Accent Conversion via Formant-based Spectral Mapping and Pitch Contour Modification

DANG CONG ZHENG

A thesis submitted in partial fulfilment of requirements of the University of Wolverhampton for the degree of Master of Philosophy

June 2011

This work or any part thereof has not previously been presented in any form to the University or any other body whether for the purposes of assessment, publication or for any other purpose (unless previously indicated). Save for any express acknowledgements, references and/or bibliographies cited in the work, I confirm that the intellectual content of the work is the result of my own efforts and of no other person.

The right of Dang Cong Zheng to be identified as author of this work is asserted in accordance with ss.77 and 78 the Copyright, Designs and Patents Act 1998. At this data copyright is owned by the author.

Signature……………………………………
Date…………………………………………
Abstract

Accent conversion intends to change the accent of a speaker to a desired accent and preserve the speaker’s voice identity. This technology can offer a number of useful applications. For example, integrating accent conversion to a text-to-speech system (TTS) can produce a voice with a desired accent instantly and inexpensively. Applying the technology to the film industry can change an actor’s or actress’s accent to a desired accent without hard training for the actor or actress to learn a new accent; this can be achieved by modifying the accent of the film recordings. As a foreign language learning tool, it could allow the learners to listen to their own voice with the native speaker’s accent and to mimic that accent. Hence, enhance learning experience and improve learning progress.

In this dissertation, a new approach in both accent analysis and conversion has been proposed. In contrast to previous approaches in accent-related research, such as in regional or foreign accent classification and identification, where the databases are formed from large groups of single-accent speakers, this study uses data from an individual who can speak in two accents. This removes the effects of inter-speaker variability and facilitates efficient identification and analysis of acoustic features of different accents.

Two British regional accents which display distinct differences to the human listener were used in this study as two typical British regional accents. Vowel based acoustic analysis was carried out to investigate the acoustic characteristics of the two accents and identify the prominent features that are most influential on the variability of accents. Acoustic characteristics such as formant frequencies, fundamental frequency and its variation slope, intensity of speech, and duration of phone were used for accent acoustic analysis.

In this dissertation, accent conversion via formants modification and pitch contour manipulation was investigated. Three different formant-based spectral mapping
algorithms, mean-variance linear conversion, $N^{th}$ order non-linear conversion and piece-wise linear transformation based on Gaussian mixture model conversion were investigated. Furthermore, the project has implemented accent conversion on a general speech analysis and synthesis system; the output speech synthesized by the three mapping algorithms was assessed by objective and subjective evaluation. The effects of spectral conversion and pitch contour conversion on accent conversion were also evaluated.

The results of the study showed that accent conversion can be achieved to some degree via formants modification and pitch contour manipulation.
# Table of Contents

Abstract .................................................................................................................. 2
Table of Contents .................................................................................................... 4
Acknowledgements ............................................................................................... 6
List of Figures ......................................................................................................... 7
List of Tables .......................................................................................................... 9
List of Abbreviations ............................................................................................ 11
Chapter 1 Introduction .......................................................................................... 12
  1.1 Objective ........................................................................................................ 14
  1.2 Structure of the thesis .................................................................................... 14
Chapter 2 Speaker individuality ........................................................................... 16
  2.1 Acoustic characteristic of speaker individuality ........................................... 16
  2.2 Speaker individuality based on accent .......................................................... 17
    2.2.1 Constitutes of an accent .......................................................................... 18
    2.2.2 Acoustic characteristic of accents .......................................................... 23
  2.3 Birmingham accent and Liverpool accent ..................................................... 25
Chapter 3 Speech analysis and synthesis ............................................................... 28
  3.1 Source filter model of speech ....................................................................... 28
  3.2 Linear prediction (LP) speech analysis and synthesis ................................... 31
Chapter 4 Acoustic analysis of accents ................................................................. 35
  4.1 Speech data preparation and recording ......................................................... 35
  4.2 Vowel-based acoustic analysis of accents .................................................... 37
    4.2.1 Vowels extraction .................................................................................. 38
    4.2.2 Acoustic features extraction and analysis .............................................. 40
      4.2.2.1 Formant frequencies extraction and analysis .................................... 40
      4.2.2.2 Fundamental frequency extraction and pitch slope analysis ............ 47
      4.2.2.3 Intensity analysis ............................................................................ 49
      4.2.2.4 Duration analysis .......................................................................... 51
Table of contents

4.3 Conclusions .............................................................................53

Chapter 5 Accent conversion .......................................................55
  5.1 Training materials .................................................................58
  5.2 Spectral mapping algorithms .................................................62
    5.2.1 Mean-variance linear conversion function .......................62
    5.2.2 Nth order conversion function .......................................65
    5.2.3 GMM conversion function .............................................66
  5.3 Spectral conversion ...............................................................72
  5.4 Pitch contour modification ....................................................73
    5.4.1 Time alignment ..............................................................75
      5.4.1.1 Dynamic time warping ............................................75
      5.4.1.2 Pitch synchronous overlap and add .........................80
    5.4.2 Pitch contour modification .............................................84

Chapter 6 Experiments and evaluations .....................................86
  6.1 Spectral mapping without pitch contour modification .............87
  6.2 Spectral mapping with pitch contour modification .................87
  6.3 Objective evaluation ..........................................................87
  6.4 Subjective evaluation ..........................................................89
  6.5 Conclusions ........................................................................96

Chapter 7 Conclusions and future work ...................................97
  7.1 Conclusions ........................................................................97
  7.2 Future work .........................................................................98

References .................................................................................100

Appendix I Text material used in this study .............................106
Appendix II Twelve vowels and the words from which the vowels
  were extracted .......................................................................108
Acknowledgements

Firstly, I would like to thank Dr. D. W. Dyke, Dr. F. Berryman and Dr. C. Morgan for their support and guidance throughout this MPhil research. I would also like to acknowledge the Department of Engineering and the School of Technology for their financial support.

I would like to thank all of those who have devoted their time and patience for speech recording and perceptual testing.

Finally, I would sincerely like to thank my husband C. J. Wang and son K. D. Wang for their support and encouragement.
List of Figures

Figure 3.1 Organs of human speech production system .................. 29
Figure 3.2 Source filter model of speech ................................. 30
Figure 3.3 Waveform of a segment of voiced speech .................... 33
Figure 3.4 The residual of the speech from 16\textsuperscript{th} LP analysis ... 33
Figure 3.5 Short-time FFT spectrum of the speech superimposed with a 16\textsuperscript{th} order LP envelope ............................................. 34
Figure 4.1 Histogram of phoneme content in the speech database .... 36
Figure 4.2 Example of manual segmentation of the phrase “more like” with software Praat ................................................................. 39
Figure 4.3 Mean values of first three formants of twelve vowels over 10 repeated utterances ......................................................... 44
Figure 4.4 F\textsubscript{2}-F\textsubscript{1} formant space of twelve vowels in accent BM and LP .................................................................................. 45
Figure 4.5 F\textsubscript{3}-F\textsubscript{2} formant space of twelve vowels in accent BM and LP .................................................................................. 45
Figure 4.6 The pitch slopes of twelve vowels over 10 repeated utterances ....................................................................................... 49
Figure 4.7 The mean value of intensity of twelve vowels .......... 51
Figure 4.8 Durations of twelve vowels in Accent BM and Accent LP .............................................................................................. 53
Figure 5.1 Flow diagram of accent conversion framework .......... 58
Figure 5.2 Histogram of $\Delta f$ values for the two accents ............ 61
Figure 5.3 Probability distribution of $\Delta f$ ................................. 61
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4</td>
<td>GMM regression line along with mean-variance linear approximation and quadratic regression for formant F1</td>
<td>70</td>
</tr>
<tr>
<td>5.5</td>
<td>GMM regression line along with mean-variance linear approximation and quadratic regression for formant F2</td>
<td>71</td>
</tr>
<tr>
<td>5.6</td>
<td>GMM regression line along with mean-variance linear approximation and quadratic regression for formant F3</td>
<td>71</td>
</tr>
<tr>
<td>5.7</td>
<td>The pitch contour of sentence “Will we ever forget it?” generated by a speaker in Accent BM (top) and Accent LP (bottom)</td>
<td>74</td>
</tr>
<tr>
<td>5.8</td>
<td>The pitch contour of sentence “There was a change now.” generated by a speaker in Accent BM (top) and Accent LP (bottom)</td>
<td>74</td>
</tr>
<tr>
<td>5.9</td>
<td>The distance matrix of MFCCs and the optimal warp path</td>
<td>78</td>
</tr>
<tr>
<td>5.10</td>
<td>A short segment of two time aligned signals</td>
<td>78</td>
</tr>
<tr>
<td>5.11</td>
<td>The waveforms of the sentence “The nightglow was treacherous to shoot by.” in Accent BM and Accent LP</td>
<td>79</td>
</tr>
<tr>
<td>5.12</td>
<td>The waveforms of two time aligned utterances using DTW method</td>
<td>80</td>
</tr>
<tr>
<td>5.13</td>
<td>Duration modification of a voiced speech segment</td>
<td>82</td>
</tr>
<tr>
<td>5.14</td>
<td>The waveforms of two time aligned utterances using PSOLA method</td>
<td>83</td>
</tr>
<tr>
<td>5.15</td>
<td>Pitch contours of sentence “there was a change now” spoken by a female speaker in Accent BM (black) and Accent LP (green) and converted pitch contour (red)</td>
<td>85</td>
</tr>
</tbody>
</table>
List of Tables

Table 2.1 Comparison of phoneme set used by CMU dictionary and BEEP.................................................................22
Table 4.1 Mean and standard deviation of the first formant F1 and its t-test of twelve vowels over 10 repeated utterances...............42
Table 4.2 Mean and standard deviation of the second formant F2 and its t-test of twelve vowels over 10 repeated utterances...............43
Table 4.3 Mean and standard deviation of the third formant F3 and its t-test of twelve vowels over 10 repeated utterances...............44
Table 4.4 The pitch slopes of twelve vowels over 10 repeated utterances........................................................................48
Table 4.5 Mean and standard deviation of intensity and its t-test of twelve vowels over 10 repeated utterances.........................50
Table 4.6 Mean and standard deviation of duration and its t-test of twelve vowels over 10 repeated utterances.........................52
Table 5.1 Comparison of the average prediction error of Gaussian mixture model with different mixture numbers..................69
Table 5.2 Comparison of the average prediction error of the three different approximation methods....................................72
Table 6.1 The normalised Euclidean distance of formant F1 between target utterance and converted utterance........................88
Table 6.2 The normalised Euclidean distance of formant F2 between target utterance and converted utterance........................89
Table 6.3 The normalised Euclidean distance of formant F3 between target utterance and converted utterance........................89
Table 6.4 Evaluation of the effect of the algorithm used by Praat software on the quality of the synthesized utterance..................91
Table 6.5 Evaluation of the three spectral mapping algorithms on accent conversion.........................................................93
Table 6.6 Evaluation of the effect of pitch contour conversion on accent conversion..........................................................94
Table 6.7 Evaluation of accent model applied on a different speaker.................................................................................95
Appendix II Table 1 Twelve vowels and the words from which the vowels were extracted.................................................108
### List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>Artificial neural network</td>
</tr>
<tr>
<td>ASR</td>
<td>Automatic speech recognition</td>
</tr>
<tr>
<td>BEEP</td>
<td>British English example pronunciation dictionary</td>
</tr>
<tr>
<td>CMU</td>
<td>Carnegie Melon University</td>
</tr>
<tr>
<td>DFW</td>
<td>Dynamic frequency warping</td>
</tr>
<tr>
<td>DTW</td>
<td>Dynamic time warping</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian mixture model</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov model</td>
</tr>
<tr>
<td>LP</td>
<td>Linear prediction</td>
</tr>
<tr>
<td>LPCC</td>
<td>Linear predictive coding coefficient</td>
</tr>
<tr>
<td>LSP</td>
<td>Line spectral pair</td>
</tr>
<tr>
<td>MAP</td>
<td>Maximum a posteriori</td>
</tr>
<tr>
<td>MFC</td>
<td>Mel-frequency cepstral</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel-frequency cepstral coefficient</td>
</tr>
<tr>
<td>MOS</td>
<td>Mean opinion score</td>
</tr>
<tr>
<td>PSOLA</td>
<td>Pitch synchronous overlap and add</td>
</tr>
<tr>
<td>STFT</td>
<td>Short time Fourier transform</td>
</tr>
<tr>
<td>STFFT</td>
<td>Short time fast Fourier transform</td>
</tr>
<tr>
<td>TTS</td>
<td>Text-to-speech</td>
</tr>
<tr>
<td>VQ</td>
<td>Vector quantization</td>
</tr>
<tr>
<td>VTLN</td>
<td>Vocal tract length normalization</td>
</tr>
</tbody>
</table>
Speech communication is very important in daily life and it is one of the most convenient, direct ways for people to communicate. Speech signals not only convey the linguistic information (the message) but also other information such as emotion, attitude, and speaker’s individuality which includes: gender, age, social and regional origin.

Speech synthesis has a long history. The earliest speech synthesizer can be dated back to the eighteenth century. In 1779, Kratzenstein, a Russian, constructed acoustic resonators similar to the human vocal tract and activated the resonators with vibrating reeds producing synthetic sound of five long vowels (/a/, /e/, /i/, /o/ and /u/). Since then, speech synthesis developed rapidly, especially in recent years. From inexpensive software programs for home computers to reading machines for the vision-impaired people, the use of the speech synthesizer is becoming more and more widespread. A text-to-speech (TTS) system is one of the human-machine interfaces using speech and this technology has developed rapidly in recent years; it has been used in many applications such as a car navigation system, information retrieval over the telephone, voice mail, and electronic dictionaries. The goal of the TTS system is to produce synthetic speech with natural human voice characteristics, and furthermore, with various speaker individualities such as accent, speaking style and different emotions.

Over the last two decades, many attempts have been carried out to incorporate emotions in synthesized speech, to improve the naturalness of synthetic speech and to make it more like a human voice [1-4]. Several prototype synthetic speech-with-emotion systems have been developed [5]. For example, the HAMLET system of Murray and Arnott (1995), Affect Editor System of Cahn (1990) and SPRUCE text to speech system of Tatham and Lewis (1992).
Most current TTS systems are based on a concatenative synthesis method due to its high-quality synthesized speech output. Unfortunately, concatenative synthesizers are usually limited to one speaker and one voice and usually require large amounts of memory to store all the phrases to be synthesized. Though some concatenative synthesizers use more than one speech corpus to produce synthetic speech with different voices, it is still limited for producing a customised voice. This is because the creation of a new synthesis database is very time-consuming and expensive and involves a huge amount of recording and labelling effort. Making a speech synthesis system produce various voices without building expensive different databases is still a big challenge. Some researchers have been making efforts to apply voice conversion techniques to TTS to create a new synthetic voice without creating a large speech database [6-8]. However, the new voices produced by voice conversion techniques are totally different from the original voice which changes the source speaker’s identity completely. Accent conversion technology could allow a TTS user to switch the accent of synthetic speech to a desired accent without losing the original speaker’s voice identity.

Accent plays an important role in daily communication and it is one of the most fascinating aspects of the acoustics of speech signals. The accent of a speaker reflects regional affiliation of the speaker; it states a speaker’s community characteristics [9]. People from different regions have a different local accent. Nowadays, electronic speech synthesizers are widely used and become more and more customized demanding. Accent modelling and synthesis is no longer regarded as a luxury for a speech synthesizer, and a multi-accent speech synthesis system can be good in some applications. For example, in telephone information systems it might be useful to have the system respond to a caller in the caller’s own accent; it would also be helpful if a reading machine for those with sight impairment could switch accents depending on the material being read aloud or the accent of the person using the system.
Chapter 1  Introduction

1.1 Objectives

This project aims to investigate accent conversion via formant-based spectral mapping and pitch contour manipulation. It will investigate acoustic characteristics of two British regional accents – the Birmingham accent and the Liverpool accent. The most accent influenced acoustic features will be identified and modified to implement accent conversion.

To achieve these objectives, the following work has been undertaken:

- Build a speech corpus which includes twenty-five short sentences and one short paragraph, uttered by a female participant in two different accents: Birmingham accent and Liverpool accent
- Identify the prominent acoustic features that are most influential on the variability of the two accents from vowel-based acoustic analysis
- Implement accent conversion via formant-based spectral mapping and pitch contour modification
- Assess the three spectral mapping methods by means of objective and subjective evaluation
- Evaluate the effect of pitch conversion on the accent conversion by subjective testing
- Apply the accent conversion model which was built from the training database to a different speaker to perform accent conversion and assess the outputs by means of subjective testing

1.2 Structure of the thesis

This thesis is divided into seven chapters with Chapter 1 being the introduction; it provides the background to this thesis such as the objective, methodology and the structure of the thesis.
Chapter 2 gives an introduction to speaker individuality, accent as one aspect of speaker individuality, and a literature review on the current accent-related research.

Chapter 3 reviews the theory of the source-filter model of speech and linear prediction (LP) speech analysis and synthesis which were applied throughout this study.

In Chapter 4 acoustic characteristics of accents were analysed with twelve specified vowels and several acoustic features. The analysis was based on two British regional accents: the Birmingham accent and the Liverpool accent.

Chapter 5 studies existing spectral mapping techniques, introducing and investigating the three methods for formants modification and pitch transplantation techniques for pitch conversion.

Chapter 6 reports on a series of experiments to produce synthesized speech and a series of objective and subjective evaluations.

Chapter 7 concludes the thesis and points out some future work.

After the main body of the text there are a list of references and a number of appendices. Appendix I gives the text materials used in the study. Appendix II includes a table listing the twelve vowels used in Chapter 2 and the words from which the vowels were extracted.
Chapter 2  
Speaker individuality

In this chapter, the acoustic characteristics of speaker individuality and speaker individuality based on accent are introduced. This thesis studies regional accent based on two specific British regional accents: the Birmingham accent and the Liverpool accent. British regional accents are introduced in section 2.2.1. The state of the current accent-related research is also covered in this chapter.

2.1 Acoustic characteristics of speaker individuality

Speaker individuality, also known as speaker identity, is the property of speech that allows one speaker to be distinguished from another. Many factors contribute to voice individuality; Kuwabara and Sagusaka [10] summarized the acoustic correlates of speech individuality based on other scientific findings [11-13]. They concluded that the acoustic parameters that are thought to have the most influence on voice individuality are the following:

Voice source characteristics:
- the average pitch frequency
- the pitch contour – the pattern of pitch frequency in time domain
- the pitch frequency fluctuation
- the glottal wave shape

Vocal tract characteristics:
- the shape of spectral envelope and spectral tilt
- the absolute values of formants frequencies
- the formant trajectories – the tracks of variations of formants frequencies
- the long-term average speech spectrum
- the formant bandwidth
Earlier studies show that there are many factors such as age, gender, height, weight and other physical properties that contribute to voice individuality of a speaker, and there is no single acoustic parameter that plays a decisive role in voice individuality. However, different acoustic parameters play a weighted contribution in voice individuality. Based on vowel samples from a group of male speakers, Matsumoto et al. [14] concluded that the fundamental frequency F0 was the most important acoustic parameter for the perception of the speaker’s voice quality with the formant frequencies the next most important followed by F0 fluctuation and spectral tilt of the glottal source. They also found that the contribution of the formants frequencies to voice quality perception varied according to the kind of vowel, whereas the contribution of the glottal source characteristics was vowel independent.

### 2.2 Speaker individuality based on accent

Previous research [10, 15] suggests that the factors that are relevant to voice individuality can be categorized in terms of sociology and physiology. Characteristics such as speaking accent and speaking style are developed by individual speakers through family, schools and community neighbourhoods; this kind of characteristic is socially conditioned, usually depending on factors such as social status, dialect, and the community to which the speaker belongs. The quality of voice is mainly determined by the physiological and anatomical properties of the speech organs, such as the overall dimension of the speaker’s vocal tract, the relative proportions between the various cavities in the tract and the properties of the vocal chords.

Accent as one of speakers’ individual characteristics, is something every speaker has. According to Wells [9], the term ‘accent’ refers to “a pattern of pronunciation used by a speaker for whom English is the native language or, more generally, by the community or social grouping to which he or she belongs”. He also pointed out that “to a very much degree accent is characteristic of people belonging to
some geographical region and /or social class; and it may well be typical of the speaker’s, age group, or level of education”. Accent is also different from dialect. The term ‘dialect’ refers to the whole speech pattern, including any differences in pronunciation, grammar and vocabulary among varieties of the same language, while accent refers to differences in pronunciation, a way of people saying the same words [9, 16].

Since accent is sociologically related, it is likely to be manifested more in conversation between friends with the same social and linguistic background than in conversation between strangers, or in formal, read speech. However, accent perception and how far a listener can categorize an accent depend upon the listener’s linguistic background and familiarity or prior knowledge of particular accents [9]. For example, it is easy for most English people to identify the region of geographic origin of a speaker of UK English, even if only to “northern” or “southern”, it might not be so easy, however, for an American or a non-native English speaker. Ikeno and Hansen [17] investigated how the listener’s own accent background affects accent perception and comprehensibility. In their experiment, they asked 33 listeners with different accent backgrounds to classify three types of native UK English accent: English, Irish and Welsh. The 33 listeners consisted of 11 British, 11 US and 11 non-native English speakers. Their experimental results showed that British listeners performed with the highest accuracy (83%), the classification accuracy of US listeners (56%) was significantly lower than that of British listeners, and the non-native English speakers group had the lowest classification accuracy (45%). This suggests that the listener’s linguistic background impacts on their ability to categorize accents. It is also difficult to agree the degree of the strength of that accent between different judges [18].

2.2.1 Constitutes of an accent

Different accents of language have distinct patterns of pronunciations and
intonation. Accents can systematically differ in the following ways [9]:

a) Differences in the phonetic realizations of accents: for the same phoneme, each accent group may have different phonetic realization rules. For example, the vowel in the word “coat” is realized as a monophthongal [o:] or shorter [o] in some accents such as in Scotland, but it is realized as a wider diphthong [ʌʊ] in others, such as in the south-east of England. Phonetic realization differences between accents can also be context-sensitive. For example, in a Canadian accent, the diphthong [ai] is used before a voiceless consonant such as in the words “nice” and “write”, and the diphthong [at] is used in other environments, the same as in most other accents. Phonetic realization differences between accents also involve consonants. For examples, aspirated consonants /p, t, k/, in the north of England and Scotland are never aspirated. An intervocalic /r/ in words such as “very, sorry, arrow” is used in most accents, but in some accents such as the working-class accent of Liverpool, an alveolar tap [ɹ] is used instead.

b) Differences in the phonemic systems of accents: phonemic systems can be different in the number and or identity of the phonemes in the system. For example, most accents of English have two distinct vowel phonemes in the close back area, a short /u/ as in “foot” and a long /u:/ as in “boot”. But in Scottish English there is only a single phoneme /u/, which corresponds to these two phonemes of other accents. British English, American English, and Australian English have different phonetic transcription systems. Yan *et al.* [19] gave an example of different phonetic transcription in British, Australian and American accents. The British accent, as transcribed by Cambridge University’s British English example pronunciation dictionary (BEEP), has five extra vowels, /ax, ea, ia, ua, and oh/, compared with the American accent as transcribed by Carnegie Melon University’s CMU
dictionary. Australian English has distinctive vowels such as /æi/ instead of /ai/ and /eə/ for /au/. Table 2.1 shows a comparison between the phoneme set used by Carnegie Melon University’s CMU dictionary and the phoneme set used by BEEP. Words “about, pair, pear, poor, pot” were transcribed as “AX B AW T/ P EA R/ P IA/ P UA/ P OH T” in BEEP, and “AH B AW T/ P EH R/ P EH R/ P UH R/ P AAT” in CMU dictionary respectively.

c) Differences in the phonotactic distribution: accents may also differ in the environments in which particular phonemes do or do not occur. For example, English accents can be divided into rhotic accent and non-rhotic accent depends on the phonotactic distribution of the consonant /r/. In the rhotic accents such as Scottish and General American, the consonant /r/ can occur before a consonant or at the end of a word, such as in the words “farm [færm], far [fɑːr]”. While in the non-rhotic accents such as RP and Australian, the consonant /r/ does not occurs in these environments, thus “farm” and “far” realized as [fɑːm] and [fɑː] respectively.

d) Differences in the lexical distribution: accents also differ in the phoneme selection for lexical representation of particular words. For example, some people use /iː/ to represent the first vowel in the word “either”, whereas other people use /ai/ instead.

e) Differences in the rhythmical characteristics of accents: apart from the differences in phonetic realizations, phonemic systems, phonotactic and lexical distributions, accents also differ in their rhythmical characteristics, such as syllable boundary (e.g. self $ ish vs sel $ fish, “$” indicates boundaries), rate of speech, lexical stress patterns (e.g. 'exquisite vs ex'quisite, “‘” indicates position of stress).

Accent is not only correlated with geographical variations; it is also related to sex, age, and social class [9]. Watt and Milroy [20] studied the distribution of vowel variants in relation to the speaker’s age, social class and gender. Their study was
based on analysis of three Newcastle vowels: the FACE, GOAT and NURSE vowels. The results of their study showed that vowel variants between different age group, social class and gender had different patterns. For example, for FACE vowel, 36% of the younger working class male using [ɪə] instead of [e] compared to 63% of the old working class male. There were 22% of the older middle class male and 14% of younger middle class male using [ɪə] respectively. Female use of the centring diphthong [ɪə] whatever the social class, however, was very rare.
Table 2.1  Comparison of phoneme set used by CMU dictionary and BEEP

<table>
<thead>
<tr>
<th>Phonemes in CMU</th>
<th>Phonemes in BEEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>odd</td>
</tr>
<tr>
<td>AE</td>
<td>at</td>
</tr>
<tr>
<td>AH</td>
<td>hut</td>
</tr>
<tr>
<td>AO</td>
<td>ought</td>
</tr>
<tr>
<td>AW</td>
<td>cow</td>
</tr>
<tr>
<td>AY</td>
<td>hide</td>
</tr>
<tr>
<td>B</td>
<td>be</td>
</tr>
<tr>
<td>CH</td>
<td>cheese</td>
</tr>
<tr>
<td>D</td>
<td>Dee</td>
</tr>
<tr>
<td>UH</td>
<td>thee</td>
</tr>
<tr>
<td>EH</td>
<td>Ed</td>
</tr>
<tr>
<td>ER</td>
<td>hurt</td>
</tr>
<tr>
<td>EY</td>
<td>ate</td>
</tr>
<tr>
<td>F</td>
<td>fee</td>
</tr>
<tr>
<td>G</td>
<td>green</td>
</tr>
<tr>
<td>HH</td>
<td>he</td>
</tr>
<tr>
<td>IH</td>
<td>it</td>
</tr>
<tr>
<td>IY</td>
<td>eat</td>
</tr>
<tr>
<td>JH</td>
<td>gee</td>
</tr>
<tr>
<td>K</td>
<td>key</td>
</tr>
<tr>
<td>L</td>
<td>lee</td>
</tr>
<tr>
<td>M</td>
<td>me</td>
</tr>
<tr>
<td>N</td>
<td>knee</td>
</tr>
<tr>
<td>NG</td>
<td>ping</td>
</tr>
<tr>
<td>OW</td>
<td>oat</td>
</tr>
<tr>
<td>OY</td>
<td>toy</td>
</tr>
<tr>
<td>P</td>
<td>pee</td>
</tr>
<tr>
<td>R</td>
<td>read</td>
</tr>
<tr>
<td>S</td>
<td>sea</td>
</tr>
<tr>
<td>SH</td>
<td>she</td>
</tr>
<tr>
<td>T</td>
<td>tea</td>
</tr>
<tr>
<td>TH</td>
<td>theta</td>
</tr>
<tr>
<td>UH</td>
<td>hood</td>
</tr>
<tr>
<td>UW</td>
<td>two</td>
</tr>
<tr>
<td>V</td>
<td>vee</td>
</tr>
<tr>
<td>W</td>
<td>we</td>
</tr>
<tr>
<td>Y</td>
<td>yield</td>
</tr>
<tr>
<td>Z</td>
<td>zee</td>
</tr>
<tr>
<td>ZH</td>
<td>seizure</td>
</tr>
<tr>
<td>AX</td>
<td>about</td>
</tr>
<tr>
<td>EA</td>
<td>pair</td>
</tr>
<tr>
<td>IA</td>
<td>peer</td>
</tr>
<tr>
<td>UA</td>
<td>poor</td>
</tr>
<tr>
<td>OH</td>
<td>pot</td>
</tr>
</tbody>
</table>
2.2.2 Acoustic characteristic of accents

Accent-related speech acoustic characteristics research has been going on for many years, and across many different languages. For example, Clopper et al. [21] studied acoustic characteristics of eleven vowel tokens produced by 48 speakers from six different regions of the United States. In their studies, acoustic measures of duration, first and second formant frequencies of eleven vowels have been used for analysis. Their statistical analysis of variance (ANOVA) results revealed that the effects of dialect differences on vowel duration are not consistent across all eleven vowels and there is a significant effect of vowel category and dialect on formants F1 and F2. Adank et al. [22] studied regional acoustic varieties in the vowel system of northern and southern standard Dutch from fifteen vowels produced by 160 speakers across four regions in the Netherlands and four regions in Flanders. Their analysis was based on acoustic parameters of duration and the first two formants frequencies of vowels. Their results showed that vowel duration varied for different vowels across both communities; measurements of the first two formant frequencies F1 and F2 at 50% of the vowel duration and their changes $\Delta F1$, $\Delta F2$ at 25% and 75% of the vowel duration also showed variations between communities and within communities. Another interesting outcome from their analysis was that more differences occurred for F2 than for F1 at 50% measurements in the both communities.

With computing technology rapidly developing in areas such as automatic speech recognition (ASR) and TTS technology, accent-related research such as accent classification and identification attracted more researchers’ interests. Arslan and Hansen [23] developed an algorithm based on isolated word and phoneme level to classify American English (AE) spoken in four variants: (i) received AE (i.e. without a foreign accent), (ii) AE with a Turkish accent, (iii) AE with a Chinese accent, and (iv) AE with a German accent. Their studies discovered that the first two formants F1 and F2 were useful parameters in accent classification and other prosodic features such as pitch contour (intonation) were shown to vary
significantly from one accent to another accent. Their experimental results also indicated that vowels carry more information than consonants in accent assessment. Zheng et al. [24] combined Mel-frequency cepstral coefficients (MFCCs) and the first three formants and their amplitudes as the accent related features for Shanghai-accented Mandarin detection; they subsequently incorporated these parameters into a speech recognition system to improve the recognition performance. Yan et al. [19] built a two-dimensional hidden Markov model (HMM) of the first two formants and a Rise/Fall/Connect (RFC) pitch model for modelling and estimating the acoustic differences between British, Australian and American accents. In their study, they noted that the phoneme duration pattern and the speaking rate also vary between the different accents and the second formant F2 is mostly affected by accent. Ghorshi et al. [25] highlighted the differences in spectral features between British, Australian and American English accents through comparison of the cross-entropies of formants and cepstrum features (MFCCs). Their studies indicated that formants are a better indicator of the three accents than MFCCs cepstrum features. Angkititrakul and Hansen [26] employed trajectory models to capture the phone temporal structure of Chinese, French, Thai, and Turkish accents in English. Huckvale [27] introduced a new metric ACCDIST (accent characterisation by comparison of distances in the inter-segment similarity table) for the quantitative assessment of the similarity of speaker’s accents. The ACCDIST metric was based on the correlations of inter-segment distance table across speakers or accent groups. The distance tables indicated the similarity of pronunciation of the same vowel in different words for a particular speaker. Since the ACCDIST was based on segment similarity within a speaker, this ensured that it was sensitive to the speaker’s pronunciation system rather than to his or her voice characteristics. The proposed metric achieved 89% recognition rate in an accent classification of 14 English regional accents of the British Isles.

Compared to research carried out in accent classification, accent conversion and synthesis are still relatively under-explored. There are several research papers
addressing accent conversion and synthesis. Felps et al. [28] proposed a voice transformation technique to convert a foreign accent to a native accent by means of pitch-synchronous decomposition of speech into glottal excitation and spectral envelope. Their accent-conversion transformation was based on the general framework of pitch-synchronous overlap and add (PSOLA) in the Fourier domain. They also proposed a pedagogical strategy of integrating accent-conversion as a form of behavioural shaping in computer assisted pronunciation training. Yamagishi et al.[29] created thousands of synthetic voices in a framework of an HMM-based speech synthesis system. Firstly, they built an “average voice model” from a number of speech corpora which include a large number of speakers, and then the model adaptation technique was applied to the “average voice model” to produce new synthetic voices. Huckvale and Yanagisawa [30] constructed a spoken language conversion system (English to Japanese), and used application of accent morphing technique to improve the intelligibility of converted speech. Their accent morphing was implemented by interpolating a series of acoustic parameters such as segmental feature LSP, pitch and rhythm between the source utterance and the target utterance. Their experiments showed that the word error rate dropped from 43% to 16% when the accent morphing technique was used.

2.3 Birmingham accent and Liverpool accent

Among native British English speakers, many different accents exist. Geographically, the UK consists of four parts, namely: England, Scotland, Wales and Northern Ireland. People living in the different parts have their own distinct cultural identity, which includes their regional accents. According to the UK’s geography, British English can be broadly classified into four main accents; England English, Scottish, Welsh and Northern Irish. However there is also significant variation within each of these four regions. England in particular has many different accents.
Birmingham accent and Liverpool accent, two of the best known regional accents of north England English have been used in this study. According to Wells [9], Birmingham and Liverpool accents are both North accents; they have some common features that North accents have, such as the absence of FOOT-STRUT split (use FOOT vowel /ʊ/ in STRUT words), absence of BATH broadening (use /a/ instead of /ɑː/), short /o/ in CLOTH words. They also have different characteristics of their own. For example, in Birmingham speech, oppositions between vowel /ʊ/ and /ʌ/, /ɒɪ/ and /ɔɪ/ exist, [ɒɪ] used instead of [ɔɪ] in CHOICE words. In Liverpool speech, words like “sing” tend to have final [ŋ], the same vowel [ɜː] is used in NURSE and SQUARE words, the final position of the FLEECE and GOOSE vowels tend to be diphthongal. The Liverpool accent is also marked out by its prosodic characteristics which use a rise in some circumstances where others use a fall. These are only some of the Birmingham and Liverpool accent features; more details about the two accents can be found in [9]. The above examples are from [9] and the symbol / / is used for phonemic transcriptions and [ ] is used for allophonic transcriptions (the actual articulation). The words in capitals such as FOOT, STRUT, FACE, GOAT, NURSE, SQUARE, GOOSE, BATCH, CHOICE, CLOTH, and FLEECE are keywords; the keyword represents a set of words which have the same vowel as it does. For examples, the FOOT words refer to “put, bush, full, good, look…” and the FOOT vowel refers to the vowel these words have; the STRUT words refer to “cup, suck, budge, blood…” and the STRUT vowel refers to the vowel these words have.

From previous literature reviews about accents, an accented speech can be defined by the following methods:

a) ‘Rule based’ method; this method involves matching a set of pronunciation rules such as described by Wells [9] and normally requires some linguistic knowledge.
b) ‘Geographical based’ method; in this method, an accented speech is defined according to the geographical origins of the speaker. For example, Birmingham accent can be defined as an accent spoken by someone who was born in Birmingham and has lived there for all of his or her life. Liverpool accent can be defined as an accent spoken by someone who was born in Liverpool and has lived there for all his or her life.

c) ‘Perceptual based’ method; in this method, an agreement has to be achieved between a set of listeners that a speaker has a particular accent. Since accent perception is affected by the listeners’ linguistic backgrounds and their prior knowledge about that particular accent [9,17], therefore, listeners who are familiar with or have knowledge about Birmingham and Liverpool accents have been selected for evaluating the recorded speech in the study.

In this thesis, the speech data were recorded from only one speaker who was born and brought up in Birmingham, but also lived in Liverpool for several years. Based on the nature of this research, the ‘perceptual based’ method will be used for defining accented speech. The accent of the recorded speech and synthesized speech will be identified and evaluated by a number of listeners based on perceptual assessment.
Chapter 3  Speech analysis and synthesis

Speech analysis is a technology of efficiently extracting important acoustic features from speech signals, while speech synthesis is the reverse procedure which aims to reproduce the original speech signal precisely from these features extracted from the analysis phase or produce a modified speech from modified acoustic features. Speech analysis can be carried out in the time domain and the frequency domain. Several speech analysis techniques exist, such as the short time Fourier transform (STFT) analysis-synthesis, Mel-frequency cepstral (MFC) analysis-synthesis and linear prediction (LP) analysis-synthesis, all commonly used techniques in the field of speech signal processing.

In this chapter, the source-filter model is described. It is one of the most common speech production models used in the current speech research, and one of the techniques used in this study. LP speech analysis and synthesis, another method used in this study, is also described in detail in this chapter.

3.1 Source-filter model of speech

The human speech production system is a very complex system; it consists of about 14 human organs. The organs involved in the production of speech are depicted in Figure 3.1. These organs can be classified into three functional components; the source of airflow, the sound source and sound source modification. These three components correspond to the lungs, the larynx and the vocal tract. The vocal tract is from glottis to lips; it includes the pharyngeal cavity, the oral cavity and nasal cavity. The oral cavity is one of the most important parts of the vocal tract. Its size, shape and acoustic characteristics can be varied by movements of the palate, the tongue, the lips, the cheeks and the teeth. The vocal cords may act in several different ways during speech. When people speak, the air
flow from the lungs is forced through the glottis between the vocal cords and the larynx to the vocal tract, and then modified by the shape of the vocal tract to produce a sound. The lips control the size and shape of the mouth opening through which speech sound is radiated.

Figure 3.1 Organs of human speech production system
(http://cnx.org/content/m18086/latest/anatomy.png)
During voiced speech such as the production of vowels and certain nasal consonants like /m/ and /n/, the vocal chords vibrate. Alternatively, in unvoiced speech such as fricatives like /f/, /s/ and unvoiced plosives like /p/, /t/, /k/, the vocal chords do not move and air is forced past the glottis, tongue, teeth and lips.

In the source filter model, the speech signal is represented by a string of sound sources filtered by digital filters with different coefficients. During the production of speech, the glottal excitation signals are modelled by either an impulse train for the voiced speech components such as vowels or white noise for the unvoiced speech such as fricatives; the effects of the shape of the vocal tract are modelled by a digital filter with different coefficients for different sounds. A source filter model of speech is depicted in Figure 3.2. The top part describes a source-filter model in the time domain and the bottom part describes the model in the frequency domain. In the time domain, the sound source is a series of glottal airflow; in the frequency domain, the energy of the speech decreases when the
frequency increases; this is due to the nature of the human speech production system.

### 3.2 Linear prediction (LP) speech analysis and synthesis

Compared to other techniques, such as STFT and MFC, the LP speech analysis technique requires fewer computing operations and can also provide a relatively accurate estimate of speech parameters [31]. Another reason for choosing the LP technique in this study is that it can be used for decomposing a speech signal into a quasi-excitation signal and quasi-stationary filter coefficients; the formants frequencies can also be extracted from the filter coefficients, which are then used as acoustic features for accent analysis and conversion in this research. The basis of linear prediction is that the current speech sample $y(n)$ can be approximated or predicted from a finite number of previous $p$ samples $y(n-1)$ to $y(n-p)$ by a linear combination with small error term $e(n)$ called the residual signal. Thus

$$y(n) = e(n) + \sum_{k=1}^{p} a_k y(n-k)$$  \hspace{1cm} (3.1)

and

$$e(n) = y(n) - \sum_{k=1}^{p} a_k y(n-k) = y(n) - \tilde{y}(n)$$  \hspace{1cm} (3.2)

$$\tilde{y}(n) = \sum_{k=1}^{p} a_k y(n-k)$$  \hspace{1cm} (3.3)

where $\tilde{y}(n)$ is the predicted value, $p$ is the linear predictor order, and $a_k$ represents the linear prediction coefficients which are determined by minimizing the sum of the squared errors between the predicted signal and the actual signal over a frame. There are several methods used for estimating these coefficients such as the autocorrelation method, the covariance method and Burg’s method. The covariance method and the autocorrelation method are the most commonly
used approaches; Burg’s method is considered when high accuracy of models is required, such as signal extrapolation and detection [32].

Taking the z transform of Eq. 3.2, we get

\[
E(z) = Y(z) - \sum_{k=1}^{p} a_k Y(z) z^{-k} = Y(z) \times [1 - \sum_{k=1}^{p} a_k z^{-k}] = Y(z) \times A(z) \quad (3.4)
\]

\[
A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \quad (3.5)
\]

Thus, the error signal \( E(z) \) can be treated as the product of the original speech signal \( Y(z) \) and the transfer function \( A(z) \). \( A(z) \) represents an all-zero digital filter, where the \( a_k \) coefficients correspond to the zeros in the filter’s \( z \)-plane.

Similarly, the original speech signal \( Y(z) \) can be presented by the product of the error signal \( E(z) \) and the transfer function \( \frac{1}{A(z)} \):

\[
Y(z) = E(z) \times \frac{1}{A(z)} \quad (3.6)
\]

The transfer function \( \frac{1}{A(z)} \) represents an all-pole digital filter, where the \( a_k \) coefficients correspond to the poles in the filter’s \( z \)-plane and the roots of \( A(z) \) represent the formant frequencies. The speech signal’s formants’ values (frequency and bandwidth) can be computed through solving polynomial equations.

In the synthesis phase, the excitation signal is either approximated by a train of impulses for voiced parts and by random noise for unvoiced regions, or uses the residual of LP analysis as the excitation signal; then the excitation signal is obtained and filtered with a digital filter for which the coefficients are \( a_k \) from LP analysis. The order of the filter for analysis and synthesis is dependent on the
sampling rate of the speech signal. In practice, for a speech signal at 8 kHz sampling rate, the filter order can be set to between 10 and 12, but for higher quality speech signal such as at 22 kHz sampling rate, the order needed is between 20 and 24 [31].

Figures 3.3-3.5 show the waveform of a segment of voiced speech, its LP residual, short-time fast Fourier transform (STFFT) spectrum and the LP spectrum. It can be seen that STFFT analysis provides a great deal of fine spectral details, while the LP analysis gives an overall envelope of the speech spectrum.

![Figure 3.3 Waveform of a segment of voiced speech](image)

![Figure 3.4 The residual of the speech from 16th LP analysis](image)
Figure 3.5  Short-time FFT spectrum of the speech superimposed with a 16\textsuperscript{th} order LP envelope
Chapter 4

Acoustic analysis of accents

In Chapter 2, the phonetic difference between accents and the acoustic difference between accents have been described briefly. Since phonetic transcription and analysis are related to linguistics, analysis of regional accents from the phonetic aspect will not be discussed in detail in this thesis. In this chapter, analysis of accents from their acoustic aspects will be studied in detail. The chapter also includes several procedures such as speech corpus building, speech segmentation, acoustic parameters extraction and analysis.

4.1 Speech data preparation and recording

The text material includes 25 short sentences which have been selected randomly from CMU_ARCTIC prompts [33] and one paragraph Group C which is from the speech accent archive [34], (see Appendix I). The 25 sentences were then divided into two groups. Group A includes 5 sentences; each sentence was uttered and recorded ten times in each accent. Group B includes 20 sentences and each sentence has been uttered and recorded once in each accent. Each recording has been checked to ensure that the recording quality and voice quality are both good. Text material of the Group A has been used for accents acoustic analysis based on vowel phoneme while Group A, Group B and Group C are used for formants transfer functions training. The statistical occurrences of phonemes in the speech database are listed in Figure 4.1.
Chapter 4  Acoustics analysis of accents

The Praat software [35] was designed and developed by Paul Boersma and David Weenink of the University of Amsterdam for speech analysis. The software is free and can be downloaded from their website. The software provides an accurate and reliable speech analysis system; it can also provide other features such as allowing users to modify several speech acoustic parameters independently. Because of these advantages, the software has become one of the most popular speech analysis tools, which has been used by many researchers for speech analysis and synthesis [36, 37]. For the same reasons, the software was chosen to be used in this study for acoustic parameters extraction, modification and speech synthesis.

In previous accent-related research, the acoustic models of accents were trained using a database from a large group of speakers [19, 23, 38]. For example, Weil et al. [38] investigated acoustic changes of the diaphones /ai/ and /ui/ between Standard American English (SAE) and English as spoken in the southern states of the USA, known as Southern English (SE). Their data was generated by an averaging process, with speech records taken from four participants. The database used by Arslan and Hansen [23] was generated from 48 males speakers across four accents: Neutral, Turkish, Chinese and German. This kind of approach has

![Figure 4.1 Histogram of phoneme content in the speech database](image_url)
also been used in other acoustic modelling; for example, Gerosa et al. [39], used a large database from a wide age range of participants developing an age-independent acoustic model to improve the performance of an automatic speech recognition system. However, due to inter-speaker variability, it is difficult to determine the proportion of acoustic parameter variation that is attributable purely to accent. In order to reduce the effect of inter-speaker variability, in this study, a different and novel approach has been taken; the speech database used in this study was from the same speaker who can speak in different accents.

A female subject was used for this study, who can speak fluently in a Birmingham accent and a Liverpool accent. The recording was conducted on a laptop, using Praat5047_win software, and a head-mounted microphone. The recording was conducted in a quiet room, and the background noise was controlled to the minimum. A sampling rate of 44.1 kHz and 16-bits per sample was used during the recording. For the remainder of the thesis, the Birmingham accent will be referred to as Accent BM, and the Liverpool accent will be referred to as Accent LP.

The recordings were perceptually tested by three native-English speakers, who are familiar with the Birmingham and Liverpool accents and can readily identify and differentiate between the two accents.

### 4.2 Vowel-based acoustic analysis of accents

In phonetics, a vowel sound in spoken language is characterized by an open configuration of the vocal tract so that there is no build up of air pressure above the glottis. A consonant sound is characterised by a constriction or closure at one or more points along the vocal tract. Since accents are most distinctive in the realization of the vowels and vowel based speech analysis technique has been widely used in previous research. In this study, nine monophthongal vowels and three diphthongal vowels are nominated for analysis. A further reason for
choosing to concentrate on the vowels in accent analysis is because the formants’ tracks in vowel regions are stable compared to those in unvoiced and weak fricatives, which are unreliable.

4.2.1 Vowels extraction

In order to extract each vowel waveform from the whole speech signal, the speech signal must be segmented at phoneme level and aligned with phonetic transcriptions. Speech segmentation and alignment can be done manually, automatically or by combining the two methods, performing automatic segmentation first, then manual correction. High accuracy segmentation can be achieved manually; however, manual segmentation and labelling is extremely time-consuming for large data sets. Automatic segmentation and labelling has become standard practice over the last decade, with the hidden Markov model (HMM)-based approaches being most widely used for automatic segmentation in speech synthesis and speech recognition [40-42]. In this method, the phonetic transcription of the sentences must be obtained at the first stage of the segmentation process and a large speech data corpus is required for training each phone model in order to achieved a reasonably accuracy. There are several software packages such as Speech Filling System (SFS) [43], Hidden Markov modelling toolkit (HTK) [44], and Praat [35], that can be used as an assistant tool to perform manual speech segmentation and labelling. For this study, since the speech data is relatively small, in order to achieve high accuracy segmentation, speech segmentation has been undertaken manually and the Praat software was used. During segmentation, the utterances were first segmented into individual words using waveforms in conjunction with spectrograms, and then the words of interest were segmented into phoneme levels. Onset and offset time of phone were taken at the points where there was evidence of phone release or closure. Figure 4.2 shows an example of segmentation. The top panel shows the waveform of the speech, the middle panel shows the spectrogram of the speech, and the bottom panel shows the segmentation at the phoneme level.
Chapter 4  Acoustics analysis of accents

Figure 4.2  Example of manual segmentation of the phrase “more like” with Praat software.

Twelve vowels which include nine monophthongal vowels: (AA as in barn, AE as in pat, AO as in born, AX as in about, ER as in burn, IH as in pit, OH as in pot, UH as in good and UW as in boon) and three diphthongal vowels (AY as in buy, EY as in bay and OW as in loan) have been manually extracted from eleven words and five utterances. The twelve vowels and the words from which the vowels were extracted are shown in Appendix II. It must be noted that the transcriptions in Appendix II are BEEP standard transcriptions, they do not reflect any accent differences, and can vary from accent to accent. For example, the word of “laughter” has a transcription “L AA F T AX” in BEEP, a Birmingham or Liverpool speaker would probably pronounce it as “L AE F T AX” [45]. In this thesis, the vowel AA refers to the first vowel in “laughter”; AE refers to the first vowel in “and”; AO refers to the first vowel in “more”; AX refers to the second vowel in “sugar”; ER refers to the first vowel in “burst”; IH refers to the first vowel in “his”; OH refers to the first vowel in “not”; UH refers to the first vowel in “sugar”; UW refers to the first vowel in “you”; AY refers to the first vowel in “like”; EY refers to the first vowel in “place” and OW refers to the first vowel in “no”.
4.2.2 Acoustic features extraction and analysis

In this study, acoustic parameters such as the first three formant frequencies F1, F2 and F3, pitch slope (derivative of the fundamental frequency), the intensity of speech and the duration of the vowels were used for determining the acoustic correlations of the two accents and identifying the most accent-related acoustic features. Software package Praat was used for the acoustic parameters extraction.

4.2.2.1 Formant frequencies extraction and analysis

Individuals configure their vocal tracts into different shapes to articulate different types of sound in different accents. Accent is a distinctive characteristic manner of pronunciation, mostly related to the shape of the vocal tract, which can be modified by the movements of the palate, the tongue, the lips, the cheeks and the teeth.

Formants correspond to resonant frequencies with the greatest amplitude, and they are dependent on the configuration of the vocal tract for articulating different types of voiced sounds, most notably vowel sound [46]. As described in [47], the specific formants F1, F2, and F3 are typically evaluated for comparison of different vowels: F1, the lowest formant, is associated with vowel height; the second formant F2 correlates roughly with tongue advancement; and both F2 and F3 vary with the degree of lip rounding and depend on the position of constriction. For a constriction in the very front part of the vocal tract, the major effect of lip rounding is to lower F3; a constriction in the palatal, velar, and uvular regions, on the other hand, has much greater effect on F2. Vowel height refers to the vertical position of the tongue relative to the roof of the mouth, whilst tongue advancement refers to the horizontal position of highest point of the tongue relative to the back of the mouth during the articulation of a vowel. Thus an open vowel has a higher first formant frequency F1 and a closed vowel has a lower first
formant frequency F1, a front vowel has a higher second formant frequency F2 and a back vowel has a lower second formant frequency F2.

Burg’s algorithm [48] was applied in the formants tracking. Burg’s algorithm estimates the reflection coefficients to minimize the sum of the forward and backward residual error in the linear prediction. The advantage of Burg’s approach is that the algorithm finds the reflection coefficients directly from the data without having to calculate the autocorrelation coefficients. Furthermore, this approach does not need to window the data [49]. A detailed description of Burg’s algorithms can be found in [32]. During the formants extraction, a value of 5500 Hz was set for maximum formant frequency, which is typical maximum value for an adult female; pre-emphasis was also applied for frequencies above 50 Hz before analysis; a 25 ms Gaussian window and 6.25 ms frame size were used in the formants tracking. Praat takes a few processing steps to perform formants tracking. Firstly, the speech signal was resampled to a sampling frequency of twice the value of the maximum formant; in this case, the resampled frequency was 11000 Hz, and high-pass filters at 50 Hz were applied on the resampled signal. Then the signal was pre-emphasised before linear prediction analysis and the linear predictive coding coefficients (LPCCs) were computed with Burg’s algorithm. Finally, formant frequencies were obtained by root solving the linear prediction polynomial.

Five formants were extracted; however, only the first three formants have been analysed in this study for accent analysis; this is because the vowels are adequately specified by the first three formants [50, 51]. The mean and standard deviations of the first three formants F1, F2 and F3 of twelve vowels over 10 repeated utterances are shown in Tables 4.1-4.3 and Figure 4.3. The t-tests were performed using statistical analysis software SPSS 16.0 with α set to 0.05.
Table 4.1  Mean and standard deviation of the first formant F1 and its \( t \)-test of twelve vowels over 10 repeated utterances

<table>
<thead>
<tr>
<th>Vowels</th>
<th>F1-mean (Hz)</th>
<th>F1-SD (Hz)</th>
<th>Paired samples ( t )-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BM</td>
<td>LP</td>
<td>BM</td>
</tr>
<tr>
<td>AA</td>
<td>753</td>
<td>955</td>
<td>21</td>
</tr>
<tr>
<td>AE</td>
<td>787</td>
<td>864</td>
<td>65</td>
</tr>
<tr>
<td>AO</td>
<td>575</td>
<td>606</td>
<td>29</td>
</tr>
<tr>
<td>AX</td>
<td>877</td>
<td>801</td>
<td>51</td>
</tr>
<tr>
<td>ER</td>
<td>591</td>
<td>496</td>
<td>26</td>
</tr>
<tr>
<td>IH</td>
<td>383</td>
<td>428</td>
<td>12</td>
</tr>
<tr>
<td>OH</td>
<td>710</td>
<td>684</td>
<td>18</td>
</tr>
<tr>
<td>UH</td>
<td>784</td>
<td>645</td>
<td>122</td>
</tr>
<tr>
<td>UW</td>
<td>484</td>
<td>443</td>
<td>26</td>
</tr>
<tr>
<td>AY</td>
<td>729</td>
<td>787</td>
<td>25</td>
</tr>
<tr>
<td>EY</td>
<td>600</td>
<td>485</td>
<td>29</td>
</tr>
<tr>
<td>OW</td>
<td>695</td>
<td>671</td>
<td>16</td>
</tr>
</tbody>
</table>
Table 4.2  Mean and standard deviation of the second formant F2 and its t-test of twelve vowels over 10 repeated utterances

<table>
<thead>
<tr>
<th>Vowels</th>
<th>F2-mean (Hz)</th>
<th>F2-SD (Hz)</th>
<th>Paired samples t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BM</td>
<td>LP</td>
<td>LP/BM</td>
</tr>
<tr>
<td>AA</td>
<td>1354</td>
<td>1717</td>
<td>1.27</td>
</tr>
<tr>
<td>AE</td>
<td>1815</td>
<td>1801</td>
<td>0.99</td>
</tr>
<tr>
<td>AO</td>
<td>1028</td>
<td>1002</td>
<td>0.97</td>
</tr>
<tr>
<td>AX</td>
<td>1838</td>
<td>2005</td>
<td>1.09</td>
</tr>
<tr>
<td>ER</td>
<td>1809</td>
<td>1776</td>
<td>0.98</td>
</tr>
<tr>
<td>IH</td>
<td>2271</td>
<td>2126</td>
<td>0.94</td>
</tr>
<tr>
<td>OH</td>
<td>1519</td>
<td>1424</td>
<td>0.94</td>
</tr>
<tr>
<td>UH</td>
<td>1939</td>
<td>1746</td>
<td>0.90</td>
</tr>
<tr>
<td>UW</td>
<td>2104</td>
<td>2166</td>
<td>1.03</td>
</tr>
<tr>
<td>AY</td>
<td>1485</td>
<td>1835</td>
<td>1.24</td>
</tr>
<tr>
<td>EY</td>
<td>2192</td>
<td>2417</td>
<td>1.10</td>
</tr>
<tr>
<td>OW</td>
<td>1628</td>
<td>1436</td>
<td>0.88</td>
</tr>
</tbody>
</table>
Chapter 4  Acoustics analysis of accents

Table 4.3  Mean and standard deviation of the third formant F3 and its $t$-test of twelve vowels over 10 repeated utterances

<table>
<thead>
<tr>
<th>Vowels</th>
<th>F3-mean</th>
<th>F3-SD</th>
<th>Paired samples $t$-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BM</td>
<td>LP</td>
<td>LP/BM</td>
</tr>
<tr>
<td>AA</td>
<td>2948</td>
<td>2887</td>
<td>0.98</td>
</tr>
<tr>
<td>AE</td>
<td>2916</td>
<td>2842</td>
<td>0.97</td>
</tr>
<tr>
<td>AO</td>
<td>2980</td>
<td>3210</td>
<td>1.08</td>
</tr>
<tr>
<td>AX</td>
<td>2866</td>
<td>3089</td>
<td>1.08</td>
</tr>
<tr>
<td>ER</td>
<td>2956</td>
<td>2928</td>
<td>0.99</td>
</tr>
<tr>
<td>IH</td>
<td>3014</td>
<td>3083</td>
<td>1.02</td>
</tr>
<tr>
<td>OH</td>
<td>3091</td>
<td>2971</td>
<td>0.96</td>
</tr>
<tr>
<td>UH</td>
<td>3131</td>
<td>3046</td>
<td>0.97</td>
</tr>
<tr>
<td>UW</td>
<td>2869</td>
<td>2951</td>
<td>1.03</td>
</tr>
<tr>
<td>AY</td>
<td>2839</td>
<td>2994</td>
<td>1.05</td>
</tr>
<tr>
<td>EY</td>
<td>2880</td>
<td>3002</td>
<td>1.04</td>
</tr>
<tr>
<td>OW</td>
<td>2869</td>
<td>3186</td>
<td>1.11</td>
</tr>
</tbody>
</table>

Figure 4.3  Mean values of first three formants of twelve vowels over 10 repeated utterances
From Tables 4.1-4.3 and Figure 4.3, it can be seen that vowels AA, AE, AO, IH, AY, the first vowel in words “laughter, and, more, his, like” have higher values of the first formant F1 for Accent LP, and there are lower F1 values in Accent LP for
AX, the second vowel in “sugar” and vowels ER, OH, UW, EY, OW, the first vowel in words “burst, not, sugar, you, place, no”. For the second formants F2, vowels AA, UW, AY, EY, the first vowel in words “laughter, you, like, place” and AX, the second vowel in “sugar” have higher values in Accent LP. For the other seven vowels F2 are higher in Accent BM. For the third formant F3, vowels AO, IH, UW, AY, EY, OW, the first vowel in word “more, his, you, like, place, no” and AX, the second vowel in “sugar” are higher in Accent LP than that in Accent BM. For vowels AA, IH, UW, AY, EY, the first vowel in words “laughter, his, sugar, like, place”, there are significant differences in the first formants F1 and second formants F2 between the two accents, as indicated by the p-values. Since the first two formants are most important in determining vowel quality [51], this may imply that these five vowels are more sensitive in discriminating between the two accents; the rest of the vowels have less distinct differences between the two accents.

A comparison of the formants for Accent BM and Accent LP provides an insight as to how people manipulate the shape of their vocal tract during articulation when they shift their accent from one to another. From Figure 4.4, it can be seen that the middle vowel AO, the first vowel in “more” and high vowel IH, the first vowel in “his” have been lowered (F1 increased) during accent shifting from BM to LP, while the middle vowels: AX, the second vowel in “sugar”, ER, the first vowel in “burst” and low vowel OH, the first vowel in “not” have been raised (F1 decreased). It also can be seen that the back vowels AA and UW, the first vowel in words “laughter, you” have been moved forward (F2 increased), while the front vowels AE, IH, the first vowel in words “and, his” and central vowel ER in “burst” have been moved further back (F2 decreased) during accent shifting from accent BM to LP. Figure 4.5 indicates that the degree of roundness of lips is reduced in Accent LP for vowels AO, IH, UW, AY, EY, OW, the first vowel in words “more, his, you, like, place, no” and AX, the second vowel in “sugar” (F3 increased).
4.2.2.2 Fundamental frequency extraction and pitch slope analysis

There are several methods for fundamental frequency estimation, such as zero-crossing rate method, peak rate counting method and slope event rate counting method in the time domain, and cepstrum analysis and multi-resolution methods in the frequency domain [52, 53]. In this study, an enhanced auto-correlation algorithm developed by Boersma [54] was used for pitch contour tracking due to its straightforwardness, flexibility and robustness. This algorithm performs a short-term autocorrelation analysis on a number of small segments taken from the signal, then identifies auto-correlation peaks for each analysis frame, and chooses a path through these peaks that avoids pitch jumps and favours high correlation peaks. This algorithm works equally well for low pitches such as a creaky voice as low as 16 Hz, middle pitches such as a female speaker at 200 Hz and high pitches such as a child speaker at 250 Hz. Its accuracy and robustness has been proved by many researchers [7, 55]. During the fundamental frequency F0 extraction in this study, the pitch range was set to 75–500Hz; a silence threshold (0.03), and voicing threshold (0.45) were also set to differentiate silence, voiced, and unvoiced frames.

The changes (rise and fall) of the fundamental frequency during speech correlate to the intonation of speech. Each language and accent has its own unique set of patterns for intonation, stress and rhythm. The role of intonation in foreign accents has been studied in past years [19, 56, 57]. Previous research has shown that intonation plays a significant role in accent establishment and, consequently, identification. The research from Yan et al. [19] showed that fundamental frequency range was closely related to intonation patterns employed by different accents. For example, British speakers have the largest fundamental frequency range and the largest initial pitch rise and final pitch fall rates compared to Austrian and American speakers. Experiments conducted by Grover et al. [57] verified that French, English and German speakers differ in the slopes of their intonations over continuous parts of the pitch contour. In their study, the slope
was computed as the maximum change in fundamental frequency divided by the
time over which this change occurs.

In this study, after the fundamental frequencies of the twelve vowels were
extracted by suing Praat software package, the pitch slope was computed as the
variation of fundamental frequency in Hertz (the value at the end minus the value
at the start) divided by duration of the vowel in seconds. It reflects the steepness
of the rise or fall in pitch over the whole vowel. The results of the pitch slope
analysis of the twelve vowels which is mean value over 10 repeated
measurements shown in Table 4.4 and Figure 4.6.

<table>
<thead>
<tr>
<th>Vowels</th>
<th>Pitch slope (Hz/sec)</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BM</td>
<td>LP</td>
<td>LP/BM</td>
</tr>
<tr>
<td>BM</td>
<td>BM</td>
<td>LP</td>
<td>BM</td>
</tr>
<tr>
<td>AA</td>
<td>-172</td>
<td>-340</td>
<td>1.98</td>
</tr>
<tr>
<td>AE</td>
<td>-160</td>
<td>-137</td>
<td>0.85</td>
</tr>
<tr>
<td>AO</td>
<td>-36</td>
<td>-1</td>
<td>0.03</td>
</tr>
<tr>
<td>AX</td>
<td>120</td>
<td>77</td>
<td>0.64</td>
</tr>
<tr>
<td>ER</td>
<td>-38</td>
<td>-18</td>
<td>0.46</td>
</tr>
<tr>
<td>IH</td>
<td>-231</td>
<td>-254</td>
<td>1.10</td>
</tr>
<tr>
<td>OH</td>
<td>-310</td>
<td>-412</td>
<td>1.33</td>
</tr>
<tr>
<td>UH</td>
<td>-463</td>
<td>-361</td>
<td>0.78</td>
</tr>
<tr>
<td>UW</td>
<td>-99</td>
<td>-97</td>
<td>0.99</td>
</tr>
<tr>
<td>AY</td>
<td>-9</td>
<td>-104</td>
<td>1.46</td>
</tr>
<tr>
<td>EY</td>
<td>-237</td>
<td>30</td>
<td>-0.12</td>
</tr>
<tr>
<td>OW</td>
<td>-63</td>
<td>-46</td>
<td>0.73</td>
</tr>
</tbody>
</table>
Figure 4.6  The pitch slopes of twelve vowels over 10 repeated utterances

From Table 4.4 and Figure 4.6, it can be seen that the pitch slope is positive in Accent LP and negative in Accent BM for the diphthongal vowel EY, the first vowel in “place”, for other vowels, the pitch slope have a similar raise or fall trend in the two accents. However, the steepness of rising or falling differs in the two accents, vowels AA, IH, OH, the first vowel in words “laughter, his, not” have steeper falls in Accent LP, while Accent BM has steeper falls for vowels AE, AO, ER, OW, the first vowel in words “and, more, burst, no” and a steeper rise for AX, the second vowel in “sugar”.

4.2.2.3 Intensity analysis

Intensity refers to the sound power per unit area and is perceived as the loudness of the sound. In order to determine correlations between intensity and accent, the intensity of the twelve vowels was also analysed and the result is shown in Table 4.5 and Figure 4.7. During the intensity calculation, the amplitude of the sound wave was first squared, and then convolved with a Gaussian analysis window. The length of the analysis window depends on the minimum pitch value settings and is calculated as dividing 3.2 by minimum pitch value. In this study,
75 Hz was used as minimum pitch value and this gives the analysis window length of about 43 ms.

There are a number of factors contributing to the intensity characteristics of speech, for example, environmental factors such as distance from the speaker, recording device settings, the number of people in the audience and the size of room; speaker individualities such as habitual vocal loudness; prosodic effect such as stress and type of utterance, and the complex interaction of these various factors [58]. However, the analysis database used in this research was from the same speaker and the recordings were also made in the same environment; this ensures that the variance of intensity was mainly due to accent.

### Table 4.5  Mean and standard deviation of intensity and it t-test of twelve vowels over 10 repeated utterances

<table>
<thead>
<tr>
<th>Vowels</th>
<th>Intensity-mean (dB)</th>
<th>Intensity-SD (dB)</th>
<th>Paired samples t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BM</td>
<td>LP</td>
<td>LP/BM</td>
</tr>
<tr>
<td>AA</td>
<td>64.10</td>
<td>65.40</td>
<td>1.02</td>
</tr>
<tr>
<td>AE</td>
<td>58.40</td>
<td>64.00</td>
<td>1.10</td>
</tr>
<tr>
<td>AO</td>
<td>62.80</td>
<td>63.40</td>
<td>1.01</td>
</tr>
<tr>
<td>AX</td>
<td>58.20</td>
<td>60.70</td>
<td>1.04</td>
</tr>
<tr>
<td>ER</td>
<td>66.50</td>
<td>68.30</td>
<td>1.03</td>
</tr>
<tr>
<td>IH</td>
<td>60.10</td>
<td>62.10</td>
<td>1.03</td>
</tr>
<tr>
<td>OH</td>
<td>60.20</td>
<td>62.60</td>
<td>1.04</td>
</tr>
<tr>
<td>UH</td>
<td>58.00</td>
<td>60.70</td>
<td>1.05</td>
</tr>
<tr>
<td>UW</td>
<td>57.20</td>
<td>63.60</td>
<td>1.11</td>
</tr>
<tr>
<td>AY</td>
<td>61.80</td>
<td>65.50</td>
<td>1.06</td>
</tr>
<tr>
<td>EY</td>
<td>61.90</td>
<td>63.60</td>
<td>1.03</td>
</tr>
<tr>
<td>OW</td>
<td>64.30</td>
<td>64.40</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Chapter 4  Acoustics analysis of accents

Figure 4.7  The mean value of intensity of twelve vowels

Form Table 4.5 and Figure 4.7, it can be seen that the average intensity also changes when individuals shift to an alternative accent. Also evident is the fact that the average intensity in Accent LP is higher than that in Accent BM for all the twelve vowels. However, the variation in intensities between the two accents is not as large as the difference in the formant frequencies and the fundamental frequencies variations for the two accents. This may indicate that variations of speech intensity can mostly be attributed to the characters of individual speakers and not their accents.

4.2.2.4 Duration analysis

Phonetic duration of a vowel varies considerably due to factors such as speaking style, stress, the locations of pauses, the boundaries of word and syllable, the manner of articulation and rhythm. In general, vowels before a voiced consonant are phonetically longer than those before the voiceless consonants for typical English and many other languages [59], for example AE in “pat” is phonetically shorter than it is in “pad”. Phonetic duration of vowels also differs with accents. Scobie et al. [59] studied patterns of Scottish English vowel duration based on the data which were collected from Glasgow and Edinburgh and in large part from
middle-class speakers, they found that Scottish English had smaller durations for the close vowels /i/ and /u/ before /d/ and /s/ compared with American English.

The results of mean and standard deviations of phonetic duration of the twelve vowels computed over ten repeated utterances analysed in this research are shown in Table 4.6 and Figure 4.8.

**Table 4.6  Mean and standard deviation of duration and its t-test of twelve vowels over 10 repeated utterances**

<table>
<thead>
<tr>
<th>Vowels</th>
<th>Duration-mean (ms)</th>
<th>Duration-SD (ms)</th>
<th>Paired samples t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BM</td>
<td>LP</td>
<td>LP-BM</td>
</tr>
<tr>
<td>AA</td>
<td>143</td>
<td>105</td>
<td>38</td>
</tr>
<tr>
<td>AE</td>
<td>79</td>
<td>73</td>
<td>6</td>
</tr>
<tr>
<td>AO</td>
<td>134</td>
<td>136</td>
<td>-2</td>
</tr>
<tr>
<td>AX</td>
<td>132</td>
<td>110</td>
<td>21</td>
</tr>
<tr>
<td>ER</td>
<td>134</td>
<td>143</td>
<td>-9</td>
</tr>
<tr>
<td>IH</td>
<td>66</td>
<td>74</td>
<td>-7</td>
</tr>
<tr>
<td>OH</td>
<td>83</td>
<td>86</td>
<td>-3</td>
</tr>
<tr>
<td>UH</td>
<td>105</td>
<td>88</td>
<td>17</td>
</tr>
<tr>
<td>UW</td>
<td>171</td>
<td>175</td>
<td>-4</td>
</tr>
<tr>
<td>AY</td>
<td>103</td>
<td>111</td>
<td>-8</td>
</tr>
<tr>
<td>EY</td>
<td>95</td>
<td>82</td>
<td>13</td>
</tr>
<tr>
<td>OW</td>
<td>117</td>
<td>95</td>
<td>22</td>
</tr>
<tr>
<td>Average</td>
<td>113</td>
<td>107</td>
<td>/</td>
</tr>
</tbody>
</table>
From duration analysis of the twelve vowels as shown in Table 4.6 and Figure 4.8, it can be seen that vowels AA, AE, UH, EY, OW, the first vowel in words “laughter, and, sugar, place, no” and AX, the second vowel in “sugar” have been shortened by a range of between 6 ms and 38 ms for Accent LP. Conversely, vowels AO, ER, IH, OH, UW and AY, the first vowel in words “more, burst, his, not, you, like” have been lengthened by a range of between 2 ms and 9 ms in Accent LP. Overall, Accent LP has a shorter average vowel duration; the average duration over all the twelve vowels are 113 ms and 107 ms for Accent BM and Accent LP respectively.

**4.3 Conclusions**

The results of vowel based acoustic analysis of the two accents showed that the formant frequencies, the pitch slope, intensity, and the phonetic duration all varied to differing degrees when individuals transfer to alternative accents. Among these acoustic features, the formant frequencies, the pitch slope and phone duration can provide better discrimination than intensity as indicated by the values of LP/BM. This also indicated that the formant frequencies and the fundamental frequency
variation contour could be the most prominent acoustic features influenced by the variability of the accents followed by phone duration.
Chapter 5

Accent conversion

Accent conversion intends to modify the part of a speaker’s acoustic characteristics that influence accent without changing the main acoustic characteristics of the speaker’s voice which allow him/her to be identified. Accent conversion is different from voice conversion. Voice conversion aims at transforming the characteristics of the speech signal in such a way that a human naturally perceives the target speaker’s own characteristics in the transformed speech [60]; this transformation exchanges the speaker’s identities both of accent and voice. Voice conversion research has been actively conducted in the past two decades; the issues concerning voice conversion have been addressed in several papers [61-63]. In a typical voice conversion system, voice conversion is achieved by mapping the spectrum between source speaker and target speaker via different spectral mapping techniques and other acoustic parameter manipulations such as fundamental frequency adjustment and duration modification to improve the similarity of the converted speech. Shuang et al. [6] proposed a voice conversion method using a frequency warping function which is generated from mapping the first four formants of the source speaker and the target speaker. In addition to frequency warping, they also used fundamental frequency adjustment, breathiness addition, and duration modification to improve the quality and similarity of the converted speech. Compared to voice conversion, accent conversion is a relatively new subject with a few papers available [28, 29]. An accent conversion procedure would involve manipulating several acoustic parameters of a recording of a speaker to change the speaker’s accent to a desired target accent as far as listeners were concerned but without changing the speaker’s voice identity.

One approach to accent conversion is phonetic elements substitution. When the phonetic content of a source utterance is known, accent conversion can be achieved by substituting those phonetic elements which are significantly different in the target accent [64]. Although the substitution method can reduce the impact
of the spectral mapping on speaker identity, this approach still needs a big speech corpus for each accent.

Another approach to accent conversion is text-independent conversion. This approach allows the user to convert any utterance to a desired target accent without knowing the phonetic contents of the source utterance. Based on the source-filter model, the speech signal can be decomposed into two components: excitation (source) part and vocal tract (filter) part via linear prediction analysis. The vocal tract filter which characterizes the spectral envelope contributes most of the linguistic information such as context message, accent identity and emotion states of the speaker; the glottal excitation signal contains most information about the speaker’s voice character [65]. Accent conversion can be achieved through spectral mapping and other acoustic parameter modification such as fundamental frequency and its contour modification. Several spectral features such as MFCCs, LPCCs and LSP can be used for spectral mapping. Since the spectral envelope of a segment of speech can be characterized by its formant frequencies, which have dominant energy over other frequencies, therefore, the formant frequencies have been used as spectral features for spectral mapping in voice conversion research by several researchers [66, 67]. The formant frequencies are also used as spectral features of speech for spectral mapping in this study.

Several spectral mapping techniques which have been used in voice conversion can be adapted to accent conversion. Codebook mapping and Gaussian mixture model (GMM), the two most basic and popular techniques for speech spectral mapping, have been used in voice conversion. The codebook mapping method [8, 68] involves replacing the source speaker’s spectral vectors with corresponding ones of the target speaker frame by frame to obtain the converted spectrum; the mapping codebook can be generated using different vector quantization (VQ) algorithms. However, this method may introduce discontinuities into the speech spectrum. In order to alleviate the discontinuities in the converted spectrum from codebook mapping, a Gaussian mixture model for spectrum conversion [69, 70]
was introduced. Unfortunately, in GMM-based voice conversion, the converted spectrum may be excessively smoothed by the statistical averaging operation; this will result in the quality of the converted speech being degraded. Combining GMM-based conversion with other algorithms such as dynamic frequency warping (DFW) [71], or maximum a posteriori (MAP) adaptation [72] avoids the over-smoothed spectrum from the simple GMM-based voice conversion. Other types of method, such as artificial neural networks (ANNs) [67, 73, 74], and vocal tract length normalization (VTLN) [75] have also been proposed.

The accent conversion framework used in this study is illustrated in Figure 5.1. The accent conversion procedure can be summarized in the following typical steps:

Step 1 Input source utterances into software Praat, apply an LP analysis with 30 orders on the source utterance to obtain the coefficients of LP filter, then apply a standard LP inverse filtering on the source utterances to separate glottal excitation signals from vocal tract spectra.

Step 2 Extract formants of the source utterance to produce a formant object. (The same settings and methods were used here as that in Chapter 4 for vowels formants tracking).

Step 3 Perform formants modification. Modify the first three formants of the source utterance using different formulas which were derived from different mapping algorithms.

Step 4 Resynthesize utterances. Apply a standard LP filter which was constructed from the modified formants onto the glottal excitation signal which was extracted from the source utterance to produce the converted utterances. At this stage, only the spectra have been modified in the converted, the pitch contour remained intact.
Step 5 Perform pitch contour modification. Extract pitch contour from the synthesized utterance (from step 4) and the target utterance, then transplant pitch contour of the target utterance to the synthesized utterance. Detailed procedures can be found in section 5.4.2.

![Flow diagram of accent conversion framework](image)

**Figure 5.1 Flow diagram of accent conversion framework**

### 5.1 Training materials

In order to estimate the transformation functions of formant frequencies between the two accents, a reasonable number of parallel utterances from each accent are needed for training. In this study, 20 short sentences and one paragraph were used for formants transfer function training (see Appendix I), the 20 short sentence
were made up of 5 sentences from Group A and the first 15 sentences from Group B. Each sentence in Group A was uttered and recorded 10 times in each accent, the sentences from group B were each uttered once, and the paragraph (Group C) was split into 7 sentences; therefore there were 72 utterances in total for each accent. The transfer functions are estimated on “word” level. The reason for choosing “word” as the basic segmental unit here was that a highly accurate segmentation at phoneme level is not only time-consuming but may also need a linguistic expert. The “word” level segmentation can save time and also achieve a high accuracy.

Firstly, each utterance was manually segmented at “word” level. The “word” here does not really mean an isolated word; sometimes it may really be two words. This depended on the waveform of the speech. When there was no obvious indicator for separating two words from the speech waveform, they were treated as one “word”. For example, the words “I ever” were treated as one word “iever”; this ensured the accuracy of alignment. In this way, 389 “words” were obtained from 72 utterances in total for each accent via manual segmentation.

In order to further reduce the misalignment during the segmentation stage, a confidence measure was conducted in a similar manner as in [42] before training. For each “word” in each utterance, the mean values of the first three formants were calculated for both accents. The mean values were acquired by averaging the formants values of all frames that fall within that “word”.

Firstly, the absolute difference of the first formant between the BMF1 (first formant of Accent BM) and LPF1 (first formant of Accent LP) $\Delta f_1$ was calculated for each “word”, as in Eq. 5.1, the absolute difference of the second formant $\Delta f_2$ and the absolute difference of the third formants $\Delta f_3$ of each “word” were also calculated in a similar way, as in Eq. 5.2-5.3. Then, $\Delta f$, the mean of $\Delta f_1, \Delta f_2$ and $\Delta f_3$ was computed for each “word”, as in Eq. 5.4. Finally,
\( \mu_{\Delta f} \), the mean of \( \Delta f \) and their standard deviation \( \delta_{\Delta f} \) were computed. Any “words” which satisfied one of the inequalities in Eq. 5.5 and Eq. 5.6 were eliminated from the training database. In Eq. 5.5 and Eq. 5.6, \( a \) was the threshold factor; a value of \( a = 1.5 \) was used in this research.

\[
\begin{align*}
\Delta f_1 &= |BMF1 - LPF1| \\
\Delta f_2 &= |BMF2 - LPF2| \\
\Delta f_3 &= |BMF3 - LPF3| \\
\Delta f &= (\Delta f_1 + \Delta f_2 + \Delta f_3) / 3
\end{align*}
\]

After application of the confidence measure, the total “words” were reduced from 389 to 361. Therefore, 361 words were used in the training phase. Figure 5.2 shows the histogram of \( \Delta f \) values of the two accents and Figure 5.3 shows its probability distribution. The “words” falling in the grey part were excluded from the training database.
Chapter 5  Accent conversion

Figure 5.2  Histogram of $\Delta f$ values for the two accents

Figure 5.3  Probability distribution of $\Delta f$
5.2 Spectral mapping algorithms

The accent conversion algorithm presented here is based on the source-filter model of speech production [15]. As described before, in this model, voiced speech is the output of a time-varying vocal-tract filter, excited by a time-varying glottal pulse signal. In order to separate the vocal-tract filter from the source signal, a linear prediction speech analysis technique [76] is applied. In this study, linear predictive coding coefficients were used for representing the vocal tract characteristics, and formant frequencies which are derived from the LPCCs were used for developing spectral mapping functions. As with the analysis in Chapter 4, it can be seen that formant frequencies are one of the prominent acoustic features for discrimination of the two British regional accents. It is possible to adjust the main formant frequencies of the source speaker to the formant frequencies of the target speaker to achieve accent conversion. From the literature review, formant-based spectral mapping were used by many researchers in voice conversion and voice morphing research [66, 77, 78]. Mizuno and Abe [66] implemented voice conversion through directly modifying formant frequencies, formant bandwidths and spectral intensity using a mapping codebook.

In this study, the following three different mapping algorithms which have been used in voice conversion were investigated and compared, and then these algorithms were applied to accent conversion.

5.2.1 Mean-variance linear conversion function

Mean-variance linear conversion has been used as a baseline method in many research activities because of its simplicity and good performance [7, 8, 79]. The method was based on the assumption that the acoustic features (here the first three formant frequencies) from the source and the target speakers have a Gaussian distribution, and can be expressed as:
where $X=[x_1, \ldots, x_N]$ represents a sequence of acoustic features extracted from the source utterances and $Y=[y_1, \ldots, y_N]$ represents a sequence of acoustic features extracted from the target utterances; $\mu_X$ and $\mu_Y$ are mean value of source utterance and target utterance respectively, while $\sigma_X$ and $\sigma_Y$ are the respective standard deviations.

A linear transformation can be defined as follows:

$$y = f(x) = ax + b$$  \hspace{1cm} (5.8)

where $x$ and $y$ are source value and predicted target value respectively, $a$ is a scaling factor and $b$ is bias value.

From Eq. 5.8 we can get:

$$x = f^{-1}(y) = \frac{y - b}{a}$$  \hspace{1cm} (5.9)

The linear transformation of the probability density function (pdf) can be written as follows:

$$p(y) = \frac{1}{a} p(x)$$  \hspace{1cm} (5.10)

Where $p(x)$ and $p(y)$ are the probability density functions of $x$ and $y$ respectively, and can be expressed as the follows:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma_x}} \exp\left(-\frac{(x - \mu_x)^2}{2\sigma_x^2}\right)$$ \hspace{1cm} (5.11)
Chapter 5  Accent conversion

\[ p(y) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left(-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right) \]  (5.12)

Apply Eq. 5.9, Eq. 5.11 and Eq. 5.12 to Eq. 5.10, the following equation can be derived

\[ \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left(-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right) = \frac{1}{a} \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{(\frac{y-b}{a}-\mu_x)^2}{2\sigma_x^2}\right) \]  (5.13)

Taking the logarithm of both sides of Eq. 5.13 gives:

\[ \log\left(\frac{a\sigma_x}{\sigma_y}\right) - \frac{(y-\mu_y)^2}{2\sigma_y^2} = -\frac{(\frac{y-b}{a}-\mu_x)^2}{2\sigma_x^2} \]  (5.14)

Then the following simultaneous equations are solved

\[ \log\left(\frac{a\sigma_x}{\sigma_y}\right) = 0 \]  (5.15)

\[ \frac{(y-\mu_y)^2}{2\sigma_y^2} = \frac{(\frac{y-b}{a}-\mu_x)^2}{2\sigma_x^2} \]  (5.16)

The following equations are thus derived

\[ a = \frac{\sigma_y}{\sigma_x} \]  (5.17)

\[ b = \mu_y - \frac{\sigma_y\mu_x}{\sigma_x} \]  (5.18)

In the mean-variance linear conversion method, the formant frequency values BMF1 (first formant of Accent BM), BMF2 (second formant of Accent BM),
BMF3 (third formant of Accent BM) of 361 “words” were used to compute the mean value and standard deviation of Accent BM. Similarly, the mean value and standard deviation of Accent LP were computed from the formant frequency values LPF1 (first formant of Accent LP), LPF2 (second formant of Accent LP) and LPF3 (third formant of Accent LP). After mean value and standard deviation of the three formants were computed for both accents, the scaling factor $a$ and bias value $b$ for the three formants were computed by Eq. 5.17 and Eq. 5.18 respectively. With $a$ and $b$ known, for any given source value, the target value can be predicted using Eq. 5.8.

### 5.2.2 N\textsuperscript{th} order conversion function

The mean-variance conversion method is based on the assumption that formant values belong to a Gaussian distribution, and the conversion function is constrained to a linear mapping. In order to ease the assumptions about the distribution of the formants values, the N\textsuperscript{th} order conversion method was investigated. Given data point $x$, data point $y$ can be estimated by an N\textsuperscript{th} order polynomial, which can be expressed as below:

$$y = \sum_{i=0}^{n} a_i x^i = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$$  \hspace{1cm} (5.19)

Where, $a_0, a_1, \ldots, a_n$ are the parameters of the N\textsuperscript{th} order polynomial and can be determined in the training phase. The training process can be outlined as below:

**Step 1** The first step is to acquire a set of training sentences spoken by the participant in two different accents. The same sentences, utterances and words as in the mean-variance method were used in this method.

**Step 2** Using the formant frequency values (F1, F2 and F3) of each “word” obtained from the training dataset, produce a scatter map by plotting the
formant frequency values of the source (Accent BM) against the formant frequency values of the target (Accent LP).

Step 3  Given all the data points, the least square method which minimizes the prediction errors was used to obtain an Nth order polynomial that best fits the data. In this study, only the second order polynomials-quadratic was investigated and used for comparisons with the other two methods.

5.2.3 GMM conversion function

A Gaussian mixture model is a multivariate probability distribution model. In this method, the distribution of the training data is approximated by multiple Gaussian distributions; it is capable of modelling an arbitrary distribution. The probability distribution of a GMM with $Q$ components is given by:

$$P(x) = \sum_{q=1}^{Q} a_q N(x; \mu_q, \Sigma_q), \quad \sum_{q=1}^{Q} a_q = 1, \quad a_q \geq 0,$$  \hspace{1cm} (5.20)

where $a_q$ denotes the prior probability of $x$ having been generated by component class $q$, $Q$ denotes the number of component classes and $N(x; \mu_q, \Sigma_q)$ denotes a normal distribution with mean vector $\mu_q$ and covariance $\Sigma_q$, its probability density given by:

$$N(x; \mu_q, \Sigma_q) = \frac{1}{\sqrt{2\pi} \sqrt{|\Sigma_q|}} \exp\left(-\frac{1}{2} (x - \mu_q)^T \Sigma_q^{-1} (x - \mu_q)\right)$$  \hspace{1cm} (5.21)

The probability of a data point $x$ belong to a particular component class $q$ can be computed using Bayes’ rule, which is expressed as [80]:

$$P(q|x) = \frac{P(x|q) P(q)}{P(x)},$$  \hspace{1cm} (5.22)
The parameters \((a, \mu, \Sigma)\) of a GMM can be estimated from the training data using the expectation maximization (EM) algorithm [81]. In this research, the joint probability density method [82, 83] was used to estimate Gaussian parameters. In this method, the parameters \((a, \mu, \Sigma)\) were determined by the joint probability density of the source and target features. To do this, during the training phase, the combination of the source and target acoustic features was used to estimate the GMM parameters \((a, \mu, \Sigma)\) for the joint probability density \(p(x, y)\), where, \(X\) represents formant frequencies of Accent BM (BMF1, BMF2, BMF3) and \(Y\) represents formant frequency of Accent LP (LPF1, LPF2 and LPF3). \(N\) represents the number of the words in the training dataset and \(K\) represents the number of formants. To start the process, set \(a_q\) equal to \(1/Q\) for all \(q=1 \ldots Q\), \(\Sigma_q\) equal to an identity matrix for all \(q\); the K-means algorithm [84] was used for estimating the initial value of \(\mu_q\). The EM algorithm was then run until either the increase in log-likelihood of \(P(x, \alpha, \mu, \Sigma)\) was less than 0.0001 or 100 iterations were reached.

After the parameters \((a, \mu, \Sigma)\) were estimated, given source feature \(x\) (BMF1, BMF2, BMF3), the target feature \(y\) (LPF1, LPF2, LPF3) can be predicted by the following equation [80]:

\[
y = E[y \mid x] = \sum_{q=1}^{Q} (\mu_q^y + \Sigma_q^{xy} (\Sigma_q^{xy})^{-1} (x - \mu_q^x)) \cdot P(\lambda_q \mid x)
\]  

(5.23)

With
\[ \Sigma_q = \begin{bmatrix} \Sigma_{qq}^{XX} & \Sigma_{qq}^{XY} \\ \Sigma_{qq}^{YX} & \Sigma_{qq}^{YY} \end{bmatrix} \quad \text{and} \quad \mu_q = \begin{bmatrix} \mu_q^X \\ \mu_q^Y \end{bmatrix} \] (5.24)

Where \( \mu_q^X \) is mean value of \( X \) in the component class \( q \), \( \mu_q^Y \) is mean value of \( Y \) in the component class \( q \), \( \Sigma_{qq}^{XY} \) is the covariance of \( X \) and \( Y \) in the component class \( q \), \( \Sigma_{qq}^{XX} \) is the variance of \( X \) in the component class \( q \), \( P(\lambda_q \mid x) \) is the probability of a data point \( x \) belonging to a particular component class \( q \).

Let

\[ M_q = \Sigma_q^{XY} (\Sigma_q^{XX})^{-1} \] (5.25)

\[ B_q = \mu_q^Y - \Sigma_q^{XY} (\Sigma_q^{XX})^{-1} \mu_q^X \] (5.26)

Eq. 5.23 became

\[ y = \sum_{q=1}^{Q} (M_q x + B_q) \cdot P(\lambda_q \mid x) \] (5.27)

where \( M_q \) is a transformation matrix and \( B_q \) is a bias vector of class \( q \).

From Eq. 5.27, it can be seen that the GMM is a piece-wise linear regression, the regression is formulated as a weighted sum of linear regression, where the weights correspond to the posterior probability of a given input belonging to a particular class.

The quality of the transformed parameters depends on the number of mixtures; the parameters of the converted speech will be over-smoothed if too few mixtures are used for developing the GMM model. When the number of mixtures increases, the model can better match the real distribution of the time sequence. However, if the number of mixtures is too big, discontinuities may be introduced into the converted speech, causing the quality of the converted speech to be degraded to various degrees [85]. Computing cost will also increase when a larger component
number is chosen, since more parameters need to be estimated. Therefore, choosing the optimal number of mixtures is important in GMM conversion. In this study, GMMs with 2 mixtures, 4 mixtures, 6 mixtures, 8 mixtures, 10 mixtures and 12 mixtures are investigated and compared.

The performance of a GMM with a different number of mixtures is evaluated by average prediction error, which is the average value of the Euclidean distance between transformed formants and target formants. The average prediction error was calculated as follows:

$$\sqrt{\frac{1}{N} \sum_{i=1}^{N} (\text{predict}_i - \text{target}_i)^2}$$  \hspace{1cm} (5.28)

where $N$ is the number of data pairs in the training dataset.

The average prediction errors of the GMM with six different mixtures are shown in Table 5.1.

**Table 5.1** Comparison of the average prediction error of Gaussian mixture model with different mixture numbers

<table>
<thead>
<tr>
<th>Number of mixtures</th>
<th>Prediction error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
</tr>
<tr>
<td>2</td>
<td>6.1003</td>
</tr>
<tr>
<td>4</td>
<td>5.9624</td>
</tr>
<tr>
<td>6</td>
<td>5.8611</td>
</tr>
<tr>
<td>8</td>
<td>5.8276</td>
</tr>
<tr>
<td>10</td>
<td>5.7723</td>
</tr>
<tr>
<td>12</td>
<td>5.8243</td>
</tr>
</tbody>
</table>

From Table 5.1, it can be seen that when the number of mixtures increases, the prediction error decreases slightly for the three formants. Since the differences in
the prediction errors are very small, (for example, for formant F1, the prediction errors are within the range of 5.77 to 6.10), and using too many mixtures may introduce discontinuities into the converted speech, therefore, a GMM with 4 mixtures was chosen for comparison with the other two algorithms for use in accent conversion.

The same training data used in the mean-variance and N\textsuperscript{th} order conversion was used in GMM. Figures 5.4-5.6 show the GMM regression with 4 mixtures along with mean-variance linear prediction and quadratic regression for formants F1, F2, F3 respectively. Black circles represent the training data; GMM approximation is illustrated in green; the quadratic approximation is illustrated in red and the linear mean-variance conversion is illustrated in black.

![Figure 5.4](image.png)

**Figure 5.4** GMM regression line along with mean-variance linear approximation and quadratic regression for formant F1
Figure 5.5 GMM regression line along with mean-variance linear approximation and quadratic regression for formant F2

Figure 5.6 GMM regression line along with mean-variance linear approximation and quadratic regression for formant F3
Comparing the approximations made by three different algorithms, it can be seen that the GMM with 4 Gaussian mixtures and quadratic approximated the distribution of the data more closely than the linear mean-variance method, especially when the data is more scattered such as the beginning and the end parts. The average prediction error of the three approximation methods was calculated by Eq. 5.28 and is listed in Table 5.2. It can be concluded that the 4 mixture GMM gives the best approximation among the three methods.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Prediction error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
</tr>
<tr>
<td>Mean-variance</td>
<td>6.2921</td>
</tr>
<tr>
<td>Quadratic</td>
<td>6.1355</td>
</tr>
<tr>
<td>4GMM</td>
<td>5.9603</td>
</tr>
</tbody>
</table>

**5.3 Spectral conversion**

Since the first three formants were used as representation of speech signal spectra in this study, the spectral conversion process was actually a process of modifying the first three formants values that are extracted from the source utterance towards that of the target utterance. In this study, the utterances spoken in Accent BM were used as source utterances and the utterances spoken in Accent LP were used as target utterances; this assumption was also used in the experiments in Chapter 6. In the conversion stage, the first three formant values of the input utterance in Accent BM were modified on a frame by frame basis using conversion functions which derived from the training data set processed by different mapping algorithms. In this study, the three mapping algorithms, as described previously, mean-variance method, quadratic non-linear method and GMM method were investigated and compared.
5.4 Pitch contour modification

Pitch contour refers to changes of the fundamental frequency F0 over time and is perceived as intonation. As F0 is only defined for voiced parts of a speech signal, the pitch contour is not continuous; it has positive values for voiced regions and gaps for unvoiced regions. It is known that pitch contour reflects expression of emotion as well as linguistic features such as the type of utterances. For example, most declaratives have an overall declining pitch contour, but yes-no questions usually have an extreme final upturn in the last intonation phrase. Figure 5.7 illustrates the pitch contours for a question sentence “Will we ever forget it?” spoken by a British female speaker in Accent BM and Accent LP, and Figure 5.8 illustrates the pitch contours for a statement sentence “There was a change now.” from the same speaker in Accent BM and Accent LP.

Integrating the conversion of certain pitch parameters into a voice conversion system for improving the naturalness of synthesized speech has been studied [7, 73]. Those studies showed that pitch contour conversion could improve the naturalness of output speech to some extent in a voice conversion system. In this section, the effect of pitch contour conversion on the quality of output speech in accent conversion was investigated. The result of vowel-based pitch analysis for Accent BM and Accent LP showed that the two accents had different pitch slopes from the same speaker. In order to determine the effect of pitch contour conversion on the perception of accent transformation, a pitch contour transplantation technique was used for modifying the pitch contour of the spectral converted utterance.
Figure 5.7  The pitch contour of sentence “Will we ever forget it?” generated by a speaker in Accent BM (top) and Accent LP (bottom)

Figure 5.8  The pitch contour of sentence “There was a change now.” generated by a speaker in Accent BM (top) and Accent LP (bottom)
5.4.1 Time alignment

The durations of linguistic units (e.g. phonemes, syllables) differ from speaker to speaker; even for the same speaker, when uttering the same sentence repeatedly, the duration differs each time. In order to ensure that the pitch contour will be accurately transplanted to the corresponding linguistic units, the two utterances must be time aligned in the time domain before performing pitch contour transplantation.

The aim of time alignment of two signals is to compress or stretch one signal in the time domain to match the length of the other signal, resulting in the two signals having the same length; the time warped speech signal should not be distorted in any way and must still be intelligible. Several methods for aligning two signals in the time domain exist; the overlap and add method [86], and the dynamic time warping (DTW) [62, 66] algorithm are the two most popular methods which have been used in previous research. Other approaches such as forced-alignment by applying hidden Markov model (HMM) [87] have also been proposed.

In this study, two time-alignment algorithms, dynamic time warping (DTW) algorithm, and pitch synchronous overlap and add (PSOLA) method, were studied.

5.4.1.1 Dynamic time warping

The dynamic time warping problem can be stated as follows: Given two time series $X$ and $Y$, of length $m$ and $n$:

$$X = x_1, x_2, x_3, \ldots, x_m$$

(5.29)

$$Y = y_1, y_2, y_3, \ldots, y_n$$

(5.30)
Construct a warp path $W$:

$$W = w_1, w_2, w_3, ..., w_K \quad \text{max}(m, n) \leq K < m + n \quad (5.31)$$

where $K$ is the length of the warp path and the $k^{th}$ element of the warp path is

$$w_k = (i, j) \quad (5.32)$$

where $i$ is an index from the time series $X$, and $j$ is an index from time series $Y$. The warp path must start at the beginning of each time series at $w_1 = (1, 1)$ and finish at the end of both time series at $w_K = (m, n)$. This ensures that every index of both time series is used in the warp path. There is also a constraint on the warp path that forces $i$ and $j$ to be monotonically increasing in the warp path, which is why the lines representing the warp path in Figure 5.9 do not overlap. The fact that every index of each time series must be used can be stated as follows:

$$w_k = (i, j) \quad w_{k+1} = (i', j') \quad i \leq i' \leq i + 1, \quad j \leq j' \leq j + 1 \quad (5.33)$$

The optimal warp path is the path which has the minimum-distance cost, and the distance of a warp path $W$ can be expressed by the following equation:

$$Dist(W) = \sum_{k=1}^{k=K} Dist(w_k, w_{k+1}) \quad (5.34)$$

$Dist(W)$ is the distance of warp path $W$, and $Dist(w_k, w_{k+1})$ is the distance between the two data point indexes (one from $X$ and one from $Y$) in the $k^{th}$ element of the warp path.

In the dynamic time warping, Mel-frequency cepstrum coefficients (MFCCs) were used as spectral representation of the speech signal. Firstly, the two speech signals which need to be time aligned were pre-emphasized with a pre-emphasis factor of $\alpha = 0.97$ to compensate for the natural spectral roll off of speech signals.
that occurs at high frequencies. Then 12\textsuperscript{th} order MFCCs were calculated by applying a Hamming window with the frame size of 1000 samples (about $1000/44100 = 0.023$ s = 23 ms) and the distances of the spectra of each frame between two signals were computed to obtain a distance matrix. Secondly, the minimum-cost path of the distance matrix with constraint of three steps (1, 1), (1, 0) and (0, 1) with equal weights was found. At the last step, the two signals were aligned using the corresponding frame index.

Figure 5.9 shows the distance matrix of MFCCs and the optimal warp path between two utterances, which is the sentence “The nightglow was treacherous to shoot by.” uttered by a speaker in the two accents, Accent BM and Accent LP. The red line is the optimal cost path from the beginning to the end of both signals. The matrix shows the distance of the MFCCs between the two speech signals, with darker blocks indicating bigger distance values, and lighter blocks indicating smaller distance values. Therefore, the DTW path tends to pick lighter blocks to minimize the distance cost, hence maximizing the matching performance. Figure 5.10 shows a short segment of two time aligned signals. For illustration purposes, each signal is represented by the value of its second MFCC in each frame, and the value of MFCC of signal $X_2$ were moved down by 10 units along the Y axis. Figure 5.11 shows the waveforms of the sentence in the two accents; it can be seen that the two utterances have different durations. Figure 5.12 shows the waveforms of the two time aligned utterances. The target utterance which is spoken in Accent LP is shown in red and the modified source utterance which is spoken in Accent BM is shown in black.
Figure 5.9 The distance matrix of MFCCs and the optimal warp path

Figure 5.10 A short segment of two time aligned signals
Figure 5.11  The waveforms of the sentence “The nightglow was treacherous to shoot by.” in Accent BM (a) and Accent LP (b)
5.4.1.2 Pitch synchronous overlap and add

The overlap and add method has been used to modify the duration of speech in the time domain in previous research [86]. However, using a simple time scale method for modifying the duration of a speech signal will introduce distortion into the modified speech, such as lowered or raised pitch. In order to overcome this problems, the pitch synchronous overlap and add (PSOLA) method was proposed [62, 88]. In this method, the input segments of speech signal are chosen pitch synchronously; this allows the modification of the signal $x(t)$ in the time domain without altering its perceived pitch frequency. A typical time domain PSOLA involves the following main procedures:

1) A pitch detection algorithm is first applied to determine the pitch markers of the speech signal, and then the speech signal is divided into separate but overlapped smaller signals. This is accomplished by windowing the signal around each pitch mark. One segment usually contains 2~4 pitch periods. As the pitch periods vary during speech, the size of segmenting window will be variable depending on the pitch periods.

Figure 5.12 The waveforms of two time aligned utterances using DTW method (black Accent BM, red Accent LP)
2) Signals are modified by either repeating or leaving out some segments, depending on target length and source length. Repeat the small overlapping segments to increase duration; eliminate the small overlapping segments to decrease duration. In order to maintain original pitch, the repeated or overlapped segments contain a whole number of pitch periods.

Figure 5.13 illustrates the process of duration modification through the overlap and add method: the following steps in the procedure are shown, a) original signal segmented by overlapped windows; each segment contains 4 pitch periods; b) duration increased by repeating part of speech segments; c) duration decreased by omitting part of speech segments.
Chapter 5  Accent conversion

a) Original signal and segmentations

b) Duration increased by repeating part of overlapped intervals

c) Duration decreased by deleting part of overlapped intervals

Figure 5.13  Duration modification of a voiced speech segment
In the analysis in this research using the PSOLA method, first, the source and the target utterances were manually segmented into “word” level. Then each word from the source utterance and target utterance was time aligned using the PSOLA method. Any silence regions were directly copied from the target utterance. After each word from the source utterance was aligned with the parallel word from the target utterance, each time aligned word was then concatenated together to form a modified speech waveform which had the same length as the target utterance. Figure 5.14 shows the waveforms of the two time aligned utterances using the PSOLA method; the same utterances used in the DTW method were used here. The target utterance which is spoken in Accent LP is shown in red and the modified source utterance which is spoken in Accent BM is shown in black.

The quality of the time warped utterances were compared with the original utterances. It was noticed that the quality of warped speech had been degraded in the DTW method and the PSOLA method achieved a better result. The PSOLA method was therefore used for time aligning the paired test utterances in the Chapter 6 experiments.
5.4.2 Pitch contour modification

Most current voice conversion systems take a simple scaling approach for pitch transformation; in this method a constant scale factor is used to adjust the pitch of the source speaker. This constant scale can be the ratio of the target speaker’s average pitch value to the source speaker’s average pitch value. Some voice conversion systems employ mean-variance linear conversion method for pitch modification [7]. Since the average pitch and its range are key acoustic features of a speaker identity and pitch contour is more accent correlated, during pitch contour modification, fundamental frequency changes for the phonetic segments are taken from the target utterance, while mean and variance of F0 are from the source utterance. This ensures that the pitch contour is copied over but the average pitch value F0, which is important to speaker identity, remains unmodified. The following steps take place during pitch contour modification:

Step 1 Use PSOLA method to modify the duration of the source utterance towards the duration of the target utterance.

Step 2 Extract pitch contour from both utterances using autocorrelation method, and then convert pitch contour to a pitch tier object. The pitch tier is one of the types of objects in the Praat software; it contains a sequence of pitch points, the pitch between points is represented by a linearly interpolated line. The pitch tier object is used for manipulating the pitch contour of a sound and for synthesizing a new sound [35]. During pitch contour modification, the target pitch tier is shifted by a constant value which is the difference between the two utterances’ average pitch values to obtain a modified pitch tier object. For example, when the average pitch values from the source utterance and the target utterance are 150 Hz and 180 Hz respectively, then the target pitch tier will be shifted down by 30 Hz to match the average value of the source utterance, thus the modified pitch
contour has the shape of the target utterance pitch contour but maintains the average fundamental frequency of the source utterance.

Step 3 Convert the sound object of the source utterance to a manipulation object, and then replace the pitch tier with the modified pitch tier.

Step 4 Get the resynthesized utterance from the manipulation object with modified pitch contour by PSOLA method.

Figure 5.15 shows the original pitch contour (black), target pitch contour (green) and converted pitch contour (red) of the sentence “there was a change now.” It can be seen that the pitch contour of the target utterance (green) has been transplanted onto the converted speech (red), while the average fundamental frequency of the source utterance is maintained without alteration.

Fig. 5.15 Pitch contour of sentence “there was a change now” spoken by a female speaker in Accent BM (black) and Accent LP (green) and converted pitch contour (red)
Chapter 6
Experiments and evaluations

In this chapter, four experiments were conducted to produce synthesized utterances for objective and subjective evaluations. In the first experiment, the speech signal was analysed and resynthesized with no modifications on either formants or pitch contour; this experiment was used for evaluating the effect of the algorithm used by the Praat software on the quality of the synthesized utterance. In the second experiment, accent conversion based on spectral mapping without pitch contour modification was performed; the aim of this experiment was to evaluate the three mapping algorithms for spectral conversion. Then the best algorithm was used in the third experiment. In the third experiment, accent conversions were performed by combining spectral mapping with pitch contour modification to evaluate the effect of pitch contour on accent conversion. The fourth experiment was designed to evaluate the performance of the accent acoustic model which was trained from a speech corpus generated by a single speaker, and was applied to a different speaker.

Objective evaluation and subjective evaluation were conducted on the converted speech and are reported in this chapter. The last five sentences of Group B in Appendix I, which were not included in the training dataset, were used for the testing. Objective evaluations were used to compare the three mapping algorithms in terms of the target parameters predictions. The subjective evaluations were used to evaluate the three mapping algorithms in terms of perception of the target accent in the converted utterance and the effect of pitch contour modification on accent conversion. The subjective evaluations mainly focused on two aspects: accent identity and voice identity. In all the experiments, utterances spoken in Accent BM were used as the source accent and utterances spoken in Accent LP were used as the target accent.
6.1 Spectral mapping without pitch contour modification

In this experiment, the linear prediction (LP) residual from the target utterance was used as excitation signal. The LP residual was calculated by applying a standard LP inverse filtering on the original speech. A previous study [89] shows that the residual and the high frequency spectrum contain important information about speaker identity. In order to ensure that the voice identity of the source speaker was affected as little as possible during conversion, the formants below 3.5 kHz were modified and the excitation signals were taken from the source utterance. The experiment follows the first four steps of the accent conversion procedure described in Chapter 5.

6.2 Spectral mapping with pitch contour modification

In order to implement the pitch contour transplantation, the two utterances need to be time aligned. Firstly, the PSOLA method was used to modify the duration of the source utterance to match the length of target utterance. Secondly, performing spectral mapping without pitch contour modification as described in section 6.1, at this stage only the spectra had been modified in the converted utterances; the pitch contour remained intact. Thirdly, the procedure of pitch contour modification in section 5.4.2 was applied to the converted utterances. Finally, the converted utterances with pitch contour modification were obtained.

6.3 Objective evaluation

In Chapter 5, an objective evaluation of the three mapping algorithms which was based on the training database was presented. Here, objective evaluation based on source utterance and converted utterance was conducted. The normalised Euclidean distance between target formants and converted formants was used for objective evaluation. The target formants were extracted from the target utterance;
the converted formants were predicted from the source utterance’s formants using
the three predicting methods.

Prior to formants extraction, the PSOLA technique was used to modify the
duration of the source utterances in Accent BM towards the duration of target
utterances in Accent LP. This procedure ensures that the converted utterance has
the same number of frames as that of the target utterance. Then normalised
Euclidean distance between converted formants and target formants were
calculated and used to compare the performance of the three algorithms.

The formula used for normalised Euclidean distance calculation was as follows:

\[
\sqrt{\frac{1}{N} \sum_{i=1}^{N} (predictF_i - targetF_i)^2}
\]  

(6.1)

where \(N\) is the number of frames in the target utterance.

The normalised Euclidean distance of Formants F1, F2 and F3 for five testing
sentences are listed in Tables 6.1, 6.2 and 6.3 respectively.

**Table 6.1  The normalised Euclidean distance of formant F1 between target
utterance and converted utterance**

<table>
<thead>
<tr>
<th>Sentence</th>
<th>F1</th>
<th>F1</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean-variance</td>
<td>Quadratic</td>
<td>GMM(4-mixtures)</td>
</tr>
<tr>
<td>s16</td>
<td>40.9652</td>
<td>39.1556</td>
<td>38.3061</td>
</tr>
<tr>
<td>s17</td>
<td>45.4192</td>
<td>43.7180</td>
<td>43.3572</td>
</tr>
<tr>
<td>s18</td>
<td>32.2526</td>
<td>30.9260</td>
<td>30.4049</td>
</tr>
<tr>
<td>s19</td>
<td>30.7788</td>
<td>29.6471</td>
<td>29.2332</td>
</tr>
<tr>
<td>s20</td>
<td>39.7523</td>
<td>38.9285</td>
<td>37.3901</td>
</tr>
</tbody>
</table>
Table 6.2  The normalised Euclidean distance of formant F2 between target utterance and converted utterance

<table>
<thead>
<tr>
<th>Sentence</th>
<th>F2 Mean-variance</th>
<th>F2 Quadratic</th>
<th>F2 GMM(4-mixtures)</th>
</tr>
</thead>
<tbody>
<tr>
<td>s16</td>
<td>33.0034</td>
<td>30.3237</td>
<td>31.4695</td>
</tr>
<tr>
<td>s17</td>
<td>56.2180</td>
<td>53.7061</td>
<td>54.7704</td>
</tr>
<tr>
<td>s18</td>
<td>25.4003</td>
<td>23.3564</td>
<td>24.1646</td>
</tr>
<tr>
<td>s19</td>
<td>28.5643</td>
<td>27.0092</td>
<td>27.6379</td>
</tr>
<tr>
<td>s20</td>
<td>28.9936</td>
<td>26.7095</td>
<td>27.4321</td>
</tr>
</tbody>
</table>

Table 6.3  The normalised Euclidean distance of formant F3 between target utterance and converted utterance

<table>
<thead>
<tr>
<th>Sentence</th>
<th>F3 Mean-variance</th>
<th>F3 Quadratic</th>
<th>F3 GMM(4-mixtures)</th>
</tr>
</thead>
<tbody>
<tr>
<td>s16</td>
<td>29.6439</td>
<td>28.6275</td>
<td>26.8433</td>
</tr>
<tr>
<td>s17</td>
<td>35.1564</td>
<td>33.9699</td>
<td>30.4132</td>
</tr>
<tr>
<td>s18</td>
<td>24.8210</td>
<td>23.0821</td>
<td>22.1922</td>
</tr>
<tr>
<td>s19</td>
<td>26.1783</td>
<td>25.4138</td>
<td>25.9854</td>
</tr>
<tr>
<td>s20</td>
<td>25.4982</td>
<td>25.0688</td>
<td>23.8849</td>
</tr>
</tbody>
</table>

From Tables 6.1-6.3, it can be seen that the GMM with 4 mixtures conversion has the smallest distance. It gave the best prediction for first formant F1 and third formant F3 among the three algorithms. For second Formant F2, quadratic prediction was slightly better than the GMM mapping. However, GMM with 4 mixtures gives a better overall performance.

6.4 Subjective evaluation

For a voice conversion system, the quality of the system can be evaluated on the output along three major dimensions: intelligibility, naturalness, and speaker recognisability. In this case, since the focus was on accent conversion, the degree
of accent identification was the most important point to be addressed. A forced choice ABX testing method was used in the subjective (perceptual) evaluation. In this test, the listeners were asked to choose an answer from A and B to which X is closest. For example, stimuli A and B were speech uttered by speaker A and speaker B respectively; stimulus X was synthesized speech by different algorithms, and the listeners were asked to choose whether X is closer to stimulus A or stimulus B in terms of accent identity or voice identity.

Ten participants conducted perceptual evaluation independently via a website (http://www.webng.com/sarahzheng/listening_testing.html) or listened to the audio files in a quiet room with the author. During perceptual evaluation, the listeners were allowed to listen to the audio files as many times as they liked before making a decision.

Case 1 Evaluation of the effect of the algorithm used by the Praat software on the quality of the synthesized utterance.

The five testing utterances from Group B in Accent BM were analysed and resynthesized using software Praat. A 30th LP analysis was used for extracting the glottal excitation signal; Burg’s algorithm was used for formants extraction; then the digital filters were constructed from the extracted formants values. Finally the constructed digital filters were applied to the excitation signal to resynthesize the speech.

The listeners were asked to give a score out of 5 on the quality of the resynthesized utterances comparing them to the original utterances in terms of intelligibility, naturalness and pleasantness. The mean opinion score (MOS) is the arithmetic mean of all the scores from each individual. The result of these evaluations is listed in Table 6.4.
The following guidance of quality score was used:

1. very poor
2. poor
3. good
4. very good
5. excellent

Table 6.4 Evaluation of the effect of the algorithm used by Praat software on the quality of the synthesized utterance

<table>
<thead>
<tr>
<th>Participant</th>
<th>Quality scores of synthesized speech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sentence 1</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
</tr>
<tr>
<td>G</td>
<td>3</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
</tr>
<tr>
<td>I</td>
<td>4</td>
</tr>
<tr>
<td>G</td>
<td>3</td>
</tr>
<tr>
<td>MOS</td>
<td>2.9</td>
</tr>
</tbody>
</table>

The evaluation results in Table 6.4 indicate that the quality of resynthesized speech was degraded when using formants and LP residuals to resynthesize the speech. This might be due to the inherent disadvantage of the formant-based speech synthesis algorithm such as non-natural sounding speech output. In a
formant-based speech synthesis system, the filter which is used to modify the spectral shape of the glottal source excitation signal is characterized by slowly time-varying formant frequencies of the vocal tract which are derived from linear prediction coefficients. However, during the extraction of formants from linear prediction coefficients, some linear prediction coefficients are ignored in the conversion to formant-bandwidth pairs. This lost information can introduce the discontinuities into the output speech and cause the quality of synthesized speech to degrade, particularly in aspects of naturalness and pleasantness.

**Case 2 Evaluation of the three spectral mapping algorithms**

A preference test was conducted to evaluate the effect of different spectral conversion algorithms on accent perception. Without pitch contour modification, the pitch contours for all the converted utterances were identical to the source utterances.

The same five testing sentences were used in this experiment. In the test, original utterances and converted utterances were played. There were 15 utterances for each algorithm (5 original utterances in Accent BM, 5 original utterances in Accent LP and 5 converted utterances), therefore a total of 45 utterances were played in the test. Two accents are named as accent A (the source accent Accent BM) and accent B (the target accent Accent LP). During listening test, the original utterances in the two accents were played first, followed by the converted utterance. The listeners were asked to identify which accent (accent A or accent B) that the converted utterance was closest to, and tick the corresponding box. The results listed in Table 6.5 show the number of participants who chose that accent.
Table 6.5  Evaluation of the three spectral mapping algorithms on accent conversion

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Mean-variance</th>
<th>Quadratic</th>
<th>GMM</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accent A</td>
<td>Accent B</td>
<td>Accent A</td>
<td>Accent B</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>3</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>1</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>0</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>0</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>1</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>Percentage</td>
<td>90%</td>
<td>10%</td>
<td>92%</td>
<td>8%</td>
</tr>
</tbody>
</table>

To analyse the evaluation results, the percentage of participants who perceived the target accent (accent B) in the converted utterances were compared. The evaluation results in Table 6.5 showed that there was not much difference between the three mapping algorithms; however the GMM with 4 mixtures has a slighter better overall performance with a 16 percent (8/50) of the target accent was identified. This is coincident with the objective evaluation results in Section 6.3.

Case 3 Evaluation of the effect of pitch contour modification on accent conversion

In the previous test, the focus was on evaluating spectral mapping methods leaving pitch contour unchanged. In this experiment, the contributions of pitch contour modification to the perception of the target accent were evaluated.

The same five testing sentences were used in this test. Only the GMM algorithm was used in this experiment. In this test, the pitch contour from the target utterance was manipulated and transplanted onto the converted utterance. During the listening test, the original utterances in the two accents were played first, and
then followed by the converted utterance. The listeners were asked to identify which accent (Accent A or Accent B) that the converted utterance was closest to. This test included 10 original utterances and 5 converted utterances. The results listed in Table 6.6 show the number of participants who chose that accent.

**Table 6.6 Evaluation of the effect of pitch contour conversion on accent conversion**

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Spectral mapping with pitch contour manipulation</th>
<th>Spectral mapping without pitch contour manipulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Speaker A</td>
<td>Speaker B</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Percentage</td>
<td>28%</td>
<td>72%</td>
</tr>
</tbody>
</table>

The evaluation results in Table 6.6 indicated that pitch contour manipulation resulted in a higher perception of accent conversion. The average perception of the target accent is 16 percent (8/50) for accent conversion based on spectral mapping without pitch contour modification. With pitch contour modification, the perception of accent conversion was increased to 72 percent (36/50).

**Case 4 Evaluation of accent model applied to a different speaker**

In this experiment, the accent model of converting Accent BM to Accent LP, which was derived from the training data with the GMM algorithm, was applied to a different speaker. The source speaker is male with a Birmingham accent (BM). The same five testing sentences were used in this experiment. Firstly, the two paired utterances from the source speaker and the target speaker were time aligned using PSOLA method. Secondly, the formant frequencies of the source
utterance were extracted and the first three formants were modified using the transfer function to obtain the modified formants. Thirdly, the LP residual of the source speaker was used as the excitation signal; the modified formant frequencies were used for constructing the vocal tract filter, applying the constructed new filter onto the excitation signal to synthesize a new utterance. Finally, the pitch contour of the target speaker was manipulated and transplanted onto the synthesized speech.

As the source utterance and the target utterance were uttered by different speakers, voice identity and accent identity were evaluated in this test. During the listening test, the source utterance and the target utterance were played first, and then the converted utterance was played. The listeners were asked to identify whether the voice identity and accent identity of the converted speech is closer to Speaker A or Speaker B. The listeners were told to ignore the meanings of the utterances and concentrate on the voice identities of the speakers and their accents. The results listed in Table 6.7 show the number of participants who chose that accent.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Voice identity</th>
<th>Accent identity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Speaker A</td>
<td>Speaker B</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Percentage</td>
<td>82%</td>
<td>18%</td>
</tr>
</tbody>
</table>

Form Table 6.7, it can be seen that voice identity of sentence 5 has been correctly identified by all ten listeners, and for sentence 1 and 4, there are 9 out of 10 listeners identified the voice identity correctly. For accent identification, the highest score is 9 out of 10 listeners having perceived accent conversion for
sentence 3. Overall, 82 percent of participants indentified the correct speaker identity and 68 percent identified the correct accent.

6.5 Conclusions

From the results of the objective and subjective evaluations conducted in sections 6.3 and 6.4, it can be seen that the three spectral mapping algorithms achieved similar scores. The GMM-based algorithm achieved slightly better results in terms of prediction of target parameters and perception rate of the target accent in the converted utterances. In Case 3, the perception of the target accent in the synthesized utterance has been dramatically increased from 16 percent to 72 percent. This result indicated that the power of converting accents based on the mapping of F1, F2 and F3 contours was very limited; however pitch contour modification which changes the pitch contour of the source utterance towards the target utterance definitely improved the degree of accent conversion.

According to the perceptual evaluation results, the accent of the source utterance was converted to the target accent to varying degrees; and most of the voice characteristics of the source speaker were preserved in the converted utterance. It can be seen from the result of voice identification in Case 4, that the highest score for voice identification is 100 percent (10 out 10) with an average of 82 percent. For accent identification, the highest score is 90 percent (9 out 10) with the average of 68 percent.

Overall, it can be concluded that accent conversion can be achieved by formant frequencies mapping with pitch contour modification, though the results still need to be improved further, such as increasing the size of the training dataset to improve the accuracy of transfer function parameters estimation, using high-dimensional acoustic features such as MFCCs, LPCCs to improve the quality of synthetic speeches.
Chapter 7

Conclusions and future work

7.1 Conclusions

This thesis has presented a new approach to identifying the acoustic characteristics of two English regional accents. The thesis identified the most accent-influential acoustic features, which were then used to develop mapping functions between the two accents. Furthermore, three mapping algorithms used for formant frequencies transfer were investigated and evaluated by means of objective and subjective evaluations.

Since the variability between speakers contributes significantly to the variants of acoustic features, in order to eliminate the influence of speaker’s variability, the utterances used for accent acoustic analysis in this study were uttered by the same speaker with different accents. This approach ensured that the difference between the spectra of phone for each accent was directly analysed without including the variances due to different speakers. The result of vowel-based acoustic analysis of the two accents indicated that the formants frequencies, the changes of the fundamental frequency are the most prominent acoustic features influenced by accent followed by phone duration.

The first three formant frequencies F1, F2 and F3 were used for spectral mapping between the two accents and the techniques of pitch contour modification and transplantation were applied in this study to investigate the effect of the pitch contour on the accent conversion. The results showed that pitch contour modification had a large positive contribution to accent conversion and identification compared with formant-based spectral conversion alone.
7.2 Future work

In this study, limited speech data were used for accent analysis and spectral training. Speech segmentation was carried out manually; this is a time-consuming and a tedious process even at word level. In future, a larger speech corpus should be considered. A relatively accurate speech segmentation at phoneme level can be achieved by means of an automatic speech segmentation tool, such as HMM based speech segmentation. This will save the tedious work of manual segmentation and a more accurate transformation function can then be developed by employing an increased size of training data set at phoneme level.

In this study, spectral mapping between the two accents was performed via model-based transformation. In this method, a model which is based on various acoustic features of speech (first three formant frequencies) was built for the two accents. Then through a training phase, a transformation function was generated. In conversion phase, the trained transformation function was used for predicting target speech features from new set of source speech features. Finally, the predicted features were used to produce the final transformed speech at the synthesis stage. The advantage of the model based conversion is that it is text independent; it can convert unconstrained words and utterances into a desired voice or accent. However, the model based conversions, are heavily dependent on the accuracy of the model, which again depends on the size of the training dataset; an increased training data set can improve the accuracy of the model. In the future, the effect of the amount of the training data and the quality of speech data such as the strength and authenticity of the accents which the speaker produced can be explored.

Since there is only a small fraction of the speech which contains discriminative information about accent, and sometimes these distinctions are quite subtle and not easy to be picked up for some untrained ears, accent conversion is more
challenging than voice conversion in aspects of accent modelling and accent perceptual evaluation.

Accent conversion and its application to purposes such as language learning or manipulation of the accent of a film recording is still a challenging subject. The research in this thesis was based on only two specific British accents; however, it provides an alternative approach to exploring accent analysis and conversion.
References


References

34. http://accent.gmu.edu/
35. www.fon.hum.uva.nl/praat/
References


90. http://mi.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html.
Appendix I

Text material used in this study

**Group A** Each sentence was uttered 10 times repeatedly in each accent

Sentence 1 It was more like sugar.
Sentence 2 This is no place for you.
Sentence 3 A burst of laughter was his reward.
Sentence 4 He was an athlete and a giant.
Sentence 5 The issue was not in doubt.

**Group B** Each sentences was uttered once in each accent

Sentence 1 Will we ever forget it?
Sentence 2 If I ever needed a fighter in my life I need one now.
Sentence 3 There was a change now.
Sentence 4 It occurred to me that there would have to be an accounting.
Sentence 5 I had faith in them.
Sentence 6 She turned in at the hotel.
Sentence 7 I was the only one who remained sitting.
Sentence 8 We’ll have to watch our chances.
Sentence 9 Meanwhile I’ll go out to breathe a spell.
Sentence 10 He moved away as quietly as he had come.
Sentence 11 It was a curious coincidence.
Sentence 12 There was nothing on the rock.
Sentence 13 Anyway, no one saw her like that.
Sentence 14 The men stared into each other’s face.
Sentence 15 What was the object of your little sensation?
Sentence 16 There has been a change, she interrupted him.
Sentence 17 His face was streaming with blood.
Sentence 18  The nightglow was treacherous to shoot by.
Sentence 19  For a full minute he crouched and listened.
Sentence 20  You must sleep, he urged.

**Group C**  Paragraph was uttered once in each accent

Please call Stella. Ask her to bring these things with her from the store: six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.
Appendix II

Twelve vowels and the words from which the vowels were extracted

Table 1 Twelve vowels and the words from which the vowels were extracted

<table>
<thead>
<tr>
<th>Vowels</th>
<th>Example</th>
<th>Phonetic spelling</th>
<th>Word in this study</th>
<th>Phonetic spelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>barn</td>
<td>B AA N</td>
<td>laughter</td>
<td>L AA F T AX</td>
</tr>
<tr>
<td>AE</td>
<td>pat</td>
<td>P AE T</td>
<td>and</td>
<td>AE N D</td>
</tr>
<tr>
<td>AO</td>
<td>born</td>
<td>B AO N</td>
<td>more</td>
<td>M AO</td>
</tr>
<tr>
<td>AX</td>
<td>about</td>
<td>AX B AW T</td>
<td>sugar</td>
<td>S UH G AX</td>
</tr>
<tr>
<td>ER</td>
<td>burn</td>
<td>B ER N</td>
<td>burst</td>
<td>B ER S T</td>
</tr>
<tr>
<td>IH</td>
<td>pit</td>
<td>P IH T</td>
<td>his</td>
<td>HH IH Z</td>
</tr>
<tr>
<td>OH</td>
<td>pot</td>
<td>P OH T</td>
<td>not</td>
<td>N OH T</td>
</tr>
<tr>
<td>UH</td>
<td>good</td>
<td>G UH D</td>
<td>sugar</td>
<td>S UH G AX</td>
</tr>
<tr>
<td>UW</td>
<td>boon</td>
<td>B UW N</td>
<td>you</td>
<td>Y UW</td>
</tr>
<tr>
<td>AY</td>
<td>buy</td>
<td>B AY</td>
<td>like</td>
<td>L AY K</td>
</tr>
<tr>
<td>EY</td>
<td>bay</td>
<td>B EY</td>
<td>place</td>
<td>P L EY S</td>
</tr>
<tr>
<td>OW</td>
<td>loan</td>
<td>L OW N</td>
<td>no</td>
<td>N OW</td>
</tr>
</tbody>
</table>

*British English example Pronunciation Dictionary (BEEP) [90] has been used for phonemic transcription