{x_i} + \beta_\varepsilon)_{LR} & if \quad a_1 \geq 0 \qquad (7.17) \\ (a_0 + a_1 m_{x_i} + m_\varepsilon,\ a_1\beta_{x_i} + \alpha_\varepsilon,\ a_1\alpha_{x_i} + \beta_\varepsilon)_{LR} & if \quad a_1 < 0 \qquad (7.18) \end{cases}$$

The feasible region of model FLR_{FS} contains the feasible regions of models FLR_{KC02}, FLR_{KC03}, since for any solution of models FLR_{KC02} or FLR_{KC03} $(a_0, a_1, m_\varepsilon, \alpha_\varepsilon, \beta_\varepsilon)$, we can always find an equivalent solution in model FLR_{FS} by comparing the coefficients of Eq. (7.16) with those of Eq. (7.17), and the coefficients of Eq. (7.16) and those of Eq. (7.18):

for $a_1 \geq 0$:

$$\begin{aligned} &k_0 = a_0 + m_\varepsilon, \quad k_1 = k_{ll} = k_{rr} = a_1, \\ &k_{lm} = k_{lr} = k_{rm} = k_{rl} = 0, \quad c_l = \alpha_\varepsilon, \quad c_r = \beta_\varepsilon, \end{aligned} \qquad (7.19)$$

for $a_1 < 0$:

$$\begin{aligned} &k_0 = a_0 + m_\varepsilon, \quad k_1 = k_{lr} = k_{rl} = a_1, \\ &k_{lm} = k_{ll} = k_{rm} = k_{rr} = 0, \quad c_l = \alpha_\varepsilon, \quad c_r = \beta_\varepsilon, \end{aligned} \qquad (7.20)$$

(b) In Nasrabadi and Nasrabadi (2004), model FLR_{NN04} is only considered with *L*-type fuzzy numbers. For a single independent variable, $\tilde{X}_i = (m_{x_i},\ \alpha_{x_i})_L$, model FLR_{NN04} can be expressed as:

$$\begin{aligned}\hat{\tilde{Y}}_i &= \tilde{A}_0 + \tilde{A}_1\tilde{X}_i = (m_{a_0},\ \alpha_{a_0})_L + (m_{a_1},\ \alpha_{a_1})_L(m_{x_i},\ \alpha_{x_i})_L \\ &= (m_{a_0} + m_{a_1}m_{x_i},\ \alpha_{a_0} + \alpha_{a_1}\alpha_{x_i})_L\end{aligned} \tag{7.21}$$

When all fuzzy numbers are *L*-type, model FLR_{FS} can be simplified as:

$$\hat{\tilde{Y}}_i = (k_0 + k_1 m_{x_i},\ k_{ll}\alpha_{x_i} + k_{lm}m_{x_i} + c)_L \tag{7.22}$$

The feasible region of model FLR_{FS} contains the feasible region of model FLR_{NN04}, since for any solution of model FLR_{NN04} $(m_{a_0},\ \alpha_{a_0},\ m_{a_1},\ \alpha_{a_1})$, we can always find an equivalent solution in model FLR_{FS} by comparing the coefficients of Eq. (7.21) and those of Eq. (7.22):

$$k_0 = m_{a_0},\quad k_1 = m_{a_1},\quad k_{ll} = \alpha_{a_1},\quad k_{lm} = 0,\quad c = \alpha_{a_0} \tag{7.23}$$

However, not all solutions of model FLR_{FS} can have an equivalent solution in models $\text{FLR}_{KC02,}$ FLR_{KC03} or FLR_{NN04}. For instance, no solutions of model $\text{FLR}_{KC02,}$ FLR_{KC03} or FLR_{NN04} are equivalent to the solutions of model FLR_{FS}, when the spreads of the dependent variables are determined or partially determined by the centres of the independent variables.

Thus, for single regression, the feasible region of model FLR_{FS} contains the feasible region of models FLR_{KC02}, FLR_{KC03} and FLR_{NN04}. Also, the conversion formulas from the parameters of solutions of model FLR_{KC02}, FLR_{KC03} and FLR_{NN04} to the parameters of solutions of model FLR_{FS} are given in Eqs. (7.19), (7.20) and (7.23). For single regression, models FLR_{KC02}, FLR_{KC03} and FLR_{NN04} can be seen as the special cases of model FLR_{FS}. □

7.4.3 Parameters estimation

When model FLR_{FS} is adopted, the following task is to estimate the parameters. Minimising the total difference between estimated and observed dependent variables is a common criterion for parameters estimation. Various distance measurements have

been proposed to measure the total difference between the estimated and observed dependent variables in fuzzy regression.

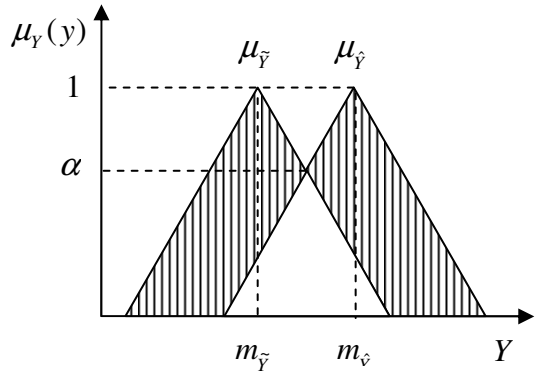


Figure 7.1 Membership functions of the estimated and observed fuzzy numbers.

In Kim and Bishu (1998), the error of estimation, E_{KB}, is defined as the ratio of the total difference between the estimated and observed membership values of response variables to the total observed membership values of the response variables, which is the shaded areas over the left triangle area, in Figure 7.1. A formularized definition of E_{KB} is given in Eq.(7.24)

$$E_{KB} = \frac{\int_{S_{\tilde{Y}} \cup S_{\hat{\tilde{Y}}}} | \mu_{\tilde{Y}}(x) - \mu_{\hat{\tilde{Y}}}(x) | dx}{\int_{S_{\tilde{Y}}} \mu_{\tilde{Y}}(x) dx} \tag{7.24}$$

where $\mu_{\tilde{Y}}(x)$ and $\mu_{\hat{\tilde{Y}}}(x)$ are the observed and estimated membership functions of the response variables; and $S_{\tilde{Y}}$ and $S_{\hat{\tilde{Y}}}$ are the supports of $\mu_{\tilde{Y}}(x)$ and $\mu_{\hat{\tilde{Y}}}(x)$.

E_{KC}, a variation of E_{KB}, was used by Kao and Chyu (2002, 2003). E_{KC} measures the total difference between the estimated and observed membership values of response variables, which include all the shaded areas in Figure 7.1.

$$E_{KC} = \int_{S_{\tilde{Y}} \cup S_{\hat{\tilde{Y}}}} | \mu_{\tilde{Y}}(x) - \mu_{\hat{\tilde{Y}}}(x) | dx \tag{7.25}$$

The similarity of fuzzy numbers is used as a criterion to evaluate the effectiveness of regression by Hojati, et al. (2005), which is defined as follows:

$$S_H = \frac{\int \min(\mu_{\tilde{Y}}(x), \mu_{\hat{\tilde{Y}}}(x)) dx}{\int \max(\mu_{\tilde{Y}}(x), \mu_{\hat{\tilde{Y}}}(x)) dx} \tag{7.26}$$

In D'Urso (2003), the squared Euclidean distance between two fuzzy numbers $\tilde{A}_1 = (m_1,\ \alpha_1,\ \beta_1)_{LR}$ and $\tilde{A}_2 = (m_2,\ \alpha_2,\ \beta_2)_{LR}$ is defined as:

$$d^2(\tilde{A}_1, \tilde{A}_2) = \| m_1 - m_2 \|^2\ \pi_c + \| (m_1 - \alpha_1) - (m_2 - \alpha_2) \|^2\ \pi_\alpha + \| (m_1 + \beta_1) - (m_2 + \beta_2) \|^2\ \pi_\beta$$

where π_c, π_α and π_β are arbitrary positive weights.

A generalized squared Euclidean distance is used by Coppi, et al. (2006), which can be described as:

$$\Delta^2(\tilde{A}_1, \tilde{A}_2) = \| m_1 - m_2 \|^2 + \| (m_1 - \lambda\alpha_1) - (m_2 - \lambda\alpha_2) \|^2 + \| (m_1 + \rho\beta_1) - (m_2 + \rho\beta_2) \|^2$$

where $\lambda = \int_0^1 L^{-1}(\omega)d\omega$ and $\rho = \int_0^1 R^{-1}(\omega)d\omega$. From the definition of Δ^2, we can see that Δ^2 weights the centres and spreads differently by means of λ and ρ.

E_{KB}, E_{KC}, d^2 and Δ^2 range from zero to positive infinity, while S_H ranges from 0 to 1. Thus, compared with E_{KB}, E_{KC}, d^2 and Δ^2, S_H can better describe the total difference between the estimated and observed response variables. Therefore, to estimate the parameters of model FLR_{FS}, our objective function is set to maximise the average similarity between the estimated and observed response variables. This is referred to as *MaxSim* solution for FLR_{FS}, which can be described as:

$$\text{Max} \quad \frac{1}{n}\sum_{i=1}^{n} S_{H_i} \tag{7.27}$$

$$\text{s.t.} \quad S_H = \frac{\int \min(\mu_{\tilde{Y}}(x), \mu_{\hat{\tilde{Y}}}(x))dx}{\int \max(\mu_{\tilde{Y}}(x), \mu_{\hat{\tilde{Y}}}(x))dx},$$

$$\hat{\tilde{Y}}_i = k_0 + k_1 m_{x_{i1}} + \cdots + k_j m_{x_{ij}} + \cdots + k_m m_{x_{im}} + \tilde{S}_i,$$

$$\tilde{S}_i = (0,\ \alpha_{s_i},\ \beta_{s_i})_{LR},$$

$$\alpha_{s_i} = \sum_{t=1}^{m} k_{llt}\alpha_{x_{it}} + \sum_{t=1}^{m} k_{lmt} m_{x_{it}} + \sum_{t=1}^{m} k_{lrt}\beta_{x_{it}} + c_l,$$

$$\beta_{s_i} = \sum_{t=1}^{m} k_{rlt}\alpha_{x_{it}} + \sum_{t=1}^{m} k_{rmt} m_{x_{it}} + \sum_{t=1}^{m} k_{rrt}\beta_{x_{it}} + c_r,$$

$$\alpha_{s_i} \geq 0,\ \beta_{s_i} \geq 0,\ i=1, 2, \ldots, n,\ j= 0, 1, \ldots, m.$$

7.5. Numerical examples

Note that the solution of the optimization problem *MaxSim* for model FLR_{FS} depends on the initial values, since the feasible region of solutions may not be continuous. In this section, we first provide a strategy of setting initial values, which is adopted in our experiments. Then, the effectiveness of model FLR_{FS} will be demonstrated on four datasets: dataset1 that has been shown in Table 7.1 in Section 7.1, and other three commonly used datasets (one is a single crisp input-fuzzy output dataset (dataset2) from (Tanaka, et al., 1989); another is a single fuzzy input-fuzzy output dataset (dataset3) from (Sakawa and Yano, 1992b); and the other is a multiple fuzzy inputs-fuzzy output real world dataset (dataset4) from (D'Urso, 2003)).

7.5.1 Initial values setting

For practical reasons, we provide an initial value setting strategy. The experimental results in Section 7.5.2 are based on this strategy. The purpose to provide this initial value setting strategy here is neither to demonstrate it is the best strategy to set initial values nor to claim that it guarantees to achieve the global optimisation.

For simplicity, we only introduce the initial value setting strategy for single regression FLR_{FS}. For model FLR_{FS} dealing with multiple regression, the initial values can be set in a similar way.

Given observations $(\tilde{X}_i, \tilde{Y}_i)$, where $\tilde{X}_i = (m_{x_i}, \alpha_{x_i}, \beta_{x_i})_{LR}$; $\tilde{Y}_i = (m_{y_i}, \alpha_{y_i}, \beta_{y_i})_{LR}$; and i=1, 2,…, n, the task of fuzzy regression is to find the parameters of model FLR_{FS}, which maximises Eq. (7.27), subject to its constraints.

In model FLR_{FS}, line $k_0 + k_1 m_{x_i}$ describes the relationship between m_{x_i} (the centre of independent variable) and m_{y_i} (the centre of dependent variable). Thus, we apply the conventional least-squares estimation to get the linear relationship between m_{x_i} and m_{y_i}, i.e. $m_{y_i} = b_0 + b_1 m_{x_i}$. b_0 and b_1 can be set as the initial values of k_0 and k_1.

In model FLR_{FS}, there are three factors that can affect the spreads of response variables: α_{x_i}, β_{x_i} and m_{x_i}. k_{ll} describes how α_{y_i} (the left spread of dependent variable) depends on α_{x_i} (the left spread of independent variable). Then, the conventional least-squares estimation is applied to α_{y_i} and α_{x_i} to get their linear relationship: $\alpha_{y_i} = p_{ll} + b_{ll}\alpha_{x_i}$. If equal weights is set to the three factors, α_{x_i}, β_{x_i} and m_{x_i}, $b_{ll}/3$ can be set as the initial value of k_{ll}. Similarly, k_{lr} describes how α_{y_i} depends on β_{x_i}. The least-squares estimation is applied to α_{y_i} and β_{x_i} to get their relationship: $\alpha_{y_i} = p_{lr} + b_{lr}\beta_{x_i}$. $b_{lr}/3$ can be set as the initial value of k_{lr}. Parameter k_{lm} describes how α_{y_i} depends on m_{x_i}. By applying the least-squares estimation to α_{y_i} and m_{x_i}, their relationship $\alpha_{y_i} = p_{lm} + b_{lm}m_{x_i}$ is obtained. The initial value of k_{lm} can be set as $b_{lm}/3$. ($p_{ll}/3 + p_{lr}/3 + p_{lm}/3$) can be set as the initial value of c_l. Unequal weights can also be assigned to the three factors α_{x_i}, β_{x_i} and m_{x_i}. For example, if the effect of m_{x_i} is significantly greater than the effects of both α_{x_i} and β_{x_i} (i.e. $b_{lm} >> b_{ll}$ and b_{lr}), the initial spreads of the estimated response variables can be set only based on m_{x_i}.

Similarly, we can set the initial values of k_{rl}, k_{rm}, k_{rr} and c_r in the same way.

7.5.2 Examples

The following experimental results are based on the initial value setting strategy described in Section 7.5.1.

Example 1 (continue)**.** As it is shown in Section 7.1, dataset1 cannot be modelled by FLR_{FF}, FLR_{KC02}, FLR_{KC03} and FLR_{NN04} properly, because the spreads of the observed response variables tend to decrease as the magnitudes of the independent variables increase. This decreasing trend in the spreads of the observed response variables can be fitted by model FLR_{FS}. Applying *L*-type fuzzy numbers based FLR_{FS} model and the initial value setting strategy given in Section 7.5.1, the following regression model is obtained.

$$\hat{\tilde{Y}}_i = -1.10 + m_{x_i} + (0,\ -0.5m_{x_i} + 1.05)_L \tag{7.28}$$

It can be seen from Eq. (7.28) that the spreads of the estimated response variables are decreasing with the increase of m_{xi}'s. The estimated response variables for the five instances are $(0.60, 0.20)_L$, $(0.70, 0.15)_L$, $(0.80, 0.10)_L$, $(0.90, 0.05)_L$ and $(1.00, 0.00)_L$, respectively.

Note that when m_x is greater than 2.10, the centre of the estimated dependent variable is greater than 1.00 and the spreads of the estimated dependent variable are less than 0, which can be interpreted as the height belongs to the concept *tall* with full confidence and a lack of uncertainty. For practical reasons, the estimated centres that are greater than 1.00 can be set to 1.00, and the estimated spreads that are less than 0 can be set to 0.

Example 2. In this example, we consider the crisp input-fuzzy output dataset given by Tanaka et al. (1989), which is shown in the left half of Table 7.2. x_i is the observed independent variable, $\tilde{Y}_i$ is the observed dependent variable, i=1, 2, …, 5.

Table 7.2 Dataset2

i	x_i	$\tilde{Y}_i = (m_{y_i},\ \alpha_{y_i})_L$	$\hat{\tilde{Y}}_i = (m_{\hat{y}_i},\ \alpha_{\hat{y}_i})_L$	S_H	E_{KC}
1	1	$(8.0, 1.8)_L$	$(6.00, 2.80)_L$	0.19	3.13
2	2	$(6.4, 2.2)_L$	$(7.75, 2.70)_L$	0.36	2.33
3	3	$(9.5, 2.6)_L$	$(9.50, 2.60)_L$	1.00	0.00
4	4	$(13.5, 2.6)_L$	$(11.25, 2.50)_L$	0.19	3.51
5	5	$(13.0, 2.4)_L$	$(13.00, 2.40)_L$	1.00	0.00
Average				0.5462	1.7932

As the observed dependent variables in dataset2 are symmetric, *L*-type fuzzy numbers based FLR_{FS} model is adopted. According to the initial value setting strategy given in Section 7.5.1, the initial values are set as:

Initial values (a): k_0=4.95; k_1=1.71; k_{ll}=0; k_{lm}=0.08; c=2.08

Then, the following regression model is obtained,

$$\hat{\tilde{Y}}_i = (4.25 + 1.75m_{x_i},\ -0.10m_{x_i} + 2.90)_L \tag{7.29}$$

Estimated response variable $\hat{\tilde{Y}}_i$, S_H and E_{KC} for each instance of dataset2 are listed in the right half of Table 7.2.

Since model FLR_{FS} is proposed to solve the *spreads increasing problem*, in this example model FLR_{FS} is compared with models FLR_{KC02} (Kao and Chyu, 2002), FLR_{KC03} (Kao and Chyu, 2003), FLR_{NN04} (Nasrabadi and Nasrabadi, 2004) and FLR_{CD08} (Chen and Dang, 2008), which are able to avoid the *spreads increasing problem* and also are provided experimental results on dataset2 by their authors. These five models are listed in Table 7.3[4], where $\tilde{E}_i$'s in model FLR_{CD08} are $\tilde{E}_1 = (-1.8, 0, 1.8)$, $\tilde{E}_2 = (-2.6, 0, 1.8)$, $\tilde{E}_3 = (-3.4, 0, 1.8)$, $\tilde{E}_4 = (-1.8, 0, 3.4)$ and $\tilde{E}_5 = (-3, 0, 1.8)$.

Table 7.3 Fuzzy regression models of dataset2

FLR_{FS}(MaxSim) Given initial values (a)	$\hat{\tilde{Y}}_i = (4.25 + 1.75m_{x_i},\ -0.10m_{x_i} + 2.90)_L$
FLR_{KC02} (Kao and Chyu, 2002)	$\hat{\tilde{Y}}_i = (4.95 + 1.71m_{x_i},\ 3.01,\ 1.80)_{LR}$
FLR_{KC03} (Kao and Chyu, 2003)	$\hat{\tilde{Y}}_i = (4.926 + 1.718m_{x_i},\ 2.320)_L$
FLR_{NN04} (Nasrabadi and Nasrabadi, 2004)	$\hat{\tilde{Y}}_i = (4.6812 + 1.73306m_{x_i},\ 2.3221)_L$
FLR_{CD08} (Chen and Dang, 2008)	$\hat{\tilde{Y}}_i = 4.95 + 1.71m_{x_i} + \tilde{E}_i$

It is worth to note that the solution of a model also depends on the objective function to be optimized. Thus, it is not easy to compare the solutions of different models that optimize different objective functions.

However, if the evaluation results of model M_A are better than these of model M_B when both objective functions of M_A and M_B are used as the evaluation criteria, we can then say M_A outperforms M_B in terms of these two evaluation criteria.

In this experiment, in order to compare model FLR_{FS} with models FLR_{KC02}, FLR_{KC03} and FLR_{NN04}, both average similarity $AveS_H$ and total error $TotE_{KC}$ are used as the evaluation criteria, which are defined as follows:

$$AveS_H = \frac{1}{n}\sum_{i=1}^{n} S_{Hi},\ \ TotE_{KC} = \sum_{i=1}^{n} E_{KC}$$

[4] In some thesis, a fuzzy number is expressed by its lower bound, centre and upper bound. To keep the notation consistency in this chapter, we express a fuzzy number by its centre, left and right spreads.

because: (i) the objective function of model FLR_{FS} is to maximise $AveS_H$ between observed response variables $\tilde{Y}$ and their estimated counterparts $\hat{\tilde{Y}}$, and (ii) the objective function of models FLR_{KC02}, FLR_{KC03} and FLR_{NN04} is to minimise $TotE_{KC}$ between $\tilde{Y}$ and $\hat{\tilde{Y}}$.

According to the definitions of $AveS_H$ and $TotE_{KC}$, a better method in terms of these two criteria should have a higher $AveS_H$ value and a lower $TotE_{KC}$ value.

Table 7.4 shows the performances of model FLR_{FS}, and models FLR_{KC02}, FLR_{KC03}, FLR_{NN04} and FLR_{CD08} on $AveS_H$ and $TotE_{KC}$. From Table 7.4, we can see that the *MaxSim* solution of model FLR_{FS} outperforms all other four models in terms of $AveS_H$, and outperforms the other three models in terms of $TotE_{KC}$ except model FLR_{CD08}.

Table 7.4 Comparison of the performance of difference methods on Dataset2

Models	$AveS_H$	$TotE_{KC}$
FLR_{FS}(*MaxSim*) (given the initial values (a))	0.5462	8.9659
FLR_{KC02} (Kao and Chyu, 2002)	0.4663	9.679
FLR_{KC03} (Kao and Chyu, 2003)	0.4095	10.089
FLR_{NN04} (Nasrabadi and Nasrabadi, 2004)	0.4388	9.771
FLR_{CD08} (Chen and Dang, 2008)	0.5198	7.857

Although model FLR_{FS} achieves better performances than other three models (FLR_{KC02}, FLR_{KC03} and FLR_{NN04}) in terms of $AveS_H$ and $TotE_{KC}$, and a better performance than FLR_{CD08} in terms of $AveS_H$, it cannot guarantee that the solution of model FLR_{FS} given in Eq.(7.29) is the global optimization. The FLR model given by Hojati et al. (2005) is shown as follows:

$$\hat{\tilde{Y}}_i = (6.75 + 1.25m_{x_i}, 1.65 + 0.15m_{x_i})_L$$

Their model obtains a higher $AveS_H$ value (0.5515) and a lower $TotE_{KC}$ value (8.3986) than model FLR_{FS} given in Eq. (7.29). This is because the *MaxSim* solution of model FLR_{FS} obtained in this experiment is a local maximum, not a global one. Given appropriate initial values, model FLR_{FS} can also find the solution obtained in Hojati, et al. (2005). This is because the solution given in Hojati, et al. (2005) can be expressed by model FLR_{FS} as follows:

$$K_0=6.75;\ k_1=1.25;\ k_{ll}=0;\ k_{lm}=0.15;\ c=1.65.$$

Although the solution given by Hojati, et al. (2005) outperforms the solution of model FLR_{FS} given in Eq.(7.29) in terms of $AveS_H$ and $TotE_{KC}$, the model proposed by Hojati, et al. (2005) has the *spreads increasing problem*.

From this example, we can see that initial values are important for finding the *MaxSim* solution of FLR_{FS}. In Section 7.6, we will give a potential solution on how to set initial values in order to find the global optimization.

Example 3. In this example, we compare model FLR_{FS} with other models using the fuzzy input-fuzzy output dataset given by Sakawa and Yano (1992b), which is shown in the left half of Table 7.5, $\tilde{X}_i$'s are the observed independent variables. $\tilde{Y}_i$'s are the observed dependent variables.

Table 7.5 Dataset3

i	$\tilde{X}_i = (m_{x_i},\ \alpha_{x_i})_L$	$\tilde{Y}_i = (m_{y_i},\ \alpha_{y_i})_L$	$\hat{\tilde{Y}}_i = (m_{\hat{y}_i},\ \alpha_{\hat{y}_i})_L$	S_H	E_{KC}
1	$(2.0, 0.5)_L$	$(4.0, 0.5)_L$	$(4.0, 0.5)_L$	1.0000	0.0000
2	$(3.5, 0.5)_L$	$(5.5, 0.5)_L$	$(4.7857, 0.5)_L$	0.0426	0.9184
3	$(5.5, 1.0)_L$	$(7.5, 1.0)_L$	$(5.833, 1.6758)_L$	0.0766	2.2952
4	$(7.0, 0.5)_L$	$(6.5, 0.5)_L$	$(6.619, 0.5)_L$	0.6341	0.2239
5	$(8.5, 0.5)_L$	$(8.5, 1.0)_L$	$(7.4048, 0.5)_L$	0.0000	1.0000
6	$(10.5, 1.0)_L$	$(8.0, 1.0)_L$	$(8.4524, 1.6758)_L$	0.4957	0.9021
7	$(11.0. 0.5)_L$	$(10.5, 0.5)_L$	$(8.7143, 0.5)_L$	0.0000	1.0000
8	$(12.5, 0.5)_L$	$(9.5, 0.5)_L$	$(9.5, 0.5)_L$	1.0000	0.0000
Average				0.4061	0.7925

The independent variables and the observed dependent variables in dataset3 are *L*-type. Thus, *L*-type fuzzy number based FLR_{FS} model is adopted. According to the initial value setting strategy given in Section 7.5.1, the initial values can be set as[5]:

Initial values (b): k_0=3.5724; k_1=0.5193; k_{ll}=1; k_{lm}=0; c=0

Then, the following regression model is obtained,

$$\hat{\tilde{Y}}_i = (2.9524 + 0.5238 m_{x_i},\ 2.3516\alpha_{x_i} - 0.6758)_L \qquad (7.30)$$

[5] By applying the least-squares estimation to α_{y_i} and α_{x_i}, and α_{y_i} and m_{x_i}, we get $\alpha_{y_i} = 1.0 * \alpha_{x_i}$ and $\alpha_{y_i} = 0.5911 + 0.0045 m_{x_i}$. Since 0.0045 << 1, we set the initial spreads of estimated response only depend on α_{x_i}.

Estimated response $\hat{\tilde{Y}}_i$, S_H and E_{KC} for each instance of dataset3 are listed in the right half of Table 7.5.

The *MaxSim* solution of model FLR_{FS} is compared with the other four models: FLR_{KC02}, FLR_{KC03}, FLR_{NN04} and FLR_{CD08}, which are listed in Table 7.6. $\tilde{E}_i$'s in model FLR_{CD08}, are $\tilde{E}_1 = (-0.234, 0, 0.234)$, $\tilde{E}_2 = (-0.234, 0, 0.234)$, $\tilde{E}_3 = (0, 0, 0.935)$, $\tilde{E}_4 = (-0.234, 0, 0.234)$, $\tilde{E}_5 = (-0.234, 0, 0.234)$, $\tilde{E}_6 = (-0.935, 0, 0)$, $\tilde{E}_7 = (-0.234, 0, 0.234)$ and $\tilde{E}_8 = (-0.234, 0, 0.234)$.

Table 7.6 Fuzzy regression models of dataset3

FLR_{FS}(MaxSim) (Given initial values (b))	$\hat{\tilde{Y}}_i = (2.9524 + 0.5238m_{x_i}, 2.3516\alpha_{x_i} - 0.6758)_L$
FLR_{KC02} (Kao and Chyu, 2002)	$\hat{\tilde{Y}}_i = (3.5724 + 0.5193m_{x_i}, 0.5193\alpha_{x_i} + 0.24)_L$
FLR_{KC03} (Kao and Chyu, 2003)	$\hat{\tilde{Y}}_i = (3.554 + 0.522m_{x_i}, 0.522\alpha_{x_i} + 0.951,\ 0.522\alpha_{x_i} + 0.949)_{LR}$
FLR_{NN04} (Nasrabadi and Nasrabadi, 2004)	$\hat{\tilde{Y}}_i = (3.5767 + 0.5467m_{x_i}, \alpha_{x_i})_L$
FLR_{CD08} (Chen and Dang, 2008)	$\hat{\tilde{Y}}_i = 3.5284 + 0.5298m_{x_i} + \tilde{E}_i$

To compare the performances of model FLR_{FS}, and models FLR_{KC02} (Kao and Chyu, 2002), FLR_{KC03} (Kao and Chyu, 2003), FLR_{NN04} (Nasrabadi and Nasrabadi, 2004) and FLR_{CD08} (Chen and Dang, 2008), $AveS_H$ and $TotE_{KC}$ are also used as evaluation criteria in this example.

The comparison of the performances of different methods on $AveS_H$ and $TotE_{KC}$ is given in Table 7.7. From Table 7.7, we can see that the average similarity of FLR_{FS} is significantly higher than that of the other methods. Also, the total error of FLR_{FS} is significantly lower than that of the other methods. Thus, model FLR_{FS} outperforms all four models (FLR_{KC02}, FLR_{KC03}, FLR_{NN04} and FLR_{CD08}) in terms of both $AveS_H$ and $TotE_{KC}$ on dataset3.

Table 7.7 Comparison of the performance of difference methods on Dataset3

Models	$AveS_H$	$TotE_{KC}$
FLR_{FS}(*MaxSim*) (given the initial values (b))	0.4061	6.3396
FLR_{KC02} (Kao and Chyu, 2002)	0.1499	7.470
FLR_{KC03} (Kao and Chyu, 2003)	0.2351	9.363
FLR_{NN04} (Nasrabadi and Nasrabadi, 2004)	0.2026	7.541
FLR_{CD08} (Chen and Dang, 2008)	0.1854	7.000

Example 4. In this example, we investigate the effectiveness of model FLR_{FS} for multiple regression on the restaurants data given by D'Urso (2003), which are listed in Table 7.8. The restaurants data are drawn from an Italian specialized book, which concerns the performances of 30 good-quality Roman restaurants, where fuzzy inputs $\tilde{X}_{i1}$'s and $\tilde{X}_{i2}$'s are *decision on cooking* and *decision on cellar*, respectively; and $\tilde{Y}$ is *decision on service*.

Table 7.8 Dataset4: Restaurants data

i	$\tilde{X}_{i1}=(m_{x_{i1}},\ \alpha_{x_{i1}},\ \beta_{x_{i1}})_{LR}$	$\tilde{X}_{i2}=(m_{x_{i2}},\ \alpha_{x_{i2}},\ \beta_{x_{i2}})_{LR}$	$\tilde{Y}_{i}=(m_{y_{i}},\ \alpha_{y_{i}},\ \beta_{y_{i}})_{LR}$
1	$(7, 0.5, 1.25)_{LR}$	$(8,0.75,1)_{LR}$	$(8,0.75,1)_{LR}$
2	$(7, 0.5, 1.25)_{LR}$	$(7,0.5,1.25)_{LR}$	$(6,0.25,0.5)_{LR}$
3	$(6, 0.25, 0.5)_{LR}$	$(7,0.5,1.25)_{LR}$	$(6,0.25,0.5)_{LR}$
4	$(8, 0.75, 1)_{LR}$	$(9,0,1)_{LR}$	$(9,0,1)_{LR}$
5	$(8, 0.75, 1)_{LR}$	$(8,0.75,1)_{LR}$	$(8,0.75,1)_{LR}$
6	$(6, 0.25, 0.5)_{LR}$	$(7,0.5,1.25)_{LR}$	$(5,0,1)_{LR}$
7	$(7, 0.5, 1.25)_{LR}$	$(8,0.75,1)_{LR}$	$(7,0.5,1.25)_{LR}$
8	$(7, 0.5, 1.25)_{LR}$	$(7,0.5,1.25)_{LR}$	$(5,0,1)_{LR}$
9	$(7, 0.5, 1.25)_{LR}$	$(8,0.75,1)_{LR}$	$(7,0.5,1.25)_{LR}$
10	$(6, 0.25, 0.5)_{LR}$	$(7,0.5,1.25)_{LR}$	$(6,0.25,0.5)_{LR}$
11	$(7, 0.5, 1.25)_{LR}$	$(8,0.75,1)_{LR}$	$(8,0.75,1)_{LR}$
12	$(7, 0.5, 1.25)_{LR}$	$(6,0.25,0.5)_{LR}$	$(6,0.25,0.5)_{LR}$
13	$(7, 0.5, 1.25)_{LR}$	$(8,0.75,1)_{LR}$	$(9,0,1)_{LR}$
14	$(7, 0.5, 1.25)_{LR}$	$(8,0.75,1)_{LR}$	$(8,0.75,1)_{LR}$
15	$(7, 0.5, 1.25)_{LR}$	$(7,0.5,1.25)_{LR}$	$(7,0.5,1.25)_{LR}$
16	$(7, 0.5, 1.25)_{LR}$	$(7,0.5,1.25)_{LR}$	$(7,0.5,1.25)_{LR}$
17	$(6, 0.25, 0.5)_{LR}$	$7,0.5,1.25)_{LR}$	$(6,0.25,0.5)_{LR}$
18	$(7, 0.5, 1.25)_{LR}$	$(8,0.75,1)_{LR}$	$(7,0.5,1.25)_{LR}$
19	$(7, 0.5, 1.25)_{LR}$	$(7,0.5,1.25)_{LR}$	$(8,0.75,1)_{LR}$
20	$(7, 0.5, 1.25)_{LR}$	$(9,0,1)_{LR}$	$(7,0.5,1.25)_{LR}$
21	$(7, 0.5, 1.25)_{LR}$	$(8,0.75,1)_{LR}$	$(7,0.5,1.25)_{LR}$
22	$(7, 0.5, 1.25)_{LR}$	$(8,0.75,1)_{LR}$	$(6,0.25,0.5)_{LR}$
23	$(7, 0.5, 1.25)_{LR}$	$(9,0,1)_{LR}$	$(7,0.5,1.25)_{LR}$
24	$(7, 0.5, 1.25)_{LR}$	$(7,0.5,1.25)_{LR}$	$(8,0.75,1)_{LR}$
25	$(7, 0.5, 1.25)_{LR}$	$(7,0.5,1.25)_{LR}$	$(6,0.25,0.5)_{LR}$
26	$(7, 0.5, 1.25)_{LR}$	$(7,0.5,1.25)_{LR}$	$(6,0.25,0.5)_{LR}$
27	$(7, 0.5, 1.25)_{LR}$	$(7,0.5,1.25)_{LR}$	$(7,0.5,1.25)_{LR}$
28	$(7, 0.5, 1.25)_{LR}$	$(8,0.75,1)_{LR}$	$(7,0.5,1.25)_{LR}$
29	$(7, 0.5, 1.25)_{LR}$	$(7,0.5,1.25)_{LR}$	$(7,0.5,1.25)_{LR}$
30	$(6, 0.25, 0.5)_{LR}$	$(7,0.5,1.25)_{LR}$	$(6,0.25,0.5)_{LR}$

The independent variables are 2-dimensional. Thus, model FLR_{FS} for multiple regression is adopted. According to the initial value setting strategy given in Section 7.5.1, the following parameters of model FLR_{FS} are obtained,

k_0=-1.66; k_1=1.33; k_2=0.00; (7.31)

k_{ll1}=-0.91; k_{lr1}=0.00; k_{lm1}=0.63; k_{ll2}=-0.65; k_{lr2}=-1.48; k_{lm2}=-0.28; c_l=0.80;

k_{rl1}=0.37; k_{rr1}=1.00; k_{rm1}=-0.12; k_{rl2}=0.09; k_{rr2}=0.20; k_{rm2}=0.03; c_r=0.08;

The regression parameters of model $FLR_{D'Urso03}$ on dataset4 are estimated by minimising the total squared Euclidean distance d^2 between the estimated and observed response variables, which are listed as follows (D'Urso, 2003):

$$\mathbf{a} = (0.6498399, 0.4542534, 0.4924441)^T \quad (7.32)$$

$$\mathbf{r} = (-1.868527, 2.3604004, 0.7392849)^T$$

$$\mathbf{s} = (-0.233325, -0.13392, 0.1271022)^T$$

b = 0.1173197, d = -0.401173, g = 0.2306911, h = -0.650102

In order to compare model FLR_{FS} with model $FLR_{D'Urso0}$, both $AveS_H$ and *sum of* d^2 are used as the evaluation criteria, because: (i) the objective function of model FLR_{FS} is to maximise $AveS_H$ between $\tilde{Y}$ and $\hat{\tilde{Y}}$, and (ii) the objective function of $FLR_{D'Urso03}$ is to minimise total squared Euclidean distance d^2 between $\tilde{Y}$ and $\hat{\tilde{Y}}$.

The comparison of the performances of FLR_{FS} and $FLR_{D'Urso03}$ on both $AveS_H$ and total d^2 is given in Table 7.9. From Table 7.9, we can see that model FLR_{FS} outperforms model $FLR_{D'Urso03}$ in terms of both $AveS_H$ and *sum of* d^2.

Table 7.9 Comparison of the performance of difference methods on Dataset4

Models	$AveS_H$	*Sum of* d^2
FLR_{FS}(*MaxSim*)	0.5675	64.5683
$FLR_{D'Urso03}$	0.2263	73.6945

7.6 Summary

In this chapter, we have proposed model FLR_{FS} that has more flexible spreads compared with previous FLR models. A property of model FLR_{FS} is that the spreads of estimated response variables are able to fit the spreads of observed response variables, no matter if the spreads of the observed response variables are increased, decreased or unchanged when the spreads and magnitudes of the independent variables change. This property

makes model FLR_{FS} be able to avoid the *spreads increasing problem* that exists in model FLR_{FF} and overcome the deficiencies of models FLR_{KC02}, FLR_{KC03}, FLR_{NN04}, and $FLR_{D'Urso03}$, which are mentioned in Section 7.3. The number of parameters in model FLR_{FS} only proportionally increases with the increase of the dimension of independent variables, while the number of parameters in model FLR_{CD08} increases with the increase of both the dimension and the number of independent variables. This makes FLR_{FS} more suitable to dataset with large instance number than FLR_{CD08}.

The parameters of model FLR_{FS} are estimated by maximising the average similarity between the estimated and observed response variables. The experimental results show that model FLR_{FS} has a better performance than previous models in terms of $AveS_H$, $TotE_{KC}$ and $sum\ d^2$. Also, the experimental results show that initial value setting is important for parameter estimation of FLR_{FS} when the objective function is set as maximising the average similarity between the estimated and observed response variables. Although we have given a strategy for the initial value setting, it cannot guarantee that the generated regression model is the global optimisation. Our future work is to find a more sophisticated initial value setting strategy to achieve better solutions of model FLR_{FS}. Genetic algorithms may be potential solutions for the initial value setting problem of model FLR_{FS}.

Chapter 8.

Conclusions and Future Work

This chapter summarises our main research findings and highlights the contributions of our research, then discusses further research possibilities.

8.1 Summary of main findings and contributions

In this thesis, we have proposed and presented the development scenario of a Computer-Assisted sentence Stress Teaching and Learning Environment (CASTLE). By reviewing the literature of English as a Second Language (ESL) teaching and learning, we have found that the sentence stress plays an important role in mutual understanding during a conversation. However, sentence stress, as well as other suprasegmental features, has been overlooked in traditional English language teaching classes and Computer-Assisted Pronunciation Teaching (CAPT) programmes. Thus, our proposed and developed CASTLE system is a vital CAPT system to fill this gap, which is intended to help ESL learners improve their ability to correctly use sentence stress.

By investigating the needs and difficulties faced by ESL learners of sentence stress, we have proposed a framework for sentence stress teaching systems, which includes three modules: an individualised speech material providing module, a perception assistance module and a production assistance module. Based on this framework, we have designed and developed our CASTLE system.

8.1.1 Individualised speech learning material

The individualised speech learning material providing module of CASTLE can provide individualised speech material for different learners, which possesses learners' preferred voice features. This individualised speech material is to create a pleasant learning environment and increase the learners' learning interests.

In order to provide learners' favourite voice, we have investigated which voice features (i.e. gender, pitch and speech rate) make a teacher's voice preferred by the language learners to imitate. We took a single teacher's voice as the source to automatically resynthesise speech material with different voice features, which cover a wide range of pitch values, speech rates and different genders. We have found that learners' imitation preference vary, according to many factors (e.g. English background and language proficiency). Voices with similar pitch values and speech rates to learners' own voice are not always what they prefer to imitate during a language learning session. Some learners prefer to listen to voice produced by the opposite gender to themselves rather than the same gender. Learners with advanced listening proficiency are more willing to choose voices with a normal speech rate or a little fast speech rate to listen to, while novices are more willing to listen to voices with a lower speed, which might help them to catch more speech features that they are not familiar to, such as linking, assimilation.

Thus, we advocate using prosody modification techniques to automatically transform original speech learning material into individualised speech material for different learners, which have the learners' preferred voice features. Current CAPT systems either only provide a single teacher's voice or provide multiple teachers' voices as learning speech. A single teacher's voice cannot give learners an opportunity to be exposed to more variations in speech. Providing multiple teachers' voices also multiplies the effort of recording speech learning material and the storage space. Moreover, no matter how wide the range of the prosodic features of the multiple 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. In contrast, CASTLE uses a single teacher's voice as the source to automatically resynthesise voices with learners preferred voices features. This can better satisfy different learners' needs. Moreover, in CASTLE, learners are also allowed to control the prosody modification by dragging the sliders or adjusting the pitch and speech rate changing factors. This gives learners more control and autonomy in the process of learning.

8.1.2 Stress-exaggeration-based perception assistance

The perception assistance module of CASTLE can provide stress-exaggerated speech learning material which enlarges the differences between stressed and unstressed syllables in teachers' speech. In order to produce sentence stress correctly, learners need to perceive it first. However, current studies show that some ESL learners have difficulties in perceiving sentence stress. The stress-exaggerated speech learning material in the perception assistance module of CASTLE is to help ESL learners to correctly perceive sentence stress.

In perception assistance module of CASTLE, we have proposed a set of automatic stress exaggeration methods to resynthesise stress-exaggerated speech learning material, based on normal speech material. Four stress exaggeration methods have been presented in this thesis: pitch-based exaggeration, duration-based exaggeration, intensity-based exaggeration, and a combined exaggeration method that integrates the previous three single-prosody-feature-based exaggeration methods.

The results of our experiments showed that: (i) all the three exaggeration methods based on single prosodic features (i.e. pitch, duration and intensity) are able to improve ESL listeners' English stress identification accuracy; (ii) among the three single-prosody-feature-based exaggeration methods, the duration-based method is the most effective one; (iii) the combined stress exaggeration method improves the perception accuracy of listeners' English stress more significantly than every single-prosody-feature-based method; (iv) the combined method and the duration-based method help the listeners to improve their stress labeling accuracy at the significance level of 0.01 for a Student's t-Test, and the pitch-based and intensity-based methods help the listeners to improve their stress labeling accuracy at the significance level of 0.05.

8.1.3 Production assistance

The production assistance module of CASTLE consists of a clapping-based pronunciation practice assistance model and three stress-error feedback models which are intended to help ESL learners produce sentence stress correctly.

The clapping-based pronunciation practice model is to help ESL learners get familiar with the rhythm of English language and train them to acoustically emphasize syllables which are supposed to be stressed. Learning the rhythm of English language can help learners to produce sentence stress correctly, since the rhythm of English language is composed of regular occurrences of stressed syllables. Clapping, as well as other body languages such as stepping or snapping fingers, has been used in English language teaching to teach rhythm. However, in English language learning classes, the clapping-based rhythm teaching requires English teachers to extract the rhythm of English language in order to produce speech with clapping hands at each stressed syllables. This is challenging for some non-native English-speaking teachers, since they lack of professional knowledge and confidence in teaching rhythm. Thus, in CASTLE, we proposed to resynthesise *clapping-based teacher's utterances* by automatically detecting stress of the original teacher's utterances and adding a clap sound to every stressed syllable of the original teacher's utterances. By practising with the clapping-based pronunciation practice model, learners are expected to be better aware of the stress patterns of English speech and establish a habit of giving more emphasis to the syllables that are supposed to be stressed.

Instead of categorical representation of stress, we have proposed a fuzzy representation of stress. Stress is conventionally treated as a categorical concept in linguistics. However, the strict boundary between stressed and unstressed syllables in the categorical representation cannot represent the uncertainty of stress (i.e. intra-labeler uncertainty and inter-labeler uncertainty). Compared with the categorical representation of stress, the fuzzy representation can better describe the subjective nature of stress.

Based on the fuzzy representation of stress, three stress-error feedback models have been presented (i.e. Feedback$_{PC}$, Feedback$_{MC}$ and Feedback$_{DI}$) which are intended to provide learners with valuable feedback in order to help them to realise and correct their stress errors. Model Feedback$_{PC}$ is to deal with the *intra-labeler* uncertainty, which is based on prediction confidence of automatic stress detector(s). Model Feedback$_{MC}$ is intended to handle the *inter-labeler* uncertainty, which is based on multiple stress detectors. Also, based on learners' more than one imitation, model Feedback$_{DI}$ aims to provide feedback for learners' repeated stress errors.

8.1.4 Linear-Regression-based flexible boundary phoneme aligner

We have also proposed a Linear-Regression-based Flexible Boundary (LR-FB) phoneme aligner for CASTLE. Segmenting continuous speech into syllables is a crucial process in CASTLE system. The accuracy of automatic segmentation has a strong impact on the performance of CASTLE system. Moreover, phoneme duration is an importance cue for automatic stress detection. Thus, we have developed a LR-FB phoneme aligner that generates the final phoneme boundaries by combining the outcomes of multiple base phoneme aligners.

There are two distinctive features of our LR-FB phoneme aligner. (i) The LR-FB phoneme aligner minimises both phoneme boundary differences and phoneme duration differences between the predicted and reference time alignments of phonemes. (ii) In the LR-FB phoneme aligner, the relationship between two conjunctive phoneme segments is more flexible. It can be overlapped, connected or unconnected. In contrast, in conventional speech segmentation systems, the relationship between two conjunctive phoneme segments can only be connected. In other words, the starting point of a phoneme is identical to the ending point of its previous phoneme. The flexible phoneme boundary relationship in the LR-FB phoneme aligner has the ability to better model the relations between the durations of two conjunctive phonemes than the simply connected phoneme boundary relationship. This is because some phoneme transitions are ambiguous. It is not easy or impossible to find the strict boundary between two conjunctive phonemes. Therefore, the connected boundary relationship alone cannot model the ambiguity in phoneme transitions. Experimental results demonstrated the effectiveness of our LR-FB phoneme aligner. The LR-FB phoneme aligner increases phoneme segmentation accuracy from 83.35% to 86.79%.

8.1.5 An enhanced fuzzy linear regression model

In addition to the development of CASTLE system, we have also proposed an enhanced fuzzy linear regression model (FLR_{FS}) with more flexible spreads, which can overcome the spreads increasing problem encountered by previous fuzzy linear regression models. The spreads increasing problem is that with the increase of the magnitudes of

independent variables, the spreads of estimated fuzzy dependent variables are increasing, even though the spreads of observed dependent variables actually decrease or remain unchanged. In model FLR_{FS}, the spreads of estimated response variables can fit the spreads of observed response variables, no matter if the spreads of the observed response variables are increased, decreased or unchanged when the spreads and magnitudes of the independent variables change. This makes model FLR_{FS} be able to avoid the *spreads increasing problem*.

8.2 Further research

Further research, which has been suggested by the work to date, is as follows:

Firstly, we noticed that the CASTLE system can be further improved. More diverse individualised speech learning material can be provided. The individualised speech material module in CASTLE only changes the voice features of teacher's speech (i.e. gender, pitch and speech rate). However, ESL learners may prefer different accents such as American accent, British accent, or New Zealand accent. It would be more useful if a CAPT system can change the accent of any speech learning material into a target accent. Moreover, currently CASTLE system is a stand-alone system. In order to use the system, ESL learners need to setup CASTLE system in their computers. It would be more convenient if an online version of CASTLE system was available.

Secondly, frameworks and technologies to teach other suprasegmentals (e.g. intonation) are needed to be proposed and developed. In our research work, we have proposed a framework for sentence stress teaching systems, and based on this framework we have designed and developed CASTLE system that only focuses on sentence stress teaching. However, intonation is also known to be important for verbal communication, and has been overlooked by traditional ESL teaching and current CAPT systems. Although there are a few CAPT systems that are intended to support intonation teaching, they provide no feedback or only simple graphic display feedback that is difficult for ESL learners to interpret. Thus, ESL learning calls for a more effective way to teach intonation, and more CAPT systems that can support intonation teaching and provide more useful feedback.

Appendix Questionnaire

1. Age: □19-, □20-29, □30-39, □40+ (please use ☑ to indicate)
2. Gender: □Male, □Female (please use ☑ to indicate)
3. First language(s):
4. How long have you been learning English:
5. How long have you been in an English speaking country?
6. How do you rank your proficiency of English? (please use ☑ to indicate)

	Low		Moderate		High
Speaking:	□	□	□	□	□
Listening:	□	□	□	□	□
Writing:	□	□	□	□	□
Reading:	□	□	□	□	□

7. What are the methods you have used to improve your sentence pronunciation? And which do you think is the best?
8. Have you ever tried to improve your pronunciation by imitating a teacher's voice, an actor's voice in a movie, or a radio announcer's voice in news?

If you answered "yes" to question 8, please answer questions 9 and 10.

9. Suppose you have the voices of different speakers uttered same sentences, and you have to choose one of those voices to imitate. What criteria would you use to choose? If you select multiple items, please rank them from the most important factor to the least one.

   (1) Same gender, (2) Reasonable speed, (3) Attractiveness, (4) Others (please specify):
10. Is there sometime you want to imitate a portion of speech in a movie or radio news, but it is too fast for you to repeat it and it is said by an opposite gender of you? Do you wish it is to be said a little bit slowly and by the same gender as you?

References

Akahane-Yamada, R., Tohkura, Y., Bradlow, A. R. and Pisoni, D. B. (1996). Does training in speech perception modify speech production? In *Proc. International Conference on Spoken Language Processing*, Vols *1-4*, 606-609.

Ananthakrishnan, S. and Narayanan, S. S. (2008). Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16 (1), 216-228.

Anderson-Hsieh, J., Johnson, R. and Koehler, K. (1992). The relationship between native speaker judgments of non-native pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, 42 (4), 529-555.

Arnett, M.K. (1952). Does the elementary teacher have time to teach speech? *Journal of the Southern States Communication Association*, 17(3), 203-208.

Bargiela, A., Pedrycz, W. and Nakashima, T. (2007). Multiple regression with fuzzy data. *Fuzzy Sets and Systems*, 158 (19), 2169-2188.

Beskow, J. and Sjölander, K. (2000). Wavesurfer - a public domain speech tool. In *Proc. International Conference on Spoken Language Processing*, 464-467.

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

Black, A. (2007). Speech synthesis for educational technology. In *Proc. Workshop on Speech and Language Technology in Education*, 104-107.

Boersma, P. and Weenink, D. (2009). Praat: Doing phonetics by computer (version 5.1.05). Retrieved may 1, 2009, from http://www.Praat.Org/.

Bond, Z. S. (1999). *Slips of the ear: Errors in the perception of casual conversation*: Academic Press.

Bond, Z. S. and Small, L. H. (1983). Voicing, vowel, and stress mispronunciations in continuous speech. *Perception & Psychophysics*, 34 (5), 470-474.

Bongaerts, T. (1999). Ultimate attainment in l2 pronunciation: The case of very advanced late l2 learners. *Second Language Acquisition and the Critical Period Hypothesis*, 133-159.

Bradlow, A., Pisoni, D., Akahana-Yamada, R. and Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, 101 (4), 2299-2310.

Chen, S.-P. and Dang, J.-F. (2008). A variable spread fuzzy linear regression model with higher explanatory power and forecasting accuracy. *Information Sciences*, 178 (20), 3973-3988.

Clark, J., Yallop, C. and J., F. (2007). *An introduction to phonetics and phonology. 3rd edition*: Blackwell Publishing.

Coppi, R. (2008). Management of uncertainty in statistical reasoning: The case of regression analysis. *International Journal of Approximate Reasoning*, 47 (3), 284-305.

Coppi, R. and D'Urso, P. (2003). Regression analysis with fuzzy informational paradigm: A least-squares approach using membership function information. *International Journal of Pure and Applied Mathematics*, 8 (3), 279-306.

Coppi, R., D'Urso, P., Giordani, P. and Santoro, A. (2006). Least squares estimation of a linear regression model with lr fuzzy response. *Computational Statistics & Data Analysis*, 51 (1), 267-286.

Cutler, A. and Clifton, C. E. (1984). The use of prosodic information in word recognition. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance x* (pp. 183-196): Hillsdale, NJ: Lawrence Erlbaum.

D'Urso, P. (2003). Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data. *Computational Statistics & Data Analysis*, 42 (1-2), 47-72.

Dalton, C. and Seidlhofer, B. (1994). *Pronunciation*: Oxford: Oxford University Press.

Derwing, T. M. (2003). What do ESL students say about their accents? *Canadian Modern Language Review-Revue Canadienne Des Langues Vivantes*, 59 (4), 547-566.

Derwing, T. M. and Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *Tesol Quarterly*, 39 (3), 379-397.

Derwing, T. M., Munro, M. J. and Wiebe, G. (1998). Evidence in favour of a broad framework for pronunciation instruction. *Language Learning*, 48 (3), 393-410.

Derwing, T. M. and Rossiter, M. J. (2002). ESL learners' perceptions of their pronunciation needs and strategies. *System*, 30 (2), 155-166.

Deshmukh, O. D. and Verma, A. (2009). Nucleus-level clustering for word-independent syllable stress classification. *Speech Communication*, 51 (12), 1224-1233.

Diamond, P. (1988). Fuzzy least squares. *Information Sciences*, 46 (3), 141-157.

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

Dubois, D. and Prade, H. (1980). *Fuzzy sets and systems: Theory and application*. New York: Academic Press.

Dupoux, E., Pallier, C., Sebastian, N. and Mehler, J. (1997). A destressing "deafness" in French? *Journal of Memory and Language*, 36 (3), 406-421.

Dyck, C. (2002). Review of tsi karhakta: At the edge of the woods. *Language Learning & Technology*, 6 (2), 27-33.

Engelbrecht, K.-P., Quade, M. and Möller, S. (2009). Analysis of a new simulation approach to dialog system evaluation. *Speech Communication*, 51 (12), 1234-1252.

Erro, D. and Moreno, A. (2007). Weighted frequency warping for voice conversion. In *Proc. EuroSpeech*, 1965–1968.

Eskenazi, M. (1999). Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. *Language Learning & Technology*, 2 (2), 62-76.

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

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

Eskenazi, M., Ke, Y., Albornoz, J. and Probst, K. (2000). The fluency pronunciation trainer: Update and user issues. In *Proc. InSTiLL Workshop on Speech Technology in Language Learning*, 73-76.

Fan, C.-Y., Chen, C.-F. and Lin, H.-P. (1998). Helping Chinese students to develop sensitivity to English rhythm. *Studies in English Language and Literature*, 3, 13-17.

Fant, G. (1960). *Acoustic theory of speech production*: Mouton, the Hague.

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

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

Flege, J. E., Munro, M. J. and Mackay, I. R. A. (1995). Factors affecting strength of perceived foreign accent in a 2nd language. *Journal of the Acoustical Society of America*, 97 (5), 3125-3134.

Florez, M. A. C. (1998). Improving adult ESL learners' pronunciation skills. *ERIC Digest. (ED427553)* Retrieved 01/06/2010 from *ERIC database.*

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S. and Dahlgren, N. L. (1990). DARPA TIMIT acoustic-phonetic continuous speech corpus cd-rom, national institute of standards and technology.

Gholampour, I. and Nayebi, K. (1998). A new fast algorithm for automatic segmentation of continuous speech. In *Proc. International Conference on Spoken Language Processing*, 9-12.

Gong, J. (2002). Introducing English rhythm in Chinese EFL classrooms: a literature review. *Post-Script: the Postgraduate Journal of Education Research*, 3 (1).

Hahn, L. D. (2004). Primary stress and intelligibility: research to motivate the teaching of suprasegmentals. *Tesol Quarterly*, 38 (2), 201-223.

Hasegawa-Johnson, M., Chen, K., Cole, J., Borys, S., Kim, S.-S., Cohen, A., et al. (2005). Simultaneous recognition of words and prosody in the Boston University radio speech corpus. *Speech Communication*, 46 (3-4), 418-439.

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

Hincks, R. (2003). Speech technologies for pronunciation feedback and evaluation. *ReCALL Journal*, 15 (1), 3-20.

Hirose, K. (2004). Accent type recognition of Japanese using perceived mora pitch values and its use for pronunciation training system. In *Proc. International Symposium on Tonal Aspects of Languages*, 77-80.

Hişmanoğlu, M. (2006). Current perspectives on pronunciation learning and teaching. *Journal of Language and Linguistic Studies*, 2 (1), 101-110.

Hojati, M., Bector, C. R. and Smimou, K. (2005). A simple method for computation of fuzzy linear regression. *European Journal of Operational Research*, 166 (1), 172-184.

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

Hosom, J.-P. (2009). Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication*, 51 (4), 352-368.

Hung, W.-L. and Yang, M.-S. (2006). An omission approach for detecting outliers in fuzzy regression models. *Fuzzy Sets and Systems*, 157 (23), 3109-3122.

Ilčiukienė, G. (2005). Teaching English rhythm through Jazz chanting. *Pedagogy Studies*, 78, 68-72.

Imoto, K., Dantsuji, M. and Kawahara, T. (2000). Modelling of the perception of English sentence stress for computer-assisted language learning. In *Proc. International Conference on Spoken Language Processing*, 175-178.

Imoto, K., Tsubota, Y., Raux, A., Kawahara, T. and Dantsuji, M. (2002). Modeling and automatic detection of english sentence stress for computer-assisted english prosody learning system. In *Proc. International Conference on Spoken Language Processing*, 749-752.

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

Jande, P.-A. (2008). Spoken language annotation and data-driven modelling of phone-level pronunciation in discourse context. *Speech Communication*, 50 (2), 126-141.

Jarifi, S., Pastor, D. and Rosec, O. (2008). A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis. *Speech Communication*, 50 (1), 67-80.

Jenkins, J. (1998). Which pronunciation norms and models for English as an international language. *ELT Journal*, 52 (2), 119-126.

Juffs, A. (1990). Tone, syllable structure and interlanguage phonology: Chinese learners' stress errors. *International Review of Applied Linguistics*, 28 (2), 99-118.

Kabré, H., Pérennou, G. and Vigouroux, N. (1991). A nonlinear filtering method applied to automatic segmentation of multilingual speech corpora. In *Proc. European Conference on Speech Communication and Technology*, 689-702.

Kahraman, C., Beskese, A. and Bozbura, F. T. (2006). Fuzzy regression approaches and applications. *Fuzzy Applications in Industrial Engineering*, 201, 589-615.

Kamakshi Prasad, V., Nagarajan, T. and Murthy, H. A. (2004). Automatic segmentation of continuous speech using minimum phase group delay functions. *Speech Communication*, 42 (3-4), 429-446.

Kao, C. and Chyu, C.-L. (2002). A fuzzy linear regression model with better explanatory power. *Fuzzy Sets and Systems*, 126 (3), 401-409.

Kao, C. and Chyu, C.-L. (2003). Least-squares estimates in fuzzy regression analysis. *European Journal of Operational Research*, 148 (2), 426-435.

Kao, C. and Lin, P.-H. (2005). Entropy for fuzzy regression analysis. *International Journal of Systems Science*, 36 (14), 869-876.

Kim, B. and Bishu, R. R. (1998). Evaluation of fuzzy linear regression models by comparing membership functions. *Fuzzy Sets and Systems*, 100 (1-3), 343-352.

Kominek, J. and Black, A. W. (2004). A family-of-models approach to hmm-based segmentation for unit selection speech synthesis. In *Proc. International Conference on Spoken Language Processing*, 1385-1388.

Kuo, J.-W. and Wang, H.-M. (2006). Minimum boundary error training for automatic phonetic segmentation. In *Proc. Interspeech.*, 1497- 1500.

Lee, K.-S. (2006). MLP-based phone boundary refining for a TTS database. *IEEE Transactions on Audio, Speech, and Language Processing*, 14 (3), 981-989.

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

Li, C.-L., Liu, J. and Xia, S.-H. (2007). English sentence stress detection system based on hmm framework. *Applied Mathematics and Computation*, 185 (2), 759-768.

Lo, H.-Y. and Wang, H.-M. (2007). Phonetic boundary refinement using support vector machine. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 4, 933-936.

Lu, J. and Wang, R. (2009). An enhanced fuzzy linear regression model with more flexible spreads. *Fuzzy Sets and Systems*, 160 (17), 2505-2523.

Lu, J., Wang, R. and De Silva, L. C. (2007). Review of current computer-assisted pronunciation teaching (CAPT) systems. Technical report, Http://www-ist.massey.ac.nz/rwang/thesis/CAPTreview.pdf.

Malfrère, F., Deroo, O., Dutoit, T. and Ris, C. (2003). Phonetic alignment: Speech synthesis-based vs. Viterbi-based. *Speech Communication*, 40 (4), 503-515.

Menzel, W., Herron, D., Morton, R., Pezzotta, D., P., B. and Howarth, P. (2001). Interactive pronunciation training. *ReCALL Journal*, 13 (1), 67-78.

Meszaros, K., Vitez L., Szabolcs I., et al. (2005). Efficacy of conservative voice treatment in male-to-female transsexuals. *Folia Phoniatrica Logopedics*, 57, 111-118.

Modarres, M., Nasrabadi, E. and Nasrabadi, M. M. (2005). Fuzzy linear regression models with least square errors. *Applied Mathematics and Computation*, 163 (2), 977-989.

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

Mporas, I., Ganchev, T. and Fakotakis, N. (2010). Speech segmentation using regression fusion of boundary predictions. *Computer Speech & Language*, 24 (2), 273-288.

Munro, M. J. and Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 49, 285-310.

Nagamine, T. (2002). An experimental study on the teachability and learnability of english intonational aspect: Acoustic analysis on F0 and native-speaker judgment task. *Journal of Language and Linguistics*, 1 (4), 362-399.

Nagano, K. and Ozawa, K. (1990). English speech training using voice conversion. In *Proc. International Conference on Spoken Language Processing*, 1169-1172.

Nasrabadi, M. M. and Nasrabadi, E. (2004). A mathematical-programming approach to fuzzy linear regression analysis. *Applied Mathematics and Computation*, 155 (3), 873-881.

Nasrabadi, M. M., Nasrabadi, E. and Nasrabady, A. R. (2005). Fuzzy linear regression analysis: A multi-objective programming approach. *Applied Mathematics and Computation*, 163 (1), 245-251.

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

Negrin-Cristiani, J. (1997). *Do non-native speakers of English acquire English stress patterns?* Retrieved May 1, 2009, from http://www.eric.ed.gov/.

Neri, A., Cucchiarini, C., Strik, H. and Boves, L. (2002). The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning*, 15 (5), 441-467.

Nolan, F. (2003). Intonational equivalence: An experimental evaluation of pitch scales. In *Proc. International Congress of Phonetic Sciences*, 771-774.

Orton, J. (2000). The teaching of rhythm: A key link in successful language classes. *Foreign Language Teaching Abroad*, 4, 1-7.

Ostendorf, M., Price, P. J. and Shattuck-Hufnagel, S. (1995). *The Boston University radio news corpus*. Boston: Boston University.

Özelkan, E. C. and Duckstein, L. (2000). Multi-objective fuzzy regression: A general framework. *Computers & Operations Research*, 27 (7-8), 635-652.

Park, S. S. and Kim, N. S. (2007). On using multiple models for automatic speech segmentation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15 (8), 2202-2212.

Pellom, B. L. and Hansen, J. H. L. (1998). Automatic segmentation of speech recorded in unknown noisy channel characteristics. *Speech Communication*, 25 (1-3), 97-116.

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

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.

Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of speech recognition*: Prentice Hall.

Rangarajan Sridhar, V. K., Bangalore, S. and Narayanan, S. S. (2008). Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. *IEEE Transactions on Audio, Speech, and Language Processing*, 16 (4), 797-811.

Redden, D. T. and Woodall, W. H. (1994). Properties of certain fuzzy linear regression methods. *Fuzzy Sets and Systems*, 64 (3), 361-375.

Sakawa, M. and Yano, H. (1992a). Fuzzy linear regression and its applications. In J. Kacprzyk & M. Fedrizzi (Eds.), *Fuzzy regression analysis* (pp. 61-80). Heidelberg: Physica-Verlag.

Sakawa, M. and Yano, H. (1992b). Multiobjective fuzzy linear regression analysis for fuzzy input-output data. *Fuzzy Sets and Systems*, 47 (2), 173-181.

Salvi, G. (2006). Segment boundary detection via class entropy measurements in connectionist phoneme recognition. *Speech Communication*, 48 (12), 1666-1676.

Scovel, T. (2000). A critical review of the critical period research. *Annual Review of Applied Linguistics*, 20, 213-223.

Seferoğlu, G. (2003). Recent trends in teaching pronunciation: To what extent is diversity acceptable? In *Proc. Multiculturalism in ELT Practices: Unity & Diversity, An International Conference*, 1-6.

Seferoğlu, G. (2005). Improving students' pronunciation through accent reduction software. *British Journal of Educational Technology*, 36 (2), 303-316.

Shakouri, H., Nadimi, R. and Ghaderi, F. (2007). Fuzzy linear regression models with absolute errors and optimum uncertainty. In *Proc. IEEE International Conference on Industrial Engineering and Engineering Management*, 917-921.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., et al. (1992). ToBI: A standard for labeling English prosody. In *Proc. International Conference on Spoken Language Processing*, 867-870.

Sluijter, A. M. C. and van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100, 2471-2485.

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. *Journal of the Acoustical Society of America,* 101, 503-513.

Solé Sabater, M. J. (1991). Stress and rhythm in English. *Revista Alicantina de Estudios Ingleses*, 4, 145-162.

Sundström, A. (1998). Automatic prosody modification as a means for foreign language pronunciation training. In *Proc. ISCA Workshop on Speech Technology in Language Learning*, 49-52.

Tamburini, F. and Caini, C. (2005). An automatic system for detecting prosodic prominence in American English continuous speech. *International Journal of speech technology*, 8 (1), 33-44.

Tanaka, H., Hayashi, I. and Watada, J. (1989). Possibilistic linear regression analysis for fuzzy data. *European Journal of Operational Research*, 40 (3), 389-396.

Tanaka, H., Uejima, S. and AsaL, K. (1982). Linear regression analysis with fuzzy model. *IEEE Transactions on Systems, Man and Cybernetics*, 12 (6), 903-907.

Taylor, P. (2000). Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, 107 (3), 1697-1714.

Todaka, Y. (1990). *An error analysis of Japanese students' intonation and its pedagogical applications.* University of California, Los Angeles.

Todaka, Y. (1995). A preliminary study of voice quality differences between Japanese and American English: Some pedagogical suggestions. *JALT Journal*, 17 (2), 261-268.

Toledano, D. T., Gomez, L. A. H. and Grande, L. V. (2003). Automatic phonetic segmentation. *IEEE Transactions on Speech and Audio Processing*, 11 (6), 617-625.

Tran, L. and Duckstein, L. (2002). Multiobjective fuzzy regression with central tendency and possibilistic properties. *Fuzzy Sets and Systems*, 130 (1), 21-31.

van Katwijk, A. (1974). *Accentuation in Dutch: An experimental linguistic study*. Amsterdam: Van Gorcum.

Wang, L., Zhao, Y., Chu, M., Zhou, J. and Cao, Z. (2004). Refining segmental boundaries for TTS database using fine contextual-dependent boundary models. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 641-644.

Wang, Q. (2008). L2 stress perception: The reliance on different acoustic cues. In *Proc. Conference on Speech Prosody*, 635-638.

Wennerstrom, A. (1994). Intonational meaning in English discourse - a study of nonnative speakers. *Applied Linguistics*, 15 (4), 399-420.

Wightman, C. (2002). ToBI or not ToBI. In *Proc. International Conference on Speech Prosody*.

Wightman, C. W. and Ostendorf, M. (1994). Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, 2 (4), 469-481.

Wightman, C. W. and Talkin, D. T. (1997). The aligner: Text-to-speech alignment using markov models. In J. P. H. Van Santen, R. W. Sproat, J. P. Olive & J. Hirschberg (Eds.), *Progress in speech synthesis* (pp. 313-323): Springer.

Witten, I. and Frank, E. (2005). *Data mining: Practical machine learning tools and techniques, 2nd ed*: Morgan Kaufmann.

Xie, H., Andreae, P., Zhang, M. and Warren, P. (2004a). Detecting stress in spoken English using decision trees and support vector machines. In *Proc. Australian Computer Science Communications*, 145-150.

Xie, H., Andreae, P., Zhang, M. and Warren, P. (2004b). Learning models for English speech recognition. In *Proc. Australian Computer Science Communications*, 323-330.

Yang, M.-S. and Lin, T.-S. (2002). Fuzzy least-squares linear regression analysis for fuzzy input-output data. *Fuzzy Sets and Systems*, 126 (3), 389-399.

Yavas, M. (2006). *Applied English phonology*: Oxford: Blackwell.

Yoon, K. (2008). Synthesis and evaluation of prosodically exaggerated utterances: A preliminary study. In *Proc. Spring Conference of the Association of Modern British & American Language & Literature*.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., et al. (2006). *The HTK book*. Cambridge, U.K: Cambrige University.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8 (3), 338-353.

Zimmermann, H. J. (1991). *Fuzzy set theory and its application*. Dordrecht: Kluwer Academic Press.

Publications Related to This Research

Published papers

1. Lu, J. and Wang, R. (2009). An enhanced fuzzy linear regression model with more flexible spreads. *Fuzzy Sets and Systems*, 160 (17), 2505-2523.

2. Wang, R. and Lu, J. Investigation of the golden speakers for second language learners from the imitation preference perspective. *Speech Communication*. (Accepted)

3. Lu, J., Wang, R., De Silva, L. C. and Gao, Y. (2009). Syllable nucleus durations estimation using linear regression based ensemble model. In Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Taipei, Taiwan, 4849-4852.

4. Lu, J., Wang, R., De Silva, L. C., Gao, Y. and Liu J. (2010). CASTLE: a computer-assisted stress teaching and learning environment for learners of English as a second language. *INTERSPEECH*, Makuhari, Japan. (Accepted)

Submitted papers

5. Lu, J., Wang, R., De Silva, L. C., and Gao, Y. Automatic stress exaggeration to assist language learners perceive sentence stress. *ACM - Transactions on Speech and Language Processing*.

Appendix D

MASSEY UNIVERSITY
Application for Approval of Request to Embargo a Thesis
(Pursuant to AC98/168 (Revised 2), Approved by Academic Board 17/02/99)

Name of Candidate: Jingli LU ID Number: 05182816

Degree: PhD Dept/Institute/School: SEAT

Thesis title:

CASTLE: a Computer-Assisted sentence Stress Teaching and Learning Environment

Name of Chief Supervisor: Ruili Wang Telephone Ext: 2548

As author of the above named thesis, I request that my thesis be embargoed from public access until (date): 01/06/2012 for the following reasons:

☒ Thesis contains commercially sensitive information.

☐ Thesis contains information which is personal or private and/or which was given on the basis that it not be disclosed.

☐ Immediate disclosure of thesis contents would not allow the author a reasonable opportunity to publish all or part of the thesis.

☐ Other (specify):

Please explain here why you think this request is justified:

This research project is supported by TIF (Technology for Industry Fellowships) of the Foundation of Research, Science & Technology, which is conducted cooperatively by Massey University and Matahoo Ltd. This thesis contains the core technology to develop a commercial pronunciation learning systems for learners of English as a second language.

Signed (Candidate): Jingli Lu Date 22/06/2010

Endorsed (Chief Supervisor): Ruili Wang Date 22/06/2010

Approved/Not Approved (Representative of VC): Date 23/6/2010

Note: Copies of this form, once approved by the representative of the Vice-Chancellor, must be bound into every copy of the thesis.