

Mandarin Concurrent Syllable Recognition for Native Listeners: Contributions of Consonant,
Vowel and Tone

by

WANG, Tingyi

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy
in Industrial Engineering and Logistics Management

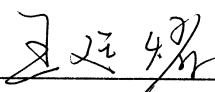
August 2020, Hong Kong

Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

A handwritten signature in black ink, appearing to read 'Wang Tingyi', is written over a horizontal line.

WANG, Tingyi


24 Aug 2020

Mandarin Concurrent Syllable Recognition for Native Listeners: Contributions of Consonant,
Vowel and Tone

by

WANG, Tingyi

This is to certify that I have examined the above PhD thesis
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by
the thesis examination committee have been made.



Professor Richard H.Y. SO (Supervisor)



Professor Guillermo GALLEGO (Head of the Department)

Department of Industrial Engineering and Decision Analytics

24 Aug 2020

Acknowledgements

First of all, I would like to extend my sincerest gratitude to my supervisor, Professor Richard SO for giving me years of guidance to gradually improve my understanding in the academic field and encouragement for me to keep studying. Thank you for letting me become more cautious and critical in reading and thinking. Thank you for your advice that make me pay more attention to every detail. Every lesson that I studied from you will benefit my future life the whole time.

High tribute shall be paid to my thesis examination committee members, Prof. Xiangtong QI, Prof. Ning CAI, Prof. Andrew HORNER and Prof. Gregory O'BEIRNE. Thank you for your profound knowledge and valuable comments for me to revise my works.

I'm also deeply indebted to all the tutors and teachers from the courses that I took during my study in HKUST for their patient and careful help.

Special thanks go to my groupmates who have helped me solve problems and provided their unique suggestions for me to conquer difficulties. Thank you Jun HUI, for the cooperation with me on the human-model comparison.

I shall also extend my gratitude to all subjects for giving patience and efforts when participated in my experiments.

Last but not least, I would like to thank my beloved family and all my friends for their encouragement and support during difficult times.

Table of Contents

| | |
|---|-----|
| Title Page | i |
| Authorization Page..... | ii |
| Signature Page | iii |
| Acknowledgements..... | iv |
| Table of Contents..... | v |
| List of Figures | ix |
| Abstract..... | xii |
| CHAPTER 1 INTRODUCTION..... | 1 |
| Summary..... | 1 |
| 1.1 Introduction and Motivation | 1 |
| 1.2 Definitions of concepts | 2 |
| 1.2.1 Vowel | 2 |
| 1.2.2 Consonant..... | 2 |
| 1.2.3 Tone..... | 5 |
| 1.2.4 Syllable..... | 5 |
| 1.2.5 Concurrent Vowel/Syllable Recognition | 5 |
| 1.2.6 Recognition accuracy | 6 |
| 1.3 Outline of the thesis | 6 |
| CHAPTER 2 LITERATURE REVIEW | 8 |
| Summary..... | 8 |
| 2.1 Introduction | 8 |
| 2.2 Speech perception and recognition..... | 8 |
| 2.3 Concurrent Vowel Recognition | 9 |
| 2.3.1 Concurrent Vowel Recognition in English | 9 |
| 2.4 Concurrent Syllable Recognition..... | 12 |
| 2.5 Isolated Syllable Recognition in Mandarin | 12 |
| 2.6 Speech Separation of Deep-learning Models | 13 |
| 2.7 Research Gaps | 14 |
| CHAPTER 3 EXPERIMENT ONE: EFFECTS OF SPECTRAL CUE ON MANDARIN CONCURRENT-VOWEL RECOGNITION | 16 |
| Summary..... | 16 |
| 3.1 Introduction | 16 |
| 3.2 Hypothesis | 16 |
| 3.2.1 Hypothesis 1 | 16 |
| 3.2.2 Hypothesis 2..... | 16 |
| 3.3 Method | 16 |

| | |
|--|----|
| 3.3.1 Variables..... | 16 |
| 3.3.2 Stimuli | 17 |
| 3.3.3 Sound calibration..... | 22 |
| 3.3.4 Tasks and procedure..... | 22 |
| 3.3.5 Subjects | 23 |
| 3.4 Results and analysis..... | 24 |
| 3.4.1 Effect of vowel spectral contrast on correct recognition rate..... | 25 |
| 3.4.2 Effect of vowel category on correct recognition rate | 27 |
| 3.4.3 Comparison of different measurement of correct recognition rate | 28 |
| 3.4.4 Effect of tone category on tone perception | 29 |
| 3.5 Discussions and conclusions | 30 |
| CHAPTER 4 EXPERIMENT TWO: EFFECTS OF SPECTRAL CUE ON MANDARIN CONCURRENT-SYLLABLE RECOGNITION: A CRITICAL CASE WITH SELECTED CONSONANTS | 32 |
| Summary..... | 32 |
| 4.1 Introduction | 32 |
| 4.2 Hypothesis | 33 |
| 4.2.1 Hypothesis 1 | 33 |
| 4.2.2 Hypothesis 2..... | 33 |
| 4.3 Method | 33 |
| 4.3.1 Variables..... | 33 |
| 4.3.2 Stimuli | 34 |
| 4.3.3 Task and Procedure | 36 |
| 4.3.4 Pilot tests | 37 |
| 4.3.5 Subjects | 39 |
| 4.4 Results and Analysis..... | 39 |
| 4.4.1 Effect of spectral contrast on syllable correct recognition rate | 39 |
| 4.4.2 Effect of consonant, vowel and tone category on syllable correct recognition rate | 42 |
| 4.4.3 Comparison of syllables, vowels and tones correct recognition rate | 43 |
| 4.5 Discussions and conclusions | 44 |
| CHAPTER 5 EXPERIMENT THREE: EFFECTS OF SPECTRAL CUE ON MANDARIN CONCURRENT-SYLLABLE RECOGNITION | 46 |
| Summary..... | 46 |
| 5.1 Introduction | 46 |
| 5.2 Hypothesis | 47 |
| 5.2.1 Hypothesis 1..... | 47 |
| 5.2.2 Hypothesis 2..... | 47 |

| | |
|--|----|
| 5.2.3 Hypothesis 3..... | 47 |
| 5.3 Method | 47 |
| 5.3.1 Variables..... | 47 |
| 5.3.2 Stimuli | 48 |
| 5.3.3 Task and Procedure | 51 |
| 5.3.4 Subjects | 52 |
| 5.4 Results and analysis..... | 53 |
| 5.4.1 Effect of spectral contrast on the syllable recognition correct rate | 53 |
| 5.4.2 Comparison of syllables, vowels and tones correct recognition rate | 57 |
| 5.4.3 Effect of tone category on the tone recognition correct rate | 58 |
| 5.4.4 Evaluate relative contributions of consonants, vowels and tones by model fitting | 59 |
| 5.5 Discussions and conclusions | 60 |
| CHAPTER 6 COMPARISON OF RECOGNITION ACCURACY BETWEEN HUMAN AND DEEP-LEARNING MODEL..... | 63 |
| Summary..... | 63 |
| 6.1 Introduction | 63 |
| 6.2 Hypothesis | 64 |
| 6.2.1 Hypothesis 1..... | 64 |
| 6.2.2 Hypothesis 2..... | 64 |
| 6.3 Method | 64 |
| 6.3.1 Model description..... | 64 |
| 6.3.2 Metrics for comparison | 65 |
| 6.4 Results and analysis..... | 67 |
| 6.4.1 Comparison of performance on concurrent-vowel recognition | 67 |
| 6.4.2 Comparison of performance on concurrent-syllable recognition..... | 71 |
| 6.5 Discussions and conclusions | 77 |
| CHAPTER 7 SUMMARY OF RESULTS AND DISCUSSIONS..... | 79 |
| 7.1 Summary of findings in the current study | 79 |
| 7.2 Concurrent vowel recognition in Mandarin..... | 79 |
| 7.3 Concurrent syllable recognition in Mandarin with four consonants involved | 80 |
| 7.4 Concurrent syllable recognition in Mandarin with extended selection of consonants ... | 81 |
| | |
| 7.5 Concurrent syllable separation by deep-learning model and a comparison with human | 83 |
| CHAPTER 8 CONCLUSIONS, LIMITATIONS AND FUTURE WORKS..... | 85 |
| 8.1 Conclusions | 85 |
| 8.1.1 Effects of spectral contrasts on concurrent-vowel recognition in Mandarin ... | 85 |

| | |
|---|-----|
| 8.1.2 Effects of spectral contrasts on concurrent-syllable (consonant + vowel) recognition in Mandarin | 85 |
| 8.1.3 Relative contributions of consonant, vowels, and tones on concurrent syllable recognition with Mandarin | 86 |
| 8.1.4 Comparison of recognition accuracy between human listeners and deep-learning model..... | 86 |
| 8.2 Limitations and Future Work | 87 |
| References and Bibliography | 89 |
| Appendix..... | 94 |
| 3.1 Waveplot of single stimuli used in experiment one..... | 94 |
| 3.2 Spectrum and spectral envelope of single stimuli used in experiment one..... | 97 |
| 4.1 Usage frequency table of consonants in mandarin..... | 100 |
| 4.2 Normalized spectral contrasts of stimuli in experiment two..... | 101 |
| 4.3 Post-hoc analysis results of interactions between factors | 103 |
| 4.4 Waveplot of single stimuli used in experiment two | 107 |
| 4.5 Spectrum and spectral envelope of single stimuli used in experiment two | 109 |
| 5.1 Combination methods of consonants and vowels | 111 |
| 5.2 Usage frequency table of vowels in mandarin..... | 114 |
| 5.3 Waveplot of single stimuli used in experiment three..... | 115 |
| 5.4 Spectrum and spectral envelope of single stimuli used in experiment three..... | 121 |

List of Figures

| | |
|--|----|
| Figure 1.1 Outline of the thesis | 7 |
| Figure 3.1 Waveforms of original stimulus generated from TTS tool (left) and adjusted stimulus used for mixing (right) | 19 |
| Figure 3.2 Waveform of concurrent-vowel pair example: “a1”+ “e2” | 19 |
| Figure 3.3 Spectral density and spectral envelope of vowel “a1” | 21 |
| Figure 3.4 Spectral envelopes of vowels “a1” and “u2” | 21 |
| Figure 3.5 Settings of calibration | 22 |
| Figure 3.6 Screenshot of the GUI used in experiment | 23 |
| Figure 3.7 Picture of audiometer used for hearing tests | 24 |
| Figure 3.8 Syllable correct recognition rate as a function of normalized spectral envelope contrast..... | 26 |
| Figure 3.9 Log-transformed syllable correct recognition rate as a function of normalized spectral envelope contrast..... | 27 |
| Figure 3.10 Average syllable recognition accuracy of six vowel categories with two combination groups (same vowel and different vowel) | 28 |
| Figure 3.11 Average correct recognition rate of tone, vowel and syllable..... | 29 |
| Figure 3.12 Tone confusion matrix in concurrent Mandarin vowel recognition..... | 30 |
| Figure 4.1 Classification table of mandarin consonants (selections of consonants used for experiment two were bolded with red color)..... | 35 |
| Figure 4.2 Screenshot of GUI used for Experiment 2 | 36 |
| Figure 4.3 Error rate of recognition from three subjects in pilot tests..... | 38 |
| Figure 4.4 Scatter plot of syllable correct rate as a function of normalized spectral envelope contrast | 40 |
| Figure 4.5 Scatter plot of syllable correct rate as a function of spectral contrast with vowel difference available..... | 41 |
| Figure 4.6 Scatter plot of syllable correct rate as a function of spectral contrast with only consonant difference available | 42 |
| Figure 4.7 Bar chart describing the averaged correct rate of consonant, vowel, tone and syllable with relative significant relations | 44 |
| Figure 5.1 Selected consonants in the classification table..... | 49 |
| Figure 5.2 Screenshot of GUI used in experiment | 52 |

| | |
|--|----|
| Figure 5.3 Scatter plot of syllable correct rate as a function of normalized spectral envelope contrast | 54 |
| Figure 5.4 Replot of syllable correct rate as a function of normalized spectral envelope contrast with aberrant points removed..... | 55 |
| Figure 5.5 Scatter plot of syllable correct rate as a function of spectral contrast with only consonant differences available..... | 56 |
| Figure 5.6 Scatter plot of syllable correct rate as a function of spectral contrast with only vowel difference available..... | 57 |
| Figure 5.7 Averaged correct recognition rate of consonants, vowels, tones and syllable, *** represented for $p < 0.001$ | 58 |
| Figure 5.8 Confusion matrix of tone recognition from concurrent-syllable pair | 59 |
| Figure 6.1 Model output of STOI and PESQ scores | 68 |
| Figure 6.2 STOI score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent vowel separation | 69 |
| Figure 6.3 PESQ score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent vowel separation | 70 |
| Figure 6.4 SDR score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent vowel separation | 71 |
| Figure 6.5 STOI score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent syllable separation (DCDV group) | 72 |
| Figure 6.6 PESQ score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent syllable separation (DCDV group) | 73 |
| Figure 6.7 SDR score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent syllable separation (DCDV group) | 74 |
| Figure 6.8 STOI score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent syllable separation (SVDC and SCDV groups) | 75 |
| Figure 6.9 PESQ score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent syllable separation (SVDC and SCDV groups) | 76 |
| Figure 6.10 SDR score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent syllable separation (SVDC and SCDV groups) | 77 |

List of Tables

| | |
|---|----|
| Table 1.1 Spectral features of unvoiced consonants in Mandarin..... | 4 |
| Table 3.1 Independent and dependent variables in experiment one..... | 17 |
| Table 3.2 Ranking of vowel pairs in terms of spectral contrasts value | 20 |
| Table 4.1 List of variables in experiment two | 34 |
| Table 5.1 List of variables of experiment three..... | 48 |
| Table 6.1 Parameters in training model..... | 65 |
| Table 6.2 Evaluation level of PESQ based on real listener experience..... | 66 |
| Table 7.1 Summarization on effects of spectral contrasts | 84 |

Mandarin Concurrent Syllable Recognition for Native Listeners: Contributions of Consonant, Vowel and Tone

by WANG, Tingyi

Department of Industrial Engineering and Decision Analytics

The Hong Kong University of Science and Technology

Abstract

Perceiving a target speech in a multi-speaker environment is common. Consequently, the ability to recognize a target syllable (vowel and consonant) in the presence of another syllable is essential for speech perception. Human ability to recognize individual vowels from the simultaneous presentation has been the subject of many studies in English. Nonetheless, studies to recognize mixed consonants and vowels from simultaneous presentations have received little attention. There are few studies in English, but similar studies in Mandarin could not be found.

This study examines syllable recognition when listening to concurrent pairs in Mandarin. In experiment 1, concurrent Mandarin vowels recognition was studied. The accuracy of recognizing two concurrent tonal vowels increased with the logarithmic spectral contrasts between the two vowels. This finding is new as past studies only showed linear increases. Experiments 2 and 3 extended the range of concurrent syllables by adding an initial consonant to vowels, which formed a “consonant + vowel (CV)”-pattern syllable. Results indicated that increasing both consonant and vowel spectral contrast can improve recognition performance. In addition, the recognition accuracy was further evaluated in terms of consonants, vowels and tones separately. A power function model was fitted to construct the relationship among consonants, vowels, tones and syllables recognition performance. The weighting coefficients of the model revealed that the relative contributions of consonants were more than vowels and tones in concurrent syllable recognition. Moreover, a deep-

learning model, trained to separate speech, was programmed to separate concurrent syllables. A comparison of human performance with the deep-learning models indicated that both human and machine performance was significantly affected by spectral differences in concurrent vowels. Interestingly, in concurrent syllables separation when different categories of consonants were involved, similar findings were found for model recognizing syllables differed on consonants. The effect of spectral contrasts was not significant when both vowel differences and consonant differences were presented. Possible insights in model training are discussed.

CHAPTER 1 INTRODUCTION

Summary

In the first chapter, the background of the study, including motivation, research problems, was briefly introduced. In addition, definitions of the concepts that would be much discussed in the study were presented for a better understanding.

1.1 Introduction and Motivation

Hearing is one of the essential abilities for human to perceive the surroundings and receive valuable information. However, challenges appear in a multi-speech-source environment when competing speeches existed and might mask the target signal we want. Similar auditory scenarios include talking and chatting in restaurants and online meetings where all attendees have the chance to speak. In such cases, more considerable difficulties of perceiving target speech arise from mixing voice that deteriorates the speech features we rely on to recognize. As a result, multiple auditory cues are needed for listeners to differentiate among simultaneously presented speech information.

Past studies showed numerous clues that human listeners would use in separating and recognizing concurrent speech. However, most studies were in English. As the second popular language in the world, mandarin speech in a multi-source environment received less attention. English and mandarin are two separate language systems consisting of different grammars and word structures. It has not been thoroughly studied that if cues functioned in English could play a consistent role in mandarin concurrent speech recognition. As a result, this study aimed to start the evaluation from a specific case, concurrent syllable recognition in mandarin.

Additionally, in recent years, the development of machine learning technology allowed for more possibilities in the reproduction of human ability by machine. In the auditory field, emerging works devoted to solving real problems by automatic speech recognition (ASR), speech enhancement and speech separation. The state of the art triggered our interests to compare the human and deep-learning models on the same recognition task. Such difference

in recognition accuracy may provide useful clues in model improvement based on experience in human

1.2 Definitions of concepts

1.2.1 Vowel

From the phonetic definition, vowel is a speech sound that pronounced with the open of vocal tract (Cruttenden, 2014). Vowels are produced by airflow through the mouth without constriction during pronunciation. In mandarin, there are 35 final vowels, which are listed with International Phonetic Alphabet (IPA) as follows:

(1) 6 simple vowels: a[a], o[o], e[ɛ], i[i], u[u], ü[y]

(2) 13 complex vowels: ai[ai], ei[ei], ao[au], ou[ou], ia[ia], ie[iɛ], iao[iɑu], iou[iəu], ua[ua], uo[uo], uai[uai], uei[uei], üe[yɛ]

(3) 16 compound nasal vowels: an[an], en[ən], ang[ɑŋ], eng[ɛŋ], ong[ɔŋ], ian[iɛn], in[in], iang[iɑŋ], ing[iŋ], ion[iɔŋ], uan[uən], uen[uən], uang[uɑŋ], ueng[uɛŋ], üan[yɛn], ün[yn]

Generally, there are 6~7 formants in the spectrum of vowels, but only 2~3 formants are needed to distinguish among vowels. So the spectral characteristics of vowels can be described by 2~3 formant frequencies in the main frequency range (Ladefoged & Johnson, 2014).

Due to the complex sound features of complex vowels and compound nasal vowels, only the six simple vowels were considered in this study.

1.2.2 Consonant

Besides vowel, consonant is another phonetic element comprise syllable. The articulation of consonant was done by complete or partial closure of the vocal tract.

In mandarin, there are 21 initial consonants that can be combined with the final vowels to form a complete syllable: b[p], p[p^h], m[m], f[f], d[t], t[t^h], n[n], l[l], g[k], k[k^h], h[x], j[tɕ], q[tɕ^h], x[ɕ], zh[tʂ], ch[tʂ^h], sh[ʂ], r[ʐ], z[ts], c[ts^h], s[s]. A main feature of consonants is that, during articulation, the airflow in the mouth is subject to various obstacles. The process of pronouncing consonants could be seen as the process of how airflow encounter and conquer obstructions. Generally, consonants have relatively low loudness; the duration cannot be extended arbitrarily, and are not used for rhyming (Duanmu, 2007).

The classification of consonants mainly varied from two perspectives: place of articulation, and manner of articulation. First, categorized by part of oral organs that constitutes an obstacle during pronunciation, seven places of articulation and corresponding consonants are listed:

- (1) Bilabial: b, p, m;
- (2) Labial-dental: f;
- (3) Alveolar: d, t, n, l;
- (4) Dental: z, c, s;
- (5) Retroflex: zh, ch, sh, r;
- (6) Palatal: j, q, x
- (7) Velar: g, k, h.

Second, in terms of the manner of articulation including the category of producing sound and the modification mechanism of sound from mouth and larynx, manners could be further classified through different parts (Dow, 1972). According to the way to stop air passage, consonants are divided into:

- (1) Plosive: b, d, g, p, t, k
- (2) Fricative: f, s, sh, r, x, h
- (3) Affricate: z, zh, j, c, ch, q
- (4) Nasal: m, n
- (5) Liquid: l

Considering the aspiration, the above plosive and affricate categories are subdivided into the aspirated and unaspirated, which includes p, t, k, c, ch, q and b, d, g, z, zh, j respectively.

On the other hand, m, n, l, r are categorized as voiced consonants refer to the vibration of vocal cord, while other consonants are unvoiced. In general, voiced consonants also have musical sound properties. Like vowels, they have frequency formants on the spectrum. The unvoiced consonants are formed by part of the ruptured airflow from the articulation pronunciation organ, usually with the similar features as noise. Spectrum of such consonants is continuous and the energy only concentrated within some certain frequency ranges. The energy distribution features along frequency of each unvoiced consonants was listed in table 1.1(教育委员会 et al., 2013):

Table 1.1 Spectral features of unvoiced consonants in Mandarin

| Consonant | Frequency range with energy concentration (Hz) | Frequency of formants (Hz) |
|-----------|--|----------------------------|
| b | 200 ~ 1600 | 250 |
| d | 300 ~ 6000 | - |
| g | 250 ~ 6300 | - |
| p | 150 ~ 8000 | 1250, 2500 |
| t | 150 ~ 8000 | 200 |
| k | 120 ~ 6000 | 2000 |
| f | 150 ~ 10000 | 200, 6300 |
| s | 3000 ~ 10000 | 6300 |
| sh | 1500 ~ 8000 | 2500, 4500 |
| x | 2500 ~ 8000 | 3000, 5000 |
| h | 1000 ~ 3000 | 1500, 2500 |
| z | 3200 ~ 8000 | 6300 |
| zh | 2000 ~ 5000 | 400 |
| j | - | 5000 |

| | | |
|----|-------------|------|
| c | 4000 ~ 8000 | 6500 |
| ch | 250 ~ 8000 | 4000 |
| q | 2000 ~ 8000 | 3500 |

1.2.3 Tone

As a tonal language, mandarin mainly has four tone categories featured with distinctive pitch contours (Tsu-Lin, 1970). A syllable can be pronounced with different tones to have respective lexical meanings. In the writing, tones are often represented by a label above the vowels carried them (Sagart, 1999). Tone 1 has the feature of high level and labeled with a horizontal line in Mandarin. The voice is relatively even across the syllable when pronouncing tone 1. Tone 2 has a rising pitch contour at high level and labeled with a rising diagonal line. Tone 3 is the only tone that has two directions of pitch change, falling first and rising next, which forms distinctive dipping in perception. A line with a dipping break labels tone 3 in Mandarin. Tone 4 is a falling tone as it has high level at first but drops sharply. A dropping diagonal line was used to label it.

In general, four tones could be written as: ā á ǎ à for tone 1, 2, 3, 4 respectively.

1.2.4 Syllable

Syllables form a sequence of speech as units of organization. The elements of syllable include a nucleus which is a vowel most often, and optional initial and final margins which were consonants typically (De Jong, 2003). Words could be made up of single syllable, two syllables or multiple syllables so that are called monosyllable, disyllable or polysyllable respectively.

In Mandarin, a syllable includes a tone carried by the vowel. Since consonants are optional, a syllable can be a tonal vowel or a tonal vowel combined with initial consonants. For example, both “ā” and “shé” are syllables in mandarin.

1.2.5 Concurrent Vowel/Syllable Recognition

The concurrent vowel recognition means that, in each trial, a vowel is presented simultaneously paired with another vowel, for example, “a” and “u”, and listeners need to identify each syllable in the concurrent pair. Additionally, in Mandarin, a vowel always exist with a tone (vowel “a” with tone 1 is labeled as “a1”, pronounced as Mandarin word “阿”), so the concurrent vowel recognition in our study also included the identification of tones. The example stimulus in concurrent vowel recognition experiment is “a1” presented together with “u2”.

The concurrent syllable recognition adds consonants to form the stimuli and follows the same way in presenting syllables. In the experiment, a Mandarin syllable is presented concurrently with another, such as “sha1” and “bu2”. Listeners are asked to separate and identify each syllable by providing their choices on consonants, vowels and tones.

1.2.6 Recognition accuracy

The recognition accuracy is calculated as the percentage of correct recognition. Specifically, in concurrent vowel recognition, correct recognition of a syllable (tonal vowel) includes both correctness of tone and vowel. For example, a tonal vowel “a3” is only considered as correct when listeners recognize “a” for vowel and tone 3 (T3) for tones. In addition, when the evaluation of accuracy is further considered in terms of elements, which are vowels and tones in this case, the accuracy is counted only by the response of corresponding element. For example, a response “a2” of the presented speech “a3” was counted as correct in vowel recognition, wrong in tone recognition and thus wrong in syllable (tonal vowel) recognition.

Similarly, in concurrent syllable recognition, the accuracies of identifying syllables are based on correctness of consonants, vowels and tones. For example, recognition of syllable “ba3” is counted as correct only if the responses from listeners are “b”, “a”, “3” for consonants, vowels and tones respectively. For further analysis in terms of elements, the consonant accuracy could be calculated separately. A response of “d”, “a”, “4” for presented syllable “ba3” would be counted as wrong in consonant recognition, correct in vowel recognition, wrong in tone recognition, thus wrong in syllable recognition respectively.

1.3 Outline of the thesis

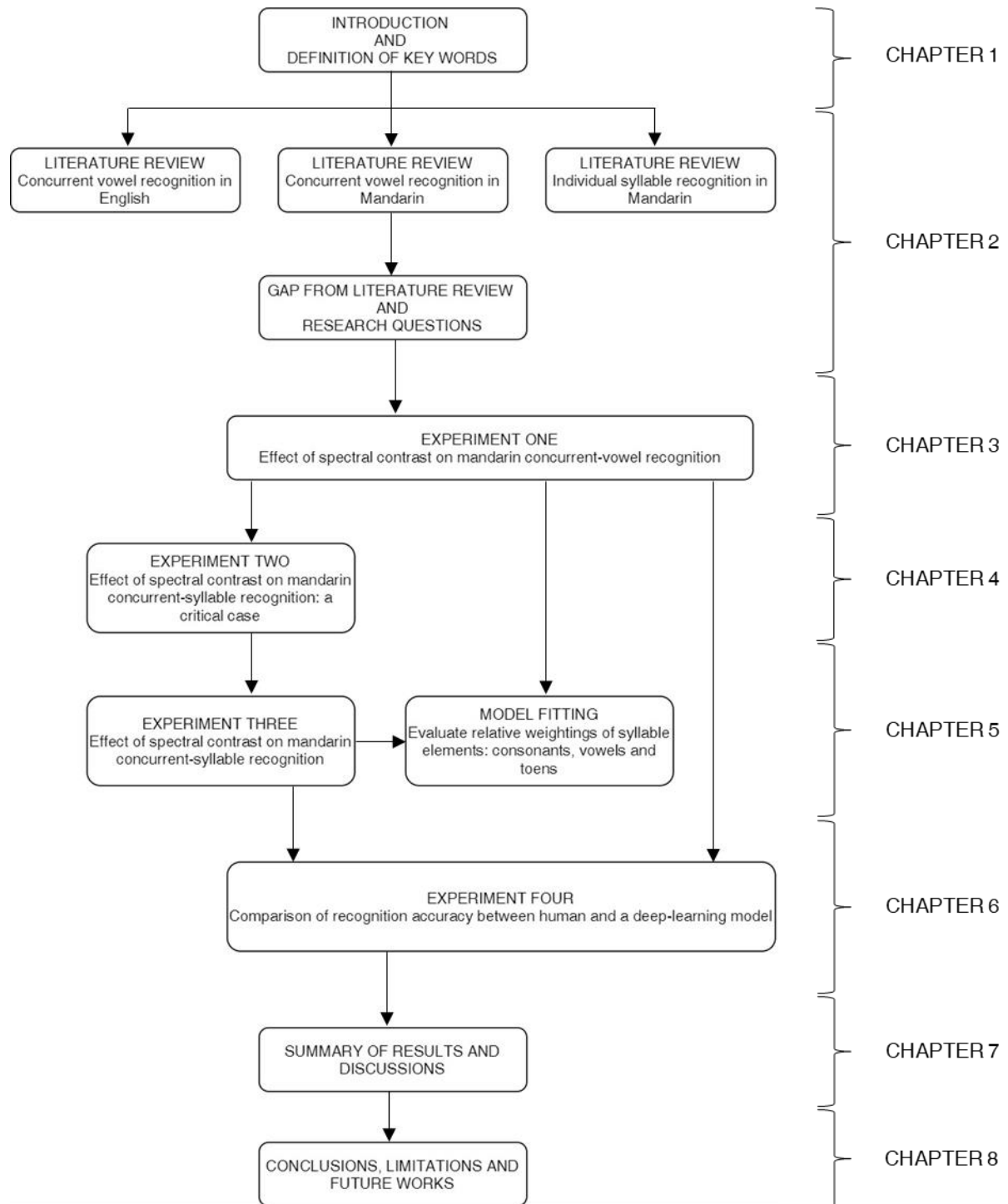


Figure 1.1 Outline of the thesis

CHAPTER 2 LITERATURE REVIEW

Summary

Studies on the perception of speech was first reviewed in understanding the importance of single syllable recognition. Acoustic cues that listeners used in concurrent vowel recognition in English and Mandarin were then discussed separately. In addition, the studies in concurrent syllable recognition were reported to show the research gaps in Mandarin. There only exist studies on contributions of consonants and vowels in isolated syllable recognition. Methodology for machine to separate speech based on deep-learning model was briefly introduced to highlight the development technology. At last, the research gaps and questions were summarized based on the literature review.

2.1 Introduction

The evaluation of human performance in speech recognition has a long history. Acoustic features of the speech sounds provide various cues for listeners to recognize. However, when competing speech existed, some cues surpass others(Fu et al., 2018) and interactions were also found between cues (Luo & Fu, 2009). Studies on speech recognition especially involved concurrent vowel recognition were reviewed to understand acoustic cues affecting human performance.

2.2 Speech perception and recognition

Cues involved in perception and recognition of speech were examined by many studies.

In a simulation of cochlear implant users' perception (Xu et al., 2005), information of temporal cues and spectral cues were both controlled at different presenting level. Results showed that both spectral cues and temporal cues enhanced speech recognition, while spectral cues outperformed temporal cues on the improvement of recognition accuracies.

Despite this, human was still able to recognize speech sounds with the largely degraded or absence of spectral cue. The study of human recognition in whispered speech revealed that manner of a start and stop provided partial cues for better performance (Fletcher, 1995). In a

case with deteriorated spectral cues, function of temporal envelopes on syllable recognition was evaluated by modulating noise with extracted temporal envelope while spectral information was largely excluded (Shannon et al., 1995). It was shown that recognition performance was sufficiently improved with the increase of temporal information, which indicated that temporal envelope played an important role in speech recognition.

Considered the phonetic elements of speech, the relative contributions of consonants and vowels to the recognition accuracy of monosyllabic words and sentences were evaluated (Fogerty & Humes, 2010; Fogerty & Kewley-Port, 2009). Both the consonants and vowels effectively affect words recognition. The performance significantly degraded when either consonants or vowels in the stimuli were replaced by noise. In recognition of sentences when context information available, vowel contributed more compared to only isolated words presented, while the contributions of consonants were equal in both conditions. Their studies showed the sentence recognition could much benefit from the information carried by vowels.

In mandarin, the perception of tones was also essential as different tones carried distinct lexical meanings even with the same structure of syllables. Studies have shown that besides the deterministic feature of tones, the pitch contour changes, cues like syllable duration, temporal envelope could provide useful information on recognizing tones (Fu & Zeng, 2000; Fu et al., 1998). The temporal envelope cues were reported to be dominant in Mandarin word recognition, while the temporal fine structure helped improve the average recognition accuracy to 90.8% in the conditions with most spectral information available (Xu & Pfingst, 2003).

2.3 Concurrent Vowel Recognition

2.3.1 Concurrent Vowel Recognition in English

In order to understand how human differentiate and recognize speech in multiple-speech-source environment, a simplified case was evaluated by plentiful studies. As the nucleus of syllable comprising speech, the recognition of concurrent vowels was evaluated through various perspectives.

Acoustic cues used for vowel identification was discussed in many studies. Frequency modulation was showed to improve the vowel recognition accuracy than unmodulated vowels (Culling & Summerfield, 1995; McAdams, 1989). Frequency modulation served as an alerting and thus made the modulated vowel prominent to be recognized (Culling & Summerfield, 1995).

The difference in fundamental frequency (F0) between vowels was another spectral cue (Duanmu, 2007; Meddis & Hewitt, 1992; Qin & Oxenham, 2005; Shackleton & Meddis, 1992; Vongpaisal & Pichora-Fuller, 2007) that could enhance the recognition of concurrent vowel pairs. Listeners performance benefit from enlarging of F0 difference between two talkers. By enlarging contrasts in F0 from 0-3 Hz, the recognition accuracies had an improvement of 30% at most. But this effect was asymptotic as the F0 differences kept increasing (Assmann & Summerfield, 1990; Chintanpalli & Heinz, 2013; Micheyl & Oxenham, 2010), representing a small increasement of F0 difference was sufficient for normal hearing listeners to utilize on differentiating and recognizing concurrent vowels.

However, the utilization of spectral information was limited due to the aging effect. Studies had shown that the ability in F0 information degraded in the older listeners group (Arehart et al., 2011; Chintanpalli et al., 2016; Snyder & Alain, 2005; Vongpaisal & Pichora-Fuller, 2007). Though the concurrent vowel recognition performance benefited from the improving difference of F0 for both younger and older groups, the elders had significant lower accuracies. Then event-related potentials (ERPs) were measured to show the aging affected listener's ability in utilizing spectral cues for segregation (Snyder & Alain, 2005).

Similarly, listeners suffered from hearing impairment had poorer performance on segregating concurrent vowels. Compared with normal hearing listeners, the enhancement that arising from the enlarging of F0 difference was less for hearing-impaired listeners in the same tasks (Arehart et al., 1997; Arehart et al., 2005; Summers & Leek, 1998).

Other cues such as harmonicity (Roberts & Holmes, 2006), onset asynchrony (Darwin, 1981; Hedrick & Madix, 2009; Hee Lee & Humes, 2012) also were shown to significantly influence recognition accuracies of concurrent vowels.

2.3.2 Concurrent Vowel Recognition in Mandarin

Studies on concurrent vowel recognition in mandarin was less than in English. Compared to English, the recognition of vowel in Mandarin involved the recognition of tone. Lexical tone contour were showed important in mandarin vowel identification (Chen et al., 2014).

Acoustic cues that was utilized by listeners in concurrent English vowels were partially studied in Mandarin. In simulated electric hearing, it was shown that spectral cues played an important role in recognition of concurrent vowels (Luo & Fu, 2009). In order to simulate the hearing of cochlear implants users, the different degree of spectral resolution was preserved in concurrently presented Mandarin vowels. In addition, the talker difference (difference of F0) was added as another factor. Results revealed that spectral resolution as well as talker difference significantly affect the recognition of concurrent vowels. Performance improved with higher degree of spectral resolution. When the spectral information was fully preserved, the male – female talker condition had higher accuracies on recognizing vowels, which was an evidence for the effect of mean F0 on concurrent vowel recognition.

Additionally, comparisons of performance between normal-hearing listeners and hearing-impaired listeners was evaluated to show different weightings of using cues for two groups (Fu, Yang, et al., 2019; Fu et al., 2018). In the recognition of concurrent Mandarin vowels, information from F0 contour difference and temporal envelope difference between two vowels in a pair significantly improved the identification performance for both group of listeners. However, for hearing-impaired listeners suffered from sensorineural hearing loss, temporal cues contributed more than spectral cues in the recognition performance. This implied a deteriorated ability in processing auditory signal through spectral information.

Another work also from Fu's team evaluated the concurrent Mandarin vowel recognition in terms of the effect from spectral contrasts between two vowels presented simultaneously (Fu, Wu, et al., 2019). For both listening groups, the enlarging of spectral differences between vowels significantly improved recognition of the tonal vowels in a linear scale. A further analysis subdivided the performance based on the recognition accuracy of tones and vowels. Relative contributions of temporal and spectral cues were showed different in two groups.

Normal hearing people relied more on temporal information in recognition task, while role of spectral cues was as important as temporal cues for hearing-impaired listeners.

2.4 Concurrent Syllable Recognition

In studies evaluating concurrent speech recognition, consecutive sentences provided contextual information that could be largely used for improvement in recognition accuracy (Cooke, 2006). However, in the recognition of concurrent syllables, such clues were missed and. As a result, it's also important to evaluate cues that dominant for concurrent syllable recognition.

The vocal tract length (VTL) and glottal pulse rate (GPR) of speakers were proved to have influence on the recognition of concurrent syllables (Brungart, 2001; Darwin et al., 2003). The bigger difference on either VTL or GPR between the two talkers speaking the concurrent sound led to higher accuracies of recognitions. Additional enhancements were found for both increasing of VTL and GPR contrasts. The interaction between two factors were also evaluated to have a trading relationship. When no other cues available, a two-semitone difference of GPR led to an equal improvement with a 20% difference of VTL (Vestergaard et al., 2009).

2.5 Isolated Syllable Recognition in Mandarin

Though the concurrent syllable recognition consisting of consonants and vowels haven't been evaluated in Mandarin, the relative importance of these two elements was discussed for listeners identifying isolated Mandarin syllables (Chen et al., 2015).

Role of consonants and vowels in speech perception were widely examined in English (Cole et al., 1996; Fogerty & Humes, 2012; Kewley-Port et al., 2007). Vowels were found to have double times of positive effect on recognizing sentences over consonants, according to the method that presented different level of noise-polluted consonants or vowels. Similar method was used to study the relative contributions of consonants and vowels in Mandarin (Chen et al., 2015). Individual words were presented as stimuli to eliminate the cue of sentence context. The performance of vowel recognition was evaluated by varying the level of noise

pollution on consonants and vowels. The results showed that recognition of words was significantly improved by reducing the noise-pollution level of vowels, while such effects were not found by increasing the intelligibility of consonants. Additionally, though all significantly correlated with word recognition accuracy, vowels and tones had higher correlation coefficient (0.72 and 0.78 respectively) than consonants (0.23). Their works showed relatively larger contributions of vowels in recognizing individual words than consonants. However, such different roles in concurrent syllable recognition was still unclear.

2.6 Speech Separation of Deep-learning Models

In recent years, large amount of studies aimed at developing technologies used for speech separation in time domain based on deep learning. (Luo & Mesgarani, 2018, 2019; Stoller et al., 2018; Yu et al., 2017). The iteration of the algorithms contributed to higher separation performance of models indicated by signal-to distortion ratio (SDR) and subjective measures (mean opinion score, MOS).

At first, works of speech recognition in single-talker cases triggered the interests on solving multi-talker speech separation, which can be reflected by a classical example called cocktail-party problem. Previous works made attempts to tackle it, while the performance on separation was limited by the speaker features to a great extent. With the progress of deep learning technology, a permutation invariant training method was proposed to eliminate the dependence on speaker so that the model could be implemented for a wider use (Yu et al., 2017). Algorithms for speech separation was achieved by accessing to spectrogram of the sound mixtures. However, limitations existed in these methods due to the ignorance on phase information. A time-domain processing system was then evaluated to largely rely on temporal information (Stoller et al., 2018).

In terms of the application, speech separation systems faced challenges especially on processing in real time and with short latencies. A recent work developed a framework using encoder and decoder to replace existing time-consuming decomposition method, which lead to an improvement on processing speed (Luo & Mesgarani, 2018). Based on their study, the latest model, Conv-TasNet, proposed by Luo & Mesgarani, 2019 could be used to separate speech from two or three talkers speaking simultaneously. A MOS value of 4.03 was

acquired by the output speech, which was very close to clean speech conditions where MOS was 4.23. The Conv-TasNat had greater advantages with smaller size of model and shorter latency in processing. Models were iterated through the optimization of algorithms and structures from work to work.

2.7 Research Gaps

Based on the literature review, it was found that the concurrent speech recognition in Mandarin received less attention than English. Studies showed a relatively weaker effect of consonants than vowels in words recognition, while the importance of consonants have not been examined when competing speech exists. As a result, the research gaps were summarized as follows:

Gap1: The concurrent-vowel recognition studies are much less on mandarin than English.

Gap2: There're studies exploring factors that affect intelligibility of mandarin syllables (consonants and vowels) (Chen et al., 2015), but no study has evaluated the recognition performance of such mandarin syllables concurrently presented.

Gap3: There're increasing studies developing algorithms and deep-learning models dealing with multi-talker speech perception (Luo & Mesgarani, 2019; Yu et al., 2017). No study has compared the human performance on concurrent syllables recognition with such deep-learning model.

2.8 Research Questions

According to the research gaps above, this study aimed at solving the following questions:

Question 1: What is the recognition performance of two concurrent mandarin syllables each comprised of an initial consonant followed by a vowel in different combinations?

Question 2: When listen to the concurrent syllables in Q1, how does the spectral envelope contrasts of two syllables influence the correct recognition rate of each of them?

Question 3: Do the spectral envelope contrasts of consonants and vowels influence the concurrent syllables recognition equally? If no, which contributes more to the correct rate of syllable recognition?

Question 4: How to use a descriptive model to predict the recognition performance of syllables concurrently presented (described in Q1) based on their consonants, vowels and tones recognition rates?

Question 5: Compared to human, what is the recognition performance of deep-learning speech separation model under the same conditions as human listeners?

CHAPTER 3 EXPERIMENT ONE: EFFECTS OF SPECTRAL CUE ON MANDARIN CONCURRENT-VOWEL RECOGNITION

Summary

The first experiment was designed to explore the effect of different characteristics originating from concurrently presented vowels on recognition performance. The setting of the experiment was consistent with previously reviewed study that six basic vowels with four tones in mandarin were used. Findings included the spectral contrast of the vowels in one pair had valuable benefits on the recognition. Different from past studies, a logarithmic relationship was found between the increase of spectral contrasts and improvement of recognition score, which could be explained by the nonlinearity of human perception.

3.1 Introduction

In order to understand human's ability in speech separation Concurrent-vowel recognition is a simplest case in simultaneously presented speech recognition since vowel is the nucleus of a syllable, which formed sentences. As a start of the study, the experiment aimed to explore how the intrinsic property of vowels would affect the detection of it when another vowel presented at the same time. The understanding of concurrent vowel recognition could form the basis of further syllable recognition.

3.2 Hypothesis

- 3.2.1 It was hypothesized that the spectral envelope contrast could facilitate the performance of concurrent syllable (tonal vowel) recognition. Increasing spectral contrasts would improve the percent correct of the syllable (tonal vowel) recognition linearly.
- 3.2.2 The accuracy of correctly recognizing syllable would be lower than the tone and vowel recognition.

3.3 Method

- 3.3.1 Variables

Dependent variables: The experiment was designed to evaluate the recognition performance of concurrent tonal-vowel pairs. As a result, the correct recognition rates including syllable, vowel and tone were three dependent variables.

Independent variables: The independent variable originated from differences of two presented tonal vowels in each pair. Specifically, vowel category and tone category of each tonal vowel. In addition, the spectral contrast between two vowels in one concurrent pair (The detailed introduction and calculation of spectral contrast would be introduced in section 3.2) was counted as another independent variable.

Variables were summarized in table 3.1

| Independent Variable | |
|-----------------------------------|-------------|
| Name | Type |
| Vowel category | Categorical |
| Tone category | Categorical |
| Spectral contrast | Continuous |
| Dependent Variable | |
| Name | Type |
| Syllable correct recognition rate | Continuous |
| Vowel correct recognition rate | Continuous |
| Tone correct recognition rate | Continuous |

Table 3.1 Independent and dependent variables in experiment one

3.3.2 Stimuli

In this experiment, the sound stimuli was generated from text-to-sound (TTS) tool (Neospeech) for better control of sound characteristics, which could also make the production of sound replicable. Six basic single vowels (“a”, “o”, “e”, “i”, “u”, “ü” in Pinyin with international phonetic alphabet (IPA) symbols [a], [o], [ɤ], [i], [u], [y] respectively) in Mandarin with four tones were produced by the TTS tool mimicking adult female speaker

with 44.1-kHz sampling rate. These four tones included T1 (high-level tone), T2 (high-rising tone), T3 (falling-rising tone) and T4 (falling tone), which differed from their fundamental frequency (F0) contours.

After acquiring generated 24 tonal vowels (6 vowels with 4 tones each) from TTS tool, three adjustments were made by Praat (Boersma, 2006) to limit the effect of other cues. First, the mean F0 of each tonal vowel was adjusted to be equal at 210Hz to eliminate the cue of mean F0 variation. The reasons for 210Hz was chosen included: (1) 210Hz is the typical average F0 for adult female talking; (2) Female speech has relatively higher intelligibility compared to male speaker (Darwin et al., 2003). Second, the durations of all vowels were adjusted to be equal at 450ms to minimize the effect of misalignment. All original stimuli was generated by TTS tool to have various durations. In natural, tone 3 has the longest duration, while tone 1,2,4 last relatively short. 450ms was a eclectic choice to remain enough features of tone 3 without much extension on other three tones. Third, the sound level of each single vowel was modified to be equal at 65dB so that cue of intensity difference was removed. After all processing, 5 listeners and experimenter listened to the single stimuli to ensure that the intelligibility of each single tonal vowel was qualified.

In the following step, the stimuli, concurrent-vowel pairs, were generated by mixing any two tonal vowels. As a result, 276 (C_{24}^2) pairs were acquired as the experiment stimuli. All the produced sound pairs were then adjusted to have the equal intensity of 65dB.

In Figure 3.1, the original and adjusted waveforms of tonal vowel “a1” were showed. The waveform of an example concurrent-syllable (tonal vowel) pair (“a1”+ “e2”) were showed in Figure 3.2.

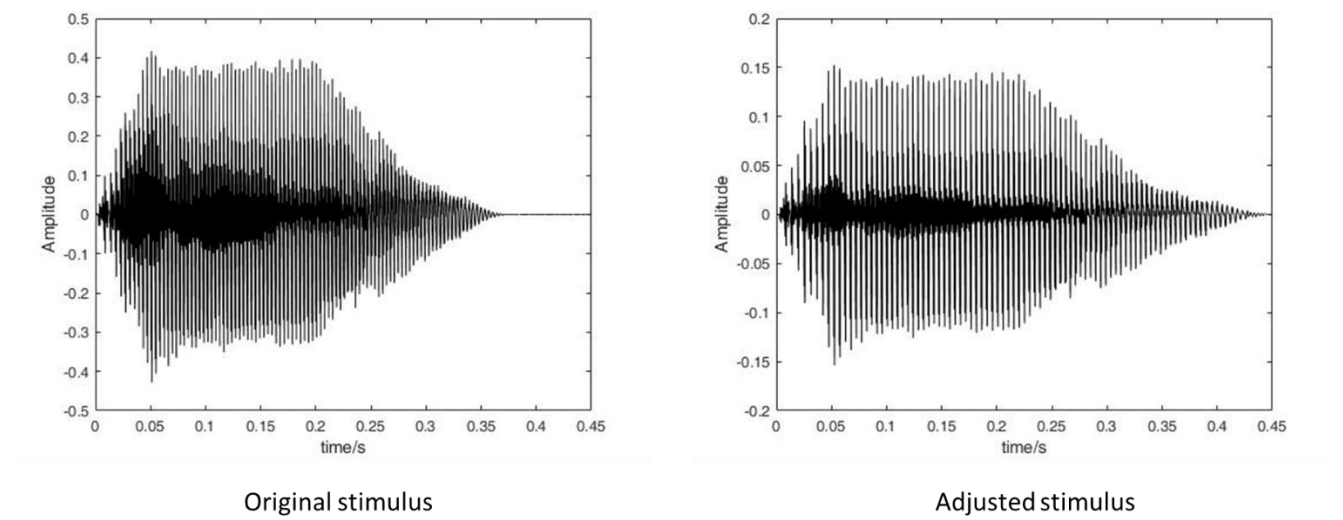


Figure 3.1 Waveforms of original stimulus generated from TTS tool (left) and adjusted stimulus used for mixing (right)

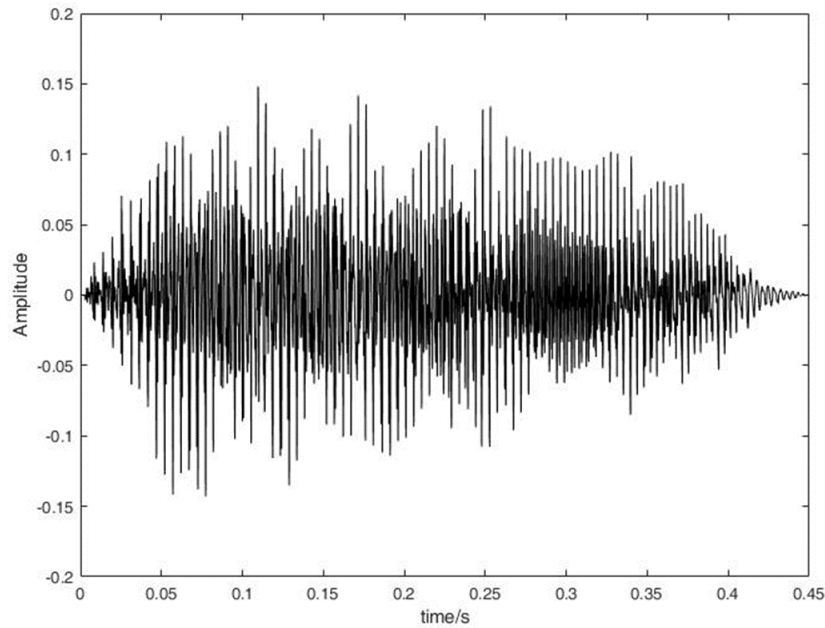


Figure 3.2 Waveform of concurrent-vowel pair example: “a1”+ “e2”

The spectral contrast of a sound pair was calculated as the mean square error (MSE) of the spectral envelope of the two tonal vowels forming the pair. For each adjusted vowel, the spectral envelope was extracted by using Cepstrum method (Song, 2013). Figure 3.3 showed

the spectral density of tonal vowel “a1” and corresponding spectral envelope. In figure 3.4, the spectral envelopes of two tonal vowels “a1” and “u2” were plotted in different colors. The spectral contrasts calculated as the mean square error of two envelopes indicated the spectral power difference of two vowels, which was expected as a cue for listeners to do the separation and recognition.

In past study (Fu, Wu, et al., 2019), spectral envelope of each syllable was extracted with ignorance of spectral information above 4kHz, as vowel energy concentrated in the first 3 formants. To be consistent with this, spectral information below 4kHz was used in this experiment. However, a comparison was conducted between this and the condition when all spectral information was reserved for spectral contrasts calculation. Table 3.2 showed the ranking of vowel pairs according to the normalized spectral contrasts in two conditions. It could be observed that besides some swaps between two consecutive pairs, no huge difference existed in the ranking.

| Frequency range below 4kHz | All frequency range |
|----------------------------|---------------------|
| au | au |
| ai | ai |
| uü | uü |
| aü | eu |
| eu | aü |
| ao | ao |
| oi | oi |
| oü | oü |
| eü | eü |
| iu | ei |
| ei | iu |
| ou | ou |
| ae | iü |
| iü | ae |
| oe | oe |
| uu | uu |
| ii | ii |
| üü | oo |
| oo | üü |
| aa | ee |
| ee | aa |

Table 3.2 Ranking of vowel pairs in terms of spectral contrasts value

The spectral contrasts were grouped according to two vowel categories comprising the concurrent pair, while tone information was averaged in each group due to the insignificant effect on spectral contrasts across tone categories. For example, vowel pair “a-u” had its specific spectral contrasts value regardless of tones.

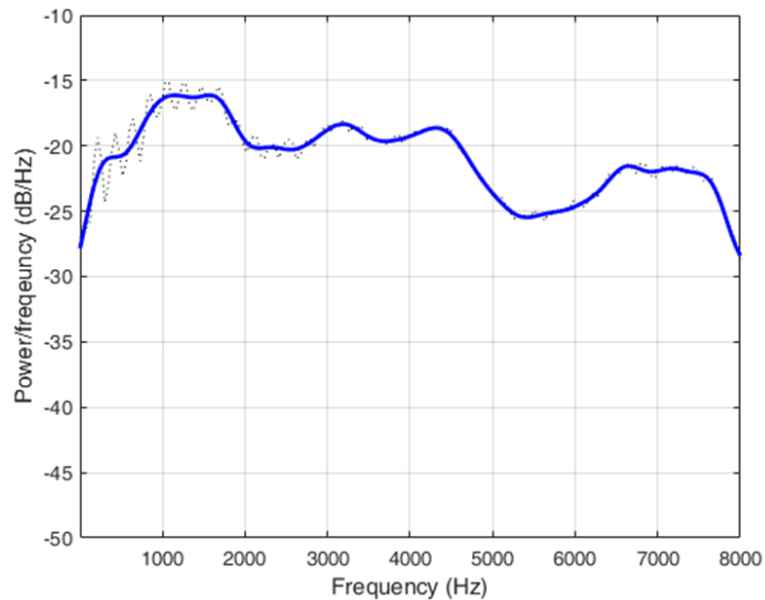


Figure 3.3 Spectral density and spectral envelope of vowel “a1”

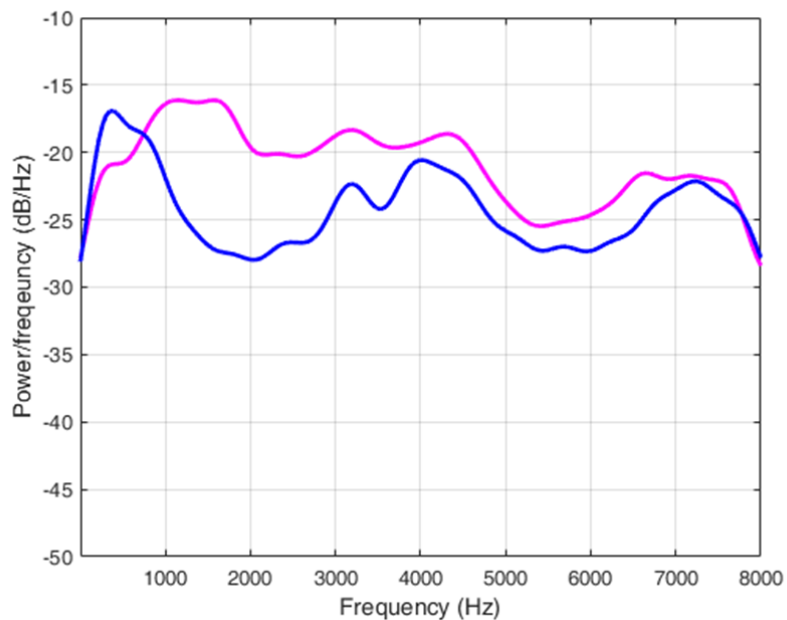


Figure 3.4 Spectral envelopes of vowels “a1” and “u2”

3.3.3 Sound calibration

In order to control the sound level reached at the subject's ear, an electric dummy head (NEUMANN) was used to mimic the real listener and receive the calibration tests. The Sennheiser HD558 headphone that would be used in the formal test was put on the mimic head, which was connected to a ZOOM audio converter that converted the analog signal recorded from the dummy head to digital signal for analysis. Stimuli used in experiment was presented to the head. Then sound files were recorded and saved on computer for analyzing. The input sound level would be adjusted to make sure the recorded wave had the intensity of 65dB. The setting of calibration is showed as Figure3.5.

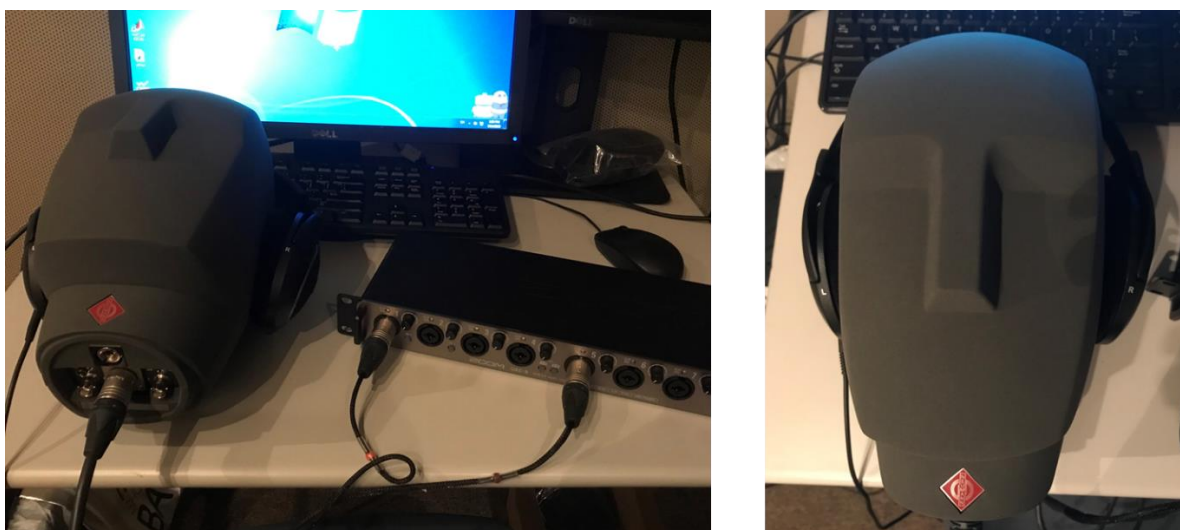


Figure 3.5 Settings of calibration

3.3.4 Tasks and procedure

The experiment was conducted in a sound-proof booth. Subjects were required to sit in front of a table facing a computer screen and listen to the stimuli binaurally through the headphone. In each test, a pair of concurrent tonal vowels was played and the subjects were asked to recognize both vowel and tone by clicking the corresponding buttons on a MATLAB graphical user interface (GUI). Figure 3.6 is the screenshot of the GUI. In each test, the sound would only be played after clicking on the “play” button. This design attempted to avoid the miss of hearing when subjects were distracted.

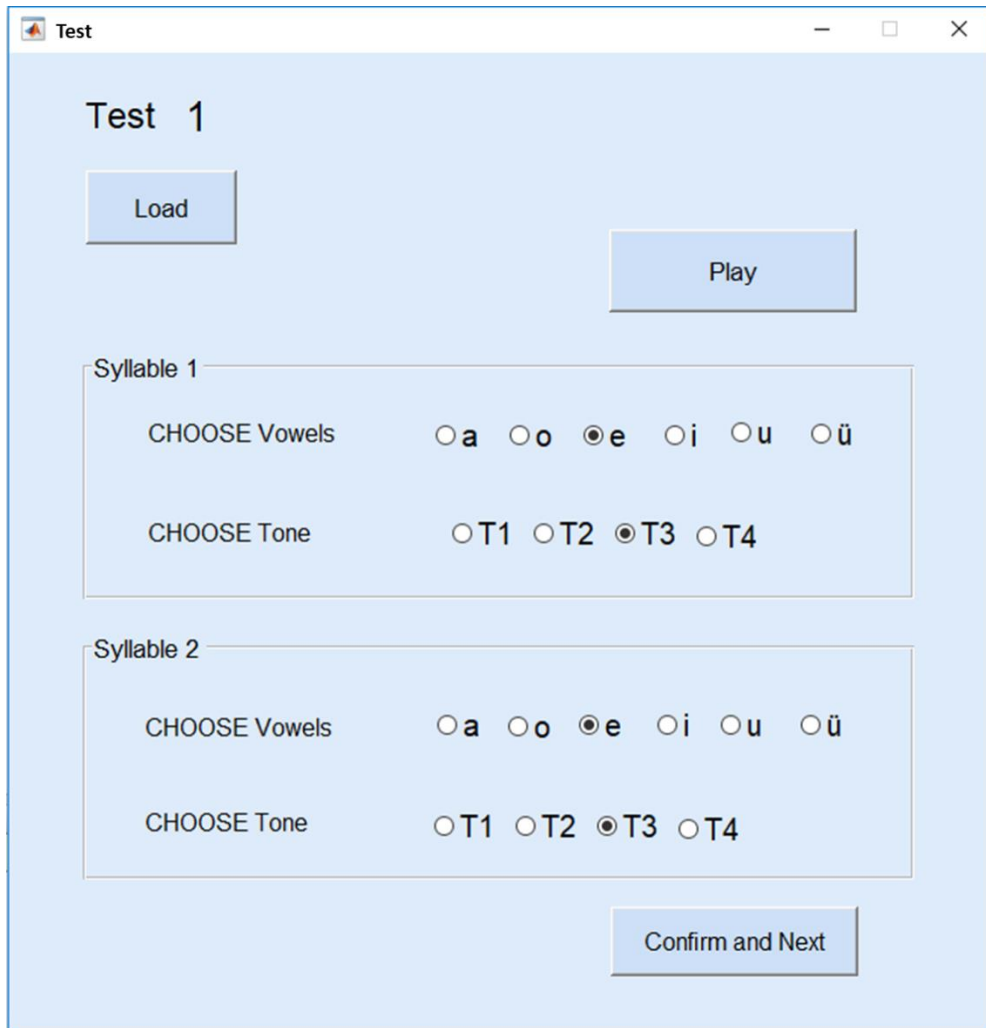


Figure 3.6 Screenshot of the GUI used in experiment

All 276 concurrent-vowel pairs were randomly assigned to seven sessions, which each of the first six sessions had 40 sound pairs and the last session had 36 sound pairs. In one session, the vowel pairs were presented one by one in random order, where each pair was labelled as one test. Subjects were allowed to listen to the stimuli only once in one test. There were 5-10 minutes break between each two sessions. The whole experiment took around 60-80 minutes to be finished.

3.3.5 Subjects

15 normal-hearing and native-Mandarin subjects (9 males and 6 females, age range 23-31, mean 25.6) participated in the experiment. All are postgraduates studying at The University of Science and Technology with no history of hearing impairment. Each of them has passed

the hearing tests to have the pure-tone detection threshold of below 20dB hearing level (HL) at octave frequency intervals from 250 to 8000Hz. For a fear of subject's dialect influencing the recognition performance, accent background of each subject was explored. Among 15 subjects, five were from southern China and others were northern residents. None of them reported difficulties in differentiating Mandarin vowels. All 15 subjects passed the training session test to ensure that all single tonal vowels can be correctly recognized.

The hearing threshold tests were done by using an audiometer (Madsen Itera-II, otometrics) showed in figure 3.7. Three-time inverses were acquired to determine the hearing threshold of each tester on each frequency.

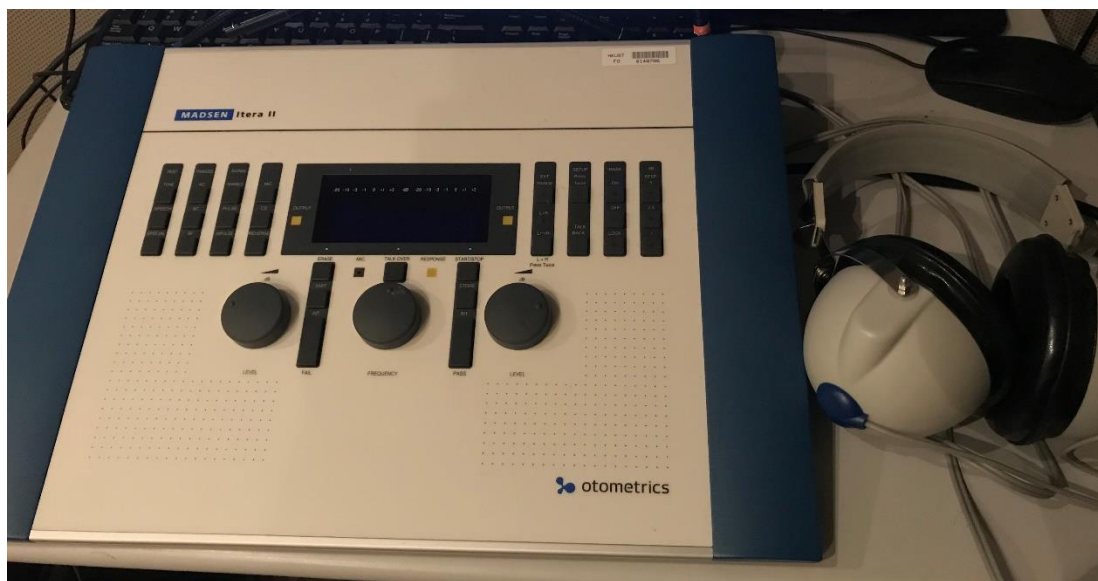


Figure 3.7 Picture of audiometer used for hearing tests

3.4 Results and analysis

In the experiment, the number of correct recognition was recorded for each subjects. The correct rate was first calculated as the number of correct recognition/total number of stimuli. The percentage was then transformed to rationalized arcsine units (RAU) to make the data suitable for statistical analysis. In this study, performance of subjects was recorded as percentages, which showed lack of statistical suitability. It was reflected in three parts: (1) Not followed normal distribution; (2) variances and means were correlated; (3) Data was not linearly related to test variability. It was shown that arcsine transform was most proper to deal with these problems. Based on this, a past study (Studebaker, 1985) further reformed the calculation to have comparable interpretation as percentage data. The RAU results presented

a nearly equal rate of correct response in recognition tasks, while also suitable for statistical analysis. As a result, RAU was chosen as the most appropriate transformation in our study.

The transformation followed two steps according to past study (Studebaker, 1985). Firstly, the arcsine units were calculated using equation (1), a further step used equation (2) to get the rationalized arcsine units. N is the number of tests presented and s is the number of correct recognition, two equations were showed below:

$$AU = \arcsin \sqrt{\frac{s}{N+1}} + \arcsin \sqrt{\frac{s+1}{N+1}} \quad (1)$$

$$RAU = \left(\frac{146}{\pi} \right) \times AU - 23 \quad (2)$$

3.4.1 Effect of vowel spectral contrast on correct recognition rate

In order to explore how the spectral contrast influenced the recognition performance, the correct rate (RAU) of syllable recognition were plotted as a function of spectral envelope contrast in figure 3.8.

A positive correlation could be observed. Pearson correlation test showed that correct rate and spectral contrast are significantly correlated ($p < 0.001$), correlation coefficient $r = 0.731$. When two vowels in the concurrent pair were from the same category (e.g. “i1i2”, “o3o4”), the correct rate of the recognition was low (e.g. 30 for aa and 21 for ii); For larger spectral contrast (larger than 0.2) sound pairs, the correct rates were all above 60 and reached 80 for the “au” vowel pairs having the largest contrast.

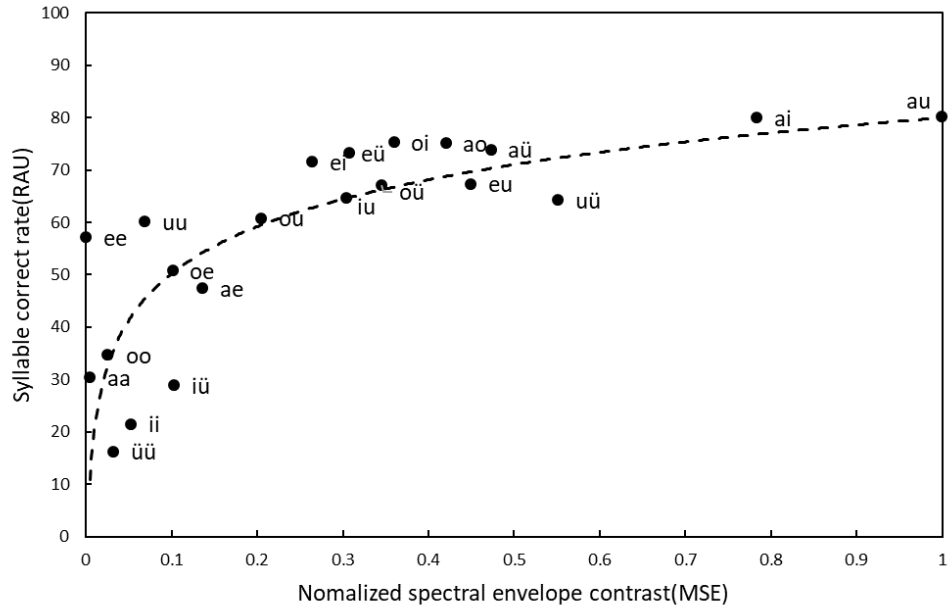


Figure 3.8 Syllable correct recognition rate as a function of normalized spectral envelope contrast

The results indicated listeners could utilize the spectral contrast of vowels to have a better recognition. However, the trend of the improvement was more of logarithmic scale (r -squared equaled to 0.71) instead of linear scale (r -squared equaled to 0.58) in the hypothesis.

Subsequently, the syllable correct rates were transformed in log scale and plotted as a function of normalized spectral envelope contrast in figure 3.9. A linear correlation could be observed. A Pearson correlation test revealed that the log-transformed correct rate was significantly correlated with the spectral envelope contrast with correlation coefficient 0.843. In addition, a good linear fitting was reflected by the residual plot showing the random pattern.

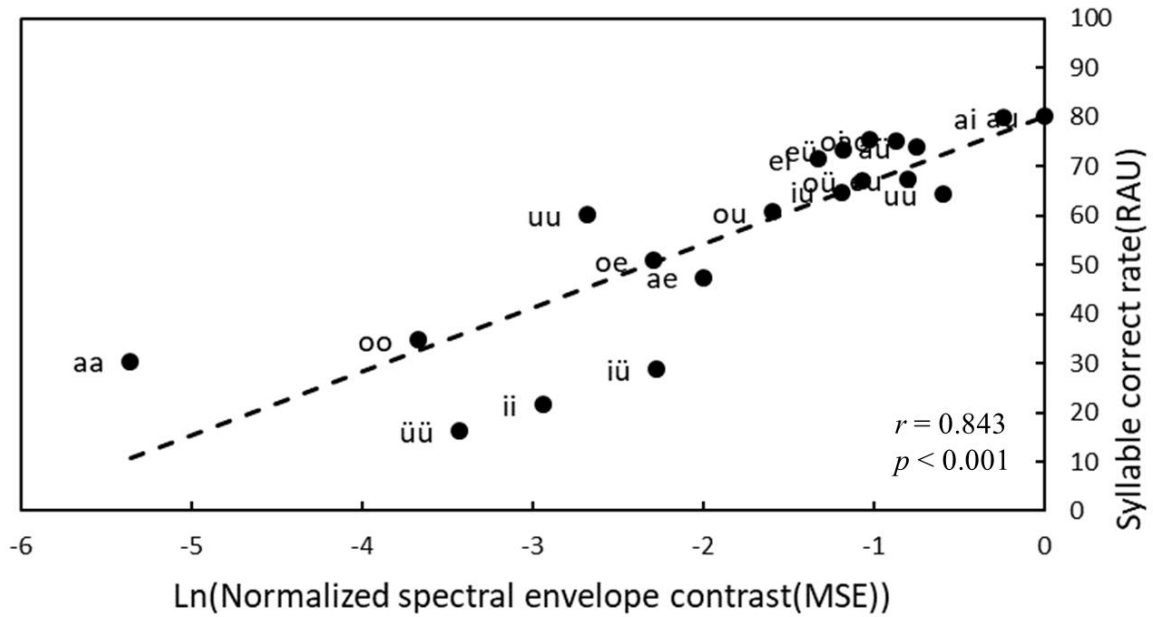


Figure 3.9 Log-transformed syllable correct recognition rate as a function of normalized spectral envelope contrast

3.4.2 Effect of vowel category on correct recognition rate

To explore how different vowel categories contribute to the recognition task, vowel pairs are classified into two groups: same vowels (SV) and different vowels (DV) for 6 basic vowels using in this experiment. For example, “a-o”, “i-u” were from the DV group, “a-a”, “o-o” belonged to the SV group.

A two-way repeated measure ANOVA was then conducted to show the main effect of vowel category was significant [$F(5,154)=9.45, p<0.001$]. And the main effect of vowel-pair group (had SV and DV as two levels) was significant [$F(1,154)=191.53, P<0.001$]. The interaction effect was also significant [$F(5,168)=9.84, p<0.001$]. A post-hoc analysis revealed that for vowel category “a”, “o”, “i”, “ü”, correct recognition rate of their DV group was significantly higher than the SV group ($p<0.001$), but such results were not significant for vowel “e” ($p=0.451$) and “u” ($p=0.984$).

Figure 3.10 showed the averaged correct recognition rate for different vowel categories from SV and DV groups. In general, accuracies were higher for vowels presented with another different one compared to vowels presented with same categories. Among six vowel

categories, vowel “i” and “ü” received the lowest accuracy possibly due to their similar sound like. Vowel “u” had the relatively high correct rate even in the SV group. A possible reason could be the distinct articulation manner that made it less confused when mixed with other vowels.

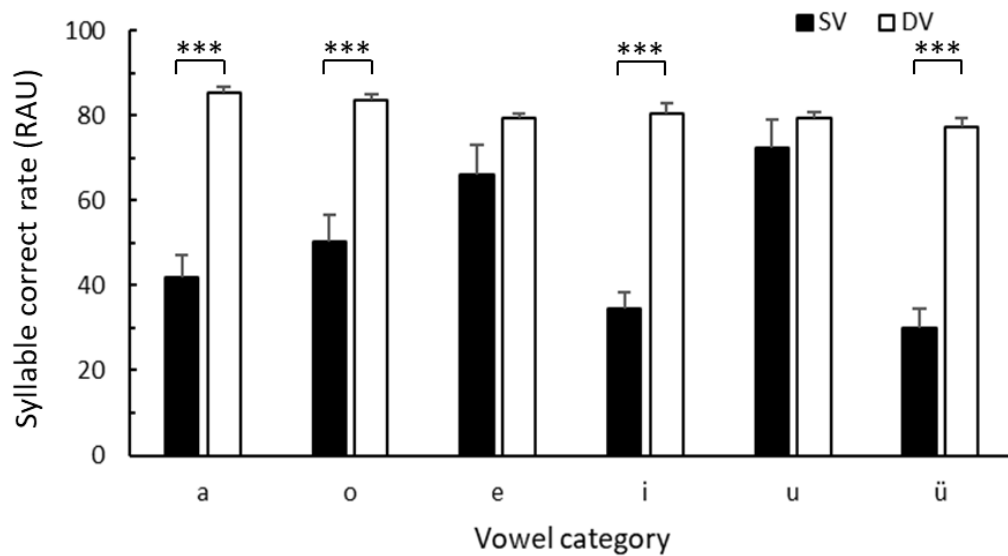


Figure 3.10 Average syllable recognition accuracy of six vowel categories with two combination groups (same vowel and different vowel)

3.4.3 Comparison of different measurement of correct recognition rate

The performance of recognition was further evaluated from the element consisting of the syllable. Besides syllable, recognition accuracies of tone and vowel were also considered in this section. Figure 3.11 showed the correct recognition rate for each measurements. One-way repeated-measure ANOVA revealed that the measurements had significant effect [$F(2,42)=16.72, p < 0.001$]. Post-hoc analysis indicated that accuracy was significantly lower for syllable than vowel and tone recognition ($p < 0.001$) but no significant difference was found between tone and vowel measurements ($p=0.781$). This finding was consistent with past studies.

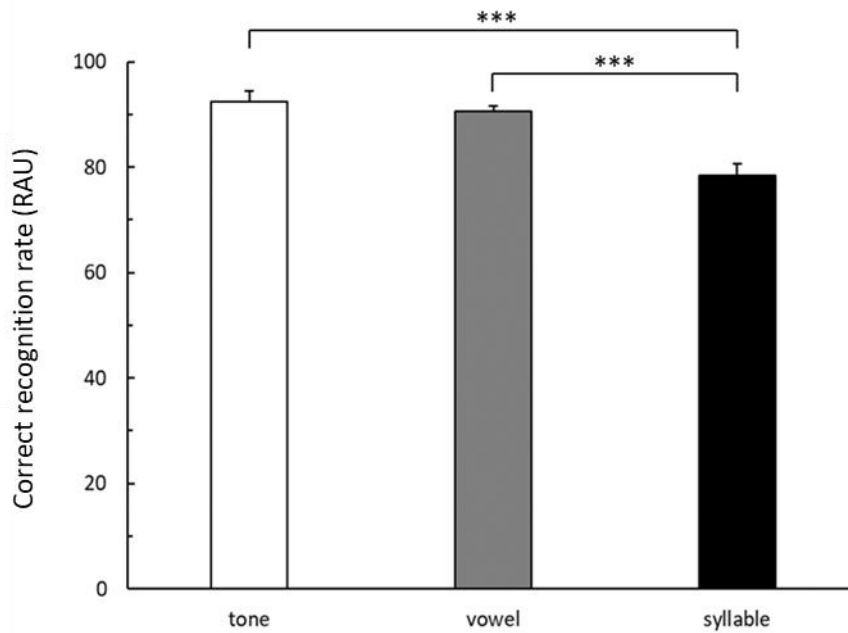


Figure 3.11 Average correct recognition rate of tone, vowel and syllable

3.4.4 Effect of tone category on tone perception

Performance of tone recognition was evaluated individually in this section. In order to show error pattern in identifying tones, figure 3.12 plotted the confusion matrix of tone recognition, where the horizontal axis showed the tone presented in the stimulus and vertical column represented for subject's response of corresponding stimulus. In the matrix, the value given in the grid "T1-T1" grid provide the information of accuracy tone 1 recognition, while the value in grid "T1-T2" was the average percentage of response that tone 1 was wrongly recognized as tone 2 by listeners. Remaining grids could be understood in the same way. From the background color of the grids, it could be seen that a darker blue stood for lower error rate while a brighter yellow represented for higher correct rate. Tone 1 and Tone 4 showed relatively high accuracy (95.8% and 96% respectively). Contrast, the most difficult tone for recognition was tone 3 with a low accuracy of 75.2% (84.3%) among four tones.

A one-way repeated-measures ANOVA showed that the main effect of tone category was significant [$F(3,42)=27.71$, $p<0.001$]. Post-hoc multiple comparisons revealed that performance of recognizing Tone 3 (T3) was significantly worse than other 3 tones ($p < 0.001$), and there were no significant difference between any other tones ($p > 0.214$).

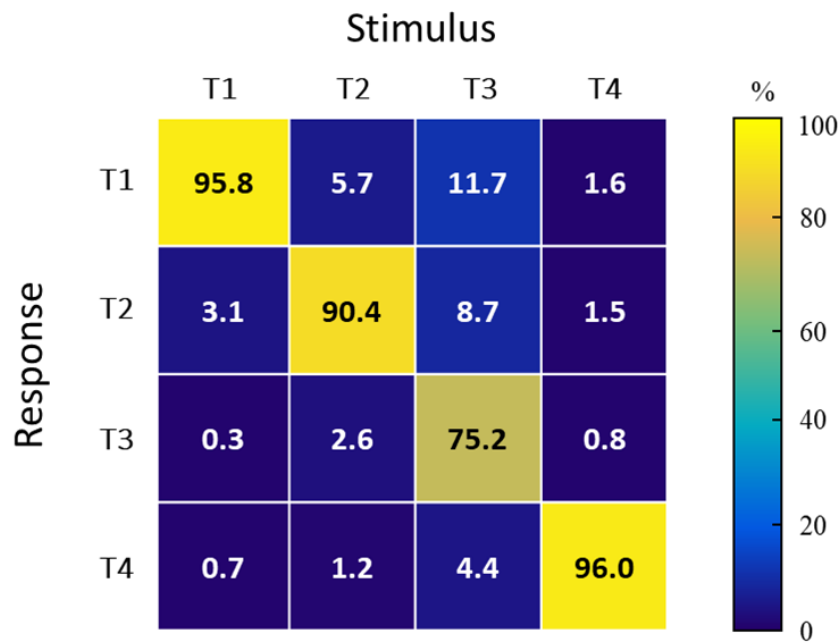


Figure 3.12 Tone confusion matrix in concurrent Mandarin vowel recognition

3.5 Discussions and conclusions

The experiment one was designed as a verification of past studies (Fu, Wu, et al., 2019). Also, it provided preliminary conclusions that can be used for extending the stimuli set to a wider range so as to cover more common cases.

In general, findings in this experiment was consistent with Fu's works except for the degree of utilization of spectral contrasts information in large contrasts range. The effect of spectral contrasts in our case was of logarithmic scale instead of linear scale. In terms of the difference on the trend, we noticed that in Fu's study, the stimuli was recorded by female speaker, while vowels were generated by TTS tool in our experiment. More speech features were controlled in our case as the synthesizer followed programmable procedure in producing stimuli. Human speaker provided more natural cues on the vowel utterance. These may serve as inferential clue for listeners differentiating concurrent vowels, leading to higher accuracy particularly for those vowel pairs having small spectral contrasts.

In terms of the logarithmic relationship in our findings, it could be understood as the using of spectral contrasts information was especially effective when two vowels in a pair differed less spectrally. Possible explanation could be the nonlinearity of human auditory system. For example, in hearing, a linear increase in sound intensity would not lead to a linear perception of loudness. Our auditory system actively adjusts the gain of the signal received according to the need. However, due to a more complicated cases rather than loudness perception in our experiment, Steven's power law could be another possible supports. When the spectral contrast of vowels in a pair was low, a small increasement of contrasts could provide useful clues for human listeners to do the differentiation. The effect of spectral power difference might surpass other cues for such vowel pairs. However, when the spectral contrasts increased to a certain degree, such as "a" and "u", the energy difference on the spectrum had weak effect on improving the recognition accuracy. It could be due to the energy suppression of hearing systems, or the spectral contrasts were so large that enough useful information had been acquired.

In the tone recognition, the tone 3 had the significantly lowest accuracy. As the tone was categorized by the pitch change temporally, tone 3 was the only tone had two consecutive changes of directions in pitch contour, which made it harder to be perceived. In addition, tone 3 had the longest duration naturally, an equalized duration of the stimuli weaken the features of tone 3 for recognition to some extent. This results were consistent with other's works (Fu, Wu, et al., 2019; Fu, Yang, et al., 2019).

Based on findings in the concurrent vowel recognition, an extension to a more general case when initial consonant was involved in the recognition could be evaluated in follow chapters.

CHAPTER 4 EXPERIMENT TWO: EFFECTS OF SPECTRAL CUE ON MANDARIN CONCURRENT-SYLLABLE RECOGNITION: A CRITICAL CASE WITH SELECTED CONSONANTS

Summary

This experiment was designed on the basis of experiment one to evaluate the effect of spectral contrasts on concurrent syllable recognition where an initial consonant with a followed vowel comprised each syllable. Four consonants and two vowels with four tones selected to generate the stimuli probed the role that cues, including spectral contrasts and syllable intrinsic properties (vowel category, consonant category), played on concurrent syllable recognition.

The findings from this experiment revealed the increasement of spectral contrast produced positive influence on syllable recognition. Whilst recognition performance could be enhanced by enlarging spectral differences between two single syllables in a pair, degree of improvement was limited much when only consonant differences available.

Different measurements of concurrent syllable recognition investigated the correct rate for every elements of syllable. It was found that consonant recognition score had the lowest accuracy, which was large-partly responsible for the confusions in syllable recognition.

Due to the restricted number of consonants and vowels used in this experiment, relative contributions of each element in concurrent syllable recognition would not be discussed in this chapter.

4.1 Introduction

Mandarin syllables typically consist of an initial consonant and a final vowel. In Experiment 1, six basic vowels had been mixed to form the concurrent sound pair and spectral contrasts between vowels were proved to be helpful for recognition. A review of literature studying the relative contributions of consonants and vowels on syllable intelligibility suggested that consonants had weak effect on identifying single syllable. In our case, syllables were not presented alone but with another competing syllable played simultaneously. The role of

consonant for distinguishing and recognizing these syllables remained unknown. As a result, the performance of concurrent syllable recognition when an initial consonant was available aroused the interests. Experiment 2 selected four consonants used in high frequency with different classifications to combine with basic vowels to form so called consonant-vowel (CV) syllables. Evaluation of human performance in recognizing concurrent CV syllable could gave promising understanding in a closer way with common cases (daily speech separation and identification).

4.2 Hypothesis

- 4.2.1 Similar as concurrent vowel recognition, spectral contrasts of consonants could also enhance syllable recognition, larger contrasts led to better performance.
- 4.2.2 When both vowel and consonant differences are available, percent correct in CV syllable recognition was high. Performance degraded when only consonant difference available (same vowel) compared to both vowel and consonant difference available.

4.3 Method

4.3.1 Variables

In this experiment, different choices of the elements forming syllables varied among sound mixtures. Thus, categories including vowel category, tone category and consonant category were three of independent variables. In order to evaluate the effect of spectral contrast on the CV syllable recognition, it was used as a continuous variable.

The recognition performance was measured according to syllable recognition, vowel recognition, consonant recognition and tone recognition four aspects.

Variables were summarized in table 4.1.

| Independent Variable | |
|------------------------------------|-------------|
| Name | Type |
| Vowel category | Categorical |
| Consonant category | Categorical |
| Tone category | Categorical |
| Spectral contrast | Continuous |
| Dependent Variable | |
| Name | Type |
| Syllable correct recognition rate | Continuous |
| Vowel correct recognition rate | Continuous |
| Consonant correct recognition rate | Continuous |
| Tone correct recognition rate | Continuous |

Table 4.1 List of variables in experiment two

4.3.2 Stimuli

Syllable presented in this experiment consisted of an initial consonant followed by a basic vowel. There are 21 initial consonants in mandarin, but considering the efficiency and subjects' capability, not all were used to form the stimuli. The selection of consonants were considered from two aspects: (1) Location in the consonant classification table; (2) frequency of use in mandarin. According to the usage frequency (the complete table was in the appendix 4.1), consonant “d” was used most frequently in modern Chinese. By mapping it on the classification table showed in figure 4.1, another three consonants were selected and marked in the figure by using bolded red color. Principle that consonant “b” was selected was that it had same manner but different place of articulation with “d”; Consonant “l” was selected as it had same place but different manner of articulation with “d”; Consonant “sh” was chosen for its different place and manner of articulation with “d”. Meanwhile, these chosen consonants were three most frequently used satisfying corresponding principles and could be combined with basic vowels grammatically.

As the number of stimuli with full combination reached over 4,000 (4 consonants \times 6 vowels \times 4 tones forms 96 single syllables, any two of 96 led to 4,560 mixtures), the basic vowels used in this experiment were only “a” and “u”. Based on the finding of the first experiment, “a” and “u” comprised the concurrent-vowel pair which had the largest spectral contrast and highest recognition score. The effect from vowel was thus weakened to reveal consonants’ role in the recognition task.

| | | Place of articulation | | | | | | |
|------------------------|-----------------------|-----------------------|--------------|----------|--------|---------------|---------|-------|
| Manner of articulation | | Bilabial | Labio-dental | Alveolar | Dental | Retroflex | Palatal | Velar |
| | Plosive Aspirated | p | | t | | | | k |
| | Plosive Unaspirated | b | | d | | | | g |
| | Fricative | | f | | s | sh , r | x | h |
| | Affricate Unaspirated | | | | z | zh | j | |
| | Affricate Aspirated | | | | c | ch | q | |
| | Nasal | m | | n | | | | |
| | Liquid | | | l | | | | |

Figure 4.1 Classification table of mandarin consonants (selections of consonants used for experiment two were bolded with red color)

In total, 32 CV syllables containing four consonants, two vowels and four tones were generated from TTS tool. Subsequent adjustments were conducted to reduce the influence of other cues. The mean fundamental frequency (F0) were equalized at 210Hz to be the typical mean for female speakers. Syllables were modified to have an equal duration of 450ms and same sound intensity of 65dB. All modifications were consistent with experiment one without changing the syllable identity that confirmed by the experimenter.

To construct experiment stimuli, any two different ones from 32 CV syllables were overlapped to form a total of 496 concurrent sound pairs. Two syllables in these pairs could

be differed in vowel only, in tone only, in consonant only or in any two or all of these three dimensions.

4.3.3 Task and Procedure

Experiment was conducted in the same acoustic chamber as experiment one. Subjects were asked to sit in front of computer screen and complete the recognition task by clicking the corresponding button on the GUI generated by MATLAB. A screenshot of the GUI was showed in figure 4.2.

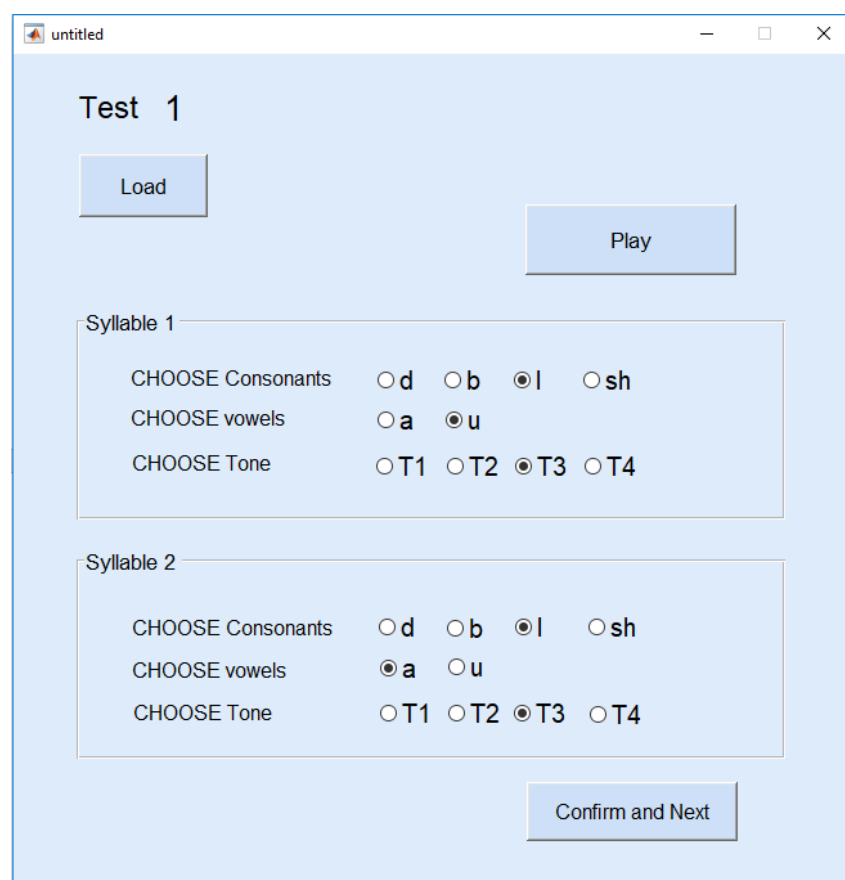


Figure 4.2 Screenshot of GUI used for Experiment 2

Before the formal tests, a training session was set to verify subjects' ability in recognizing single syllable and ensure a correct operation of subjects on the GUI. After getting familiar with single syllables, subjects needed to pass an 8-trial test on recognizing single syllables with correct rate over 95%. Only two out of sixteen subjects failed to pass the test at the first

time. They were asked to listen to single syllable again to successfully pass the second-time test.

Formal tests proceeded after the training session. In each trial, a pair of concurrent syllable was presented upon the “play” button was clicked. Subjects were allowed to listen to the stimuli only once and make the choice without time limit. Nine sessions consisting of 50 trials with another one only contained 49 trials were presented random in order. 5-10 minutes break existed between any two sessions.

4.3.4 Pilot tests

Before the formal experiment starts, a pilot study was conducted to ensure if the design in last experiment (concurrent vowel recognition) could be followed in concurrent syllable recognition. Three subjects participated in the pilot tests. All were native-Mandarin speakers studying in HKUST.

In this pilot experiment, the test-retest effect was studied by repeating the same task 3-4 times. Two subjects completed the same testing in four different days, each two days apart, while one subject repeated the same task in three separate days. In each test, subjects were asked to listen to 50 concurrent-syllable pairs consisting of same vowels but different consonants. Stimuli used in the pilot tests were processed in the same way as formal experiment.

The performance was evaluated according to the accuracy of consonants, tone and syllable recognition respectively. Specifically, the percentage of wrong responses was counted for comparison. Figure 4.3 showed the performance of three subjects in all tests. It could be observed that the error in recognizing syllables and tones decreased as the tests being repeated, which indicated that learning effect existed in identifying such concurrent syllables. Precisely, 23% and 22% improvement in recognition accuracies were obtained after 3-times retests for subject 1 and subject 2 respectively.

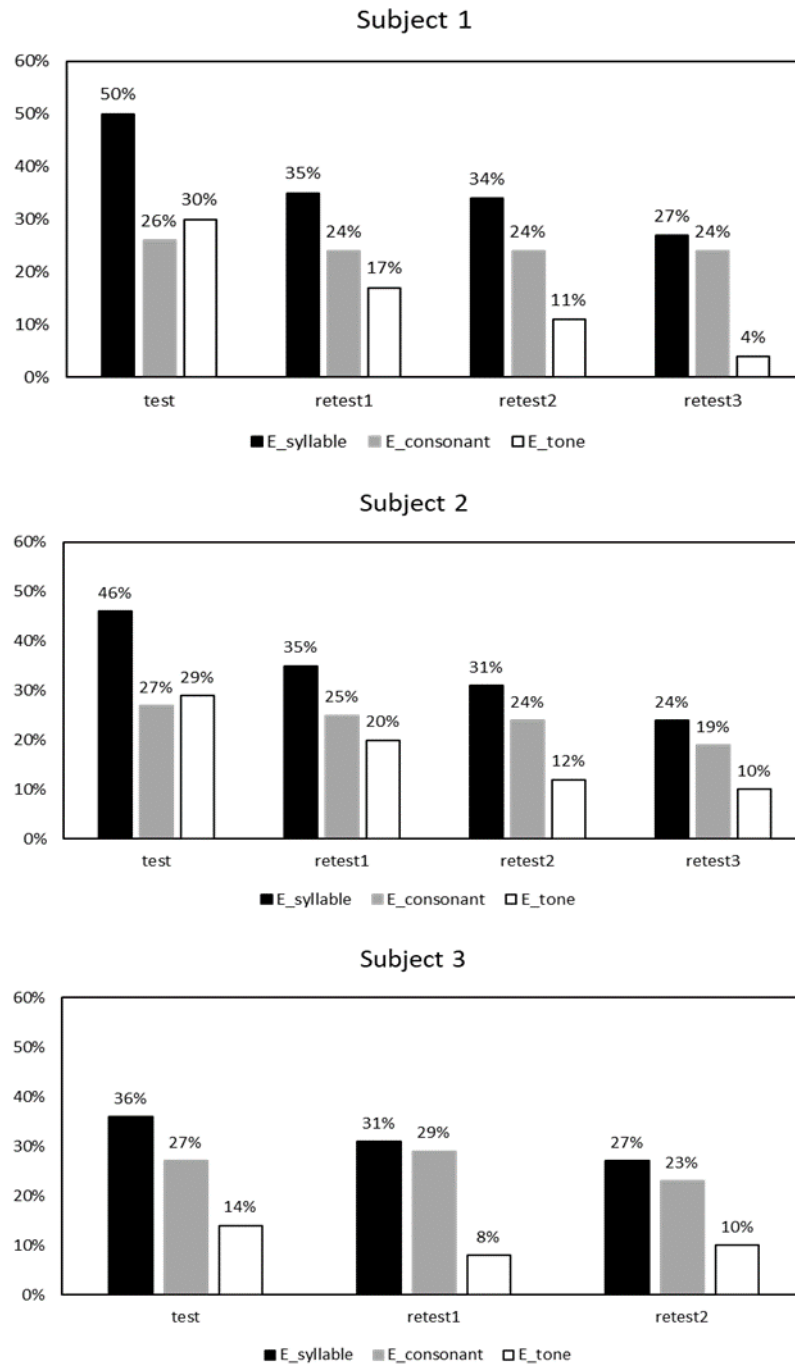


Figure 4.3 Error rate of recognition from three subjects in pilot tests

In addition, in terms of the times allowed for listening in each trial, subjects reported possible strategy for obtaining high accuracy. If one concurrent-syllable pair was allowed to listen for twice, subjects could focus on each individual syllable in each listening. As a result, in formal

tests, the stimuli was allowed to listen for only once to collect listener's instantaneous responses.

4.3.5 Subjects

In this experiment, 16 normal-hearing and native-Mandarin subjects including 8 males and 8 females participated in the experiment. The average age of subjects was 26.7 ranging between 24-31. All of them were studying at The University of Science and Technology and had no history of hearing impairment. Each subjects had the pure-tone detection threshold below 20dB hearing level (HL) at frequency 250, 500, 1000, 2000, 4000 and 8000Hz.

The accent background of each subject was recorded. 6 out of 16 subjects come from southern districts of China, whilst none of them reported difficulties in differentiating consonants or vowels that used in this experiment.

4.4 Results and Analysis

Similar as experiment one, the correct rate was calculated as the rationalized arcsine transformed correct percentage for each measurements (dependent variables).

4.4.1 Effect of spectral contrast on syllable correct recognition rate

The normalized spectral envelope contrasts of each syllable pair were calculated as the same method introduced in experiment one. The spectral contrast value of all stimuli were showed in the appendix 4.2. Fricative property of consonant “sh” produced large spectral difference when it mixed with other syllables consisting of consonants “b”, “d”, “l”. As “b” and “d” share the same place of articulation, spectral contrasts of syllable pair formed by these two consonants with same vowel were relatively low (below 0.05 in normalized value).

A scatter plot depicting the relationship of syllable correct rate and normalized spectral envelope contrast was showed as figure 4.4. It could be observed that there were two clusters showing different information on the plot. In small contrast area (contrast value less than 0.2), no clear trend could be observed. In larger contrast area (contrast value larger than 0.2), a positive relationship was found between contrast value and correct rate.

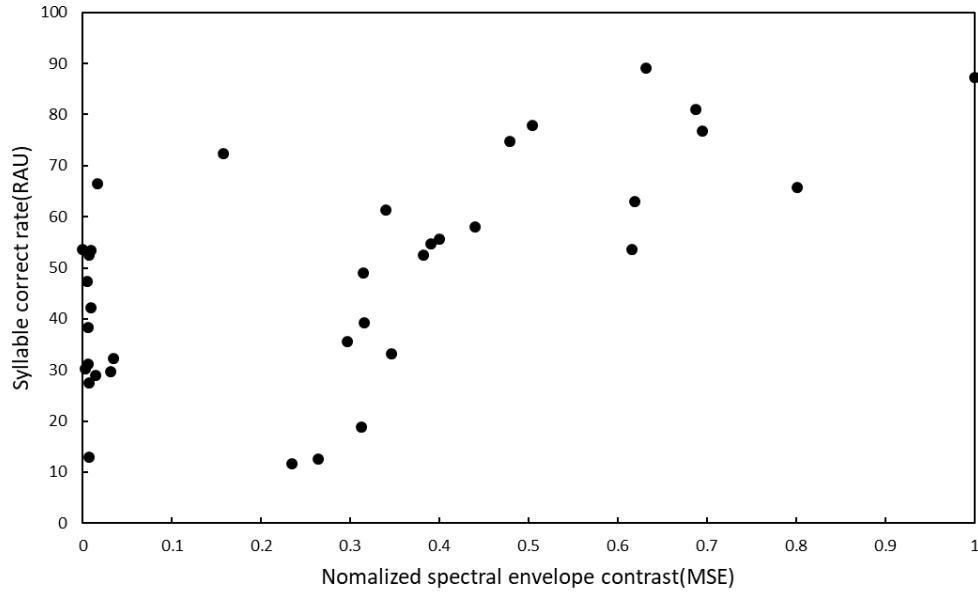


Figure 4.4 Scatter plot of syllable correct rate as a function of normalized spectral envelope contrast

In order to explore the deeper cause of the phenomenon, the concurrent syllable pair were divided into three groups:

- (1) Two syllables had different consonants and vowels or different vowels but same consonants;
- (2) Two syllables had different consonant, same vowels;
- (3) Two syllables had both same consonants and vowels.

As two vowels used in this experiment already had large spectral contrasts, syllables differed in vowels were grouped together despite the diversity of consonants. In group (1), higher spectral contrasts of syllable pairs originated from both vowel contrasts and consonant contrasts available, while in group (2), the spectral contrasts were mainly from consonant differences. When vowels and consonants of two syllables in a pair were the same, the tone played the major role as a temporal cue, which would not be introduced in this section but would be reviewed in the discussion part.

According to the grouping, figure 4.5 plotted out the relation between normalized spectral envelope contrast and syllable correct rate of group (1), all black dots showed an increasing trend of correct rate when spectral contrast became larger. The category of two paired syllables' consonants and vowels was labeled next to the data points. A “shabu” represents for stimuli formed by syllable “sha” and syllable “bu”. When vowel differences available, syllables with top three largest spectral contrast were “sha” mixed with either “bu”, “du” or “lu”, which could be an evidence that fricative articulation could change spectral property of consonants much.

The syllable pair “shashu” was marked with red color as an outlier due to the high correct rate (72.26) but low spectral contrast (0.16). A possible explanation was the limitation of consonants available for selection in the experiment. Due to the distinct sound identity of “sh”, subjects tended to exclude other three consonants in recognition when both syllables in a pair consisting of “sh”. As a result, the correct rate for syllable pair “shashu” was high.

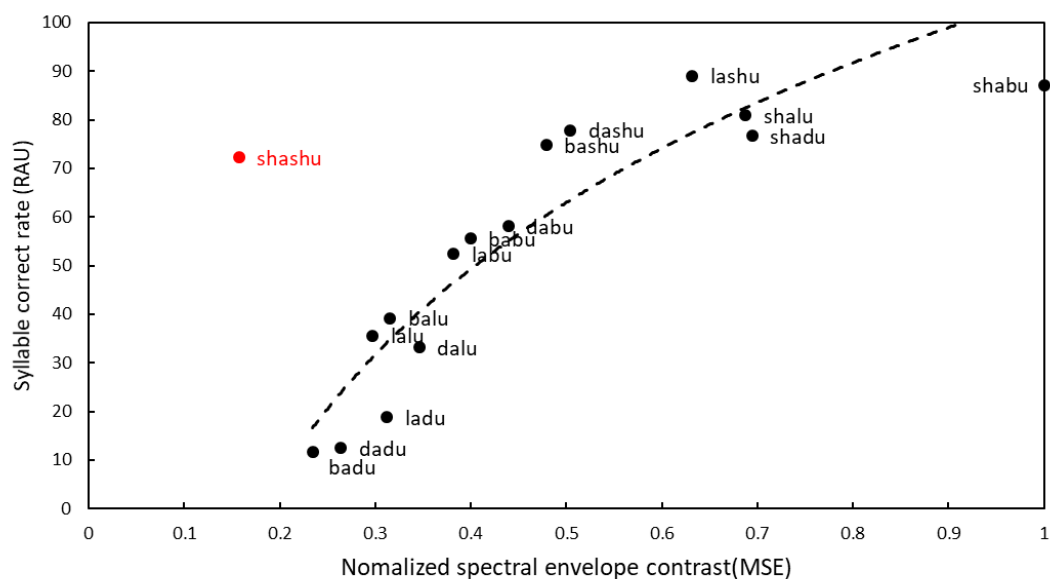


Figure 4.5 Scatter plot of syllable correct rate as a function of spectral contrast with vowel difference available

For the rest of the data points, a Pearson correlation test was conducted to reveal that the syllable correct rate was significantly ($p < 0.001$) correlated with normalized spectral envelope contrast, and the correlation coefficient was 0.846. But curve fitting showed that a

logarithmic regression had higher r-squared (0.856) than linear regression (r-squared equaled to 0.715), which is consistent with experiment one that the effect of spectral contrast was of logarithmic scale rather than linear.

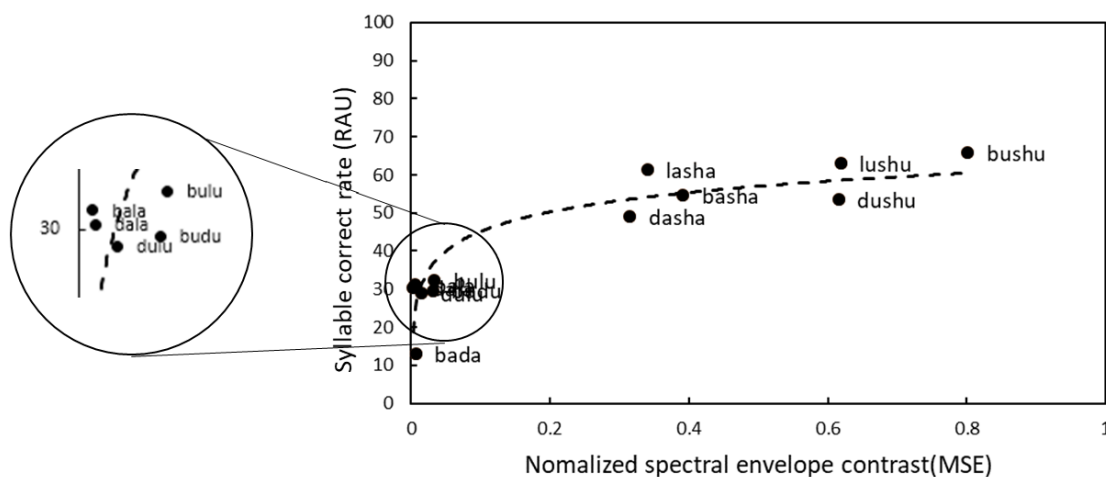


Figure 4.6 Scatter plot of syllable correct rate as a function of spectral contrast with only consonant difference available

4.4.2 Effect of consonant, vowel and tone category on syllable correct recognition rate

The categories of the element forming the syllable were intrinsic identity and were used as independent factors for analysis. A three-way repeated-measures ANOVA reported that the main effect of consonant category was significant [$F(3,474)=54.68, p<0.001$], the main effect of vowel category was significant [$F(1,474)=83.43, P<0.001$] and the main effect of tone

category was significant [$F(3,474)=20.90, P<0.001$]. The interaction effect is also significant between any two factors ([$F(9,474)=8.48, p<0.001$], [$F(3,474)=25.48, p<0.001$], [$F(3,474)=6.70, p<0.001$] for Consonant* tone, Consonant* vowel and Tone*vowel respectively). The post-hoc analysis with Sidak correction showed the correct recognition rate of consonant “sh” was significantly higher than “b”, “d”, “l”, but no significance existed among latter three consonants. The results including complex relation among three factors were attached in the appendix without detailed discussion as it was beyond the scope of the current experiment.

4.4.3 Comparison of syllables, vowels and tones correct recognition rate

We discussed how spectral contrast and category of syllables affect the recognition in earlier sections. Here the performance of recognition was evaluated from detailed measurements including syllable, consonant, vowel and tone four aspects to examine the respective correct rates. Figure 4.7 is the bar chart describing the average correct rate for each measurements. A one way repeated-measures ANOVA was conducted to show the measurements had significant effect [$F(3,45)=505.35, p < 0.001$]. Post-hoc analysis revealed that between any two measurements, the correct rate was significantly different ($p < 0.001$). It should be noticed that the averaged correct rate of vowel was quite high (reached 98.3). The accurate recognition could be explained by the disparate identity of vowel “a” and “u”, and the chance level had already reached 50%.

Among three elements, consonants had the lowest accuracy (81.4), which implied relatively a big part of difficulty in recognizing concurrent syllables came from consonant identification.

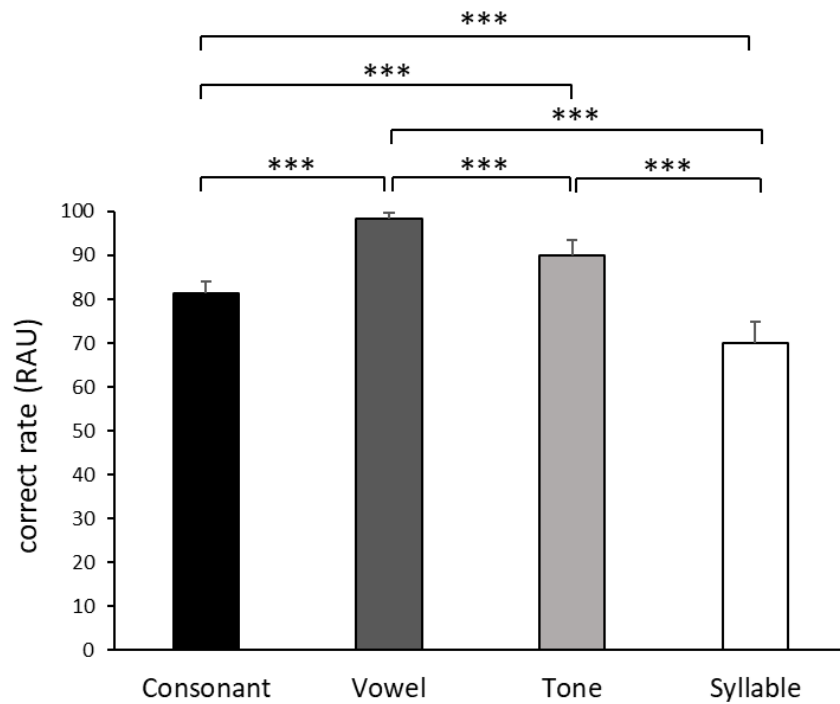


Figure 4.7 Bar chart describing the averaged correct rate of consonant, vowel, tone and syllable with relative significant relations

4.5 Discussions and conclusions

Experiment 2 was designed as a continued study of concurrent-vowel recognition by adding an initial consonant to form the syllable. An important finding in this experiment was the spectral contrast still functioned on the recognition of concurrent CV syllable. The performance affected by spectral contrasts followed a logarithmic scale especially for syllable pairs without vowel difference. A detailed grouping of syllable pairs evaluated two conditions with and without vowel differences. When both consonant difference and vowel difference available, the correct recognition rate could be largely improved with increased spectral consonants, while the room for improvement was relatively limited when only consonant difference available. It could be explained by the relative duration of consonants and vowels of the syllable. The initial consonants carried less energy than followed vowels. As spectral contrasts considered the information of energy difference from the overall spectrum, the energy change from the consonants provided finite clues for listeners to utilize.

In this experiment, consonants comprised the presented syllables were differentiated from either place of articulation or manner of articulation. Both consonant “b” and “d” had plosive aspirated articulation manner. Though pronounced from bilabial and alveolar place respectively, similar identity produced difficulty in recognition when they were presented simultaneously, especially combined with the same vowel. As the only voiced consonant among four, “l” had a relatively clear formants shape on the spectrum, while the same place of articulation as “d” limited the extent of the difference between “l” and “d”. In contrast, the fricative manner of consonant “sh” provided a noise-like feature to be recognized, which made it distinct with other three consonants. Besides the spectral contrasts that was largely different from other consonants, the noise-featured initial sound of a syllable with “sh” could surpass another consonant thus improved the correct recognition rate of concurrent syllables. From the experiment results and subject’s post review, syllable pairs formed by “b” and “d” were the most difficult stimuli to recognize, while “sh” was the easiest one to be identified among all combinations.

In summary, starting from consonant “d”, other three consonants in this experiment had obvious classified differences with “d”, which produced progressive level of difficulties on recognizing concurrent syllable pair from “sh”, “l” to “b”. Spectral contrast extract their energy differences from spectrum thus displayed a beneficial role in recognition performance as a spectral cue. However, due to the limited choice of vowels and consonants, not enough evidence could be explored to define the relative contributions among consonant, vowel and tone on the concurrent syllable recognition. The role of consonants in concurrent syllable recognition remained unclear.

CHAPTER 5 EXPERIMENT THREE: EFFECTS OF SPECTRAL CUE ON MANDARIN CONCURRENT-SYLLABLE RECOGNITION

Summary

An expansion of stimuli range was made on the basis of experiment two to evaluate the concurrent CV syllable recognition in a more general case. With more consonants and vowels involved, the relative weightings of three elements comprised syllables on recognition were investigated by fitting a power-function-based model. In addition, as a spectral cue, this experiment continually probe the effect of spectral contrasts between concurrent syllables on the recognition accuracy.

Results were consistent with last experiments (described in chapter 4) to show that, enlarging the spectral contrast enhanced the recognition of concurrent syllables, especially when both consonants and vowels differences were available. A further regression analysis revealed a new finding that, when two paired syllables only differed on consonants and tones (had identical vowel), the usage of spectral contrasts in recognition accuracy enhancement followed a linear scale, while logarithmic trends were found in other cases. Revised model with a correction coefficient added described the relative weightings among three elements. Consonants accounted for a bigger part of recognition accuracy compared with vowels and tones. Possible reasons and inspirations were introduced in the discussion section.

5.1 Introduction

Experiment two described in chapter 4 involved only four consonants and two vowels in order to mainly probe the effects of consonants with related classifications. Not enough evidence could be collected to investigate the relative contributions of consonants, vowels and tones on recognizing the concurrent syllables. In this experiment, subjects were asked to recognize syllables consisting of a wider selection of consonants and vowels categories, which extended the study to a more common case. After summarizing the results, a model based on power function was introduced for fitting and showing the relative weightings of three syllable elements on concurrent syllable recognition accuracy.

The recognition of such syllables presented simultaneously was expected to benefit from the cue of spectral contrasts, which was similar as the effects found in last two experiments. Especially when both consonants and vowels differences cues were available, listeners took advantage of the spectral contrasts to isolate and identify each syllable competing in a pair.

However, in mandarin, consonants sharing the same place or manner of articulation easily got listener confused (Duanmu, 2007). Consonant had the features that lasted short with less sound energy. The effect of spectral contrasts in concurrent-syllable recognition with more consonant categories might be weakened due to the finite energy change producing by consonant differences.

5.2 Hypothesis

- 5.2.1 In concurrent syllable recognition when more categories of consonants involved, spectral contrasts of syllables could provide cues for identifying each syllables, larger contrasts helped improve syllable recognition accuracy.
- 5.2.2 The information that spectral contrasts provided was limited. More categories of consonants aroused great difficulties that could not be resolved only by utilizing spectral contrasts.
- 5.2.3 Consonants were the hardest to detect among three syllable elements. Identification performance of consonants accounted for the most part of syllables recognition correct rate.

5.3 Method

5.3.1 Variables

Similar as last chapter, this experiment is an extension of experiment two with more categories of consonants considered. The independent variables mainly originated from the categories of elements comprised syllable pairs for identifying. These categories were vowel

category, tone category and consonant category. In addition, spectral contrast that calculated the mean squared error of two syllables spectral envelope in a pair was another independent variable to probe the effect of overall spectral difference on concurrent CV syllable recognition task.

The recognition performance was evaluated from four aspects, including syllable recognition, vowel recognition, consonant recognition and tone recognition with each counted separately.

All sets of variables were summarized in table 5.1.

| Independent Variable | |
|------------------------------------|-------------|
| Name | Type |
| Vowel category | Categorical |
| Consonant category | Categorical |
| Tone category | Categorical |
| Spectral contrast | Continuous |
| Dependent Variable | |
| Name | Type |
| Syllable correct recognition rate | Continuous |
| Vowel correct recognition rate | Continuous |
| Consonant correct recognition rate | Continuous |
| Tone correct recognition rate | Continuous |

Table 5.1 List of variables of experiment three

5.3.2 Stimuli

Syllable used in this experiment was an extension of last study to more common cases. Consonants were chosen with comprehensive consideration of frequency of use and classification. In the first step, according to the table of 2500 common used words in modern Chinese (Li et al., 2013), possible combinations of 21 consonants and 6 basic vowels with 4 tones were listed as candidates (The detailed table of each combination was in appendix5.1).

Second, only the most frequently used one would be reserved among the candidate consonants group having almost the same combinations with vowels and tones such as “z”, “c” “s” or “j”, “q”, “x”. These consonants were often pronounced with the same place but slightly differed on the manner of articulation, for example, “j” and “q” were both affricate but had unaspirated and aspirated property respectively. In mandarin, these consonants with similar articulation could only be combined with specific vowels and tones. Thus, only one of the grouped candidates was reserved as a representative to mix with other consonants from different groups. These criteria selected six consonants out to be the main consonants used for generating syllables. In addition, based on the findings of experiment two, when the identity of a consonant surpassed others, recognition would be easier by excluding other available choices. Three consonants articulated from retroflex, palatal, velar place were thus added to eliminate possible recognitions relied on exclusion. Nine consonants selected were bolded and marked with red color in figure 5.1.

| | | Place of articulation | | | | | | |
|------------------------|-----------------------|-----------------------|--------------|----------|----------|---------------|----------|----------|
| Manner of articulation | | Bilabial | Labio-dental | Alveolar | Dental | Retroflex | Palatal | Velar |
| | Plosive Aspirated | p | | t | | | | k |
| | Plosive Unaspirated | b | | d | | | | g |
| | Fricative | | f | | s | sh , r | x | h |
| | Affricate Unaspirated | | | | z | zh | j | |
| | Affricate Aspirated | | | | c | ch | q | |
| | Nasal | m | | n | | | | |
| | Liquid | | | l | | | | |

Figure 5.1 Selected consonants in the classification table

The selection of vowels depended on their possible combinations with initial vowels. In Mandarin, there are 35 final vowels that can be combined with initial consonants, which includes 6 simple vowels, 13 complex vowels and 16 compound nasal vowels. However, this

experiment was designed to extend the influence of consonants based on the study of concurrent simple vowel recognition. Only simple vowels was involved in this experiment. As complex vowels and compound nasal vowels have more complicated spectral features, future works could be conducted on such vowels recognition when they are concurrently presented. Among 6 basic vowels, syllable containing “o” only existed when combined with consonants from bilabial articulation group, while “o” had the lowest frequency of use. As a result, vowel “o” was not used in the experiment to generate stimuli.

Tone was chosen if the specific combination of a syllable belonged to the table of 2500 common used words in modern Chinese. For example, a combination “bu” only had tone 3 and tone 4 as single syllable in this experiment as it was not common used when carried tone 1 and tone 2.

Based on the elements selection, 57 single syllables were remained for mixing. A full combination of 1,596 pairs of concurrent-syllable pairs would be produced if any two mixed, whilst far exceeded subject’s tolerance in listening (Subjects attended experiment two spent around 80-100 minutes finishing tests of 496 sound pairs). A further reduction of total stimuli number was conducted by modifying the method of mix.

The mixing of single syllables followed the similar grouping in experiment two, which based on the principle that whether elements of two syllables in a pair were different. Four groups of stimuli were generated individually: (1) Same vowels, different consonants; (2) Same consonants, different vowels; (3) Same vowels and consonants, different tones; (4) Different vowels, consonants. The generation also referred to the usage frequency table of vowels (see Appendix 5.2), to remove redundant mixtures.

At last, 734 concurrent-syllable pairs were used as experiment stimuli. Each single syllable that used for generating was adjusted to control the cues functioned. Every individual syllable need three modifications after producing from TTS tool including adjusted the mean fundamental frequency to be 210Hz, equalized the duration and sound intensity at 450ms and 65dB (The reasons had been introduced in section 4.3.2). The concurrent syllable pairs were presented to the subjects through headphone binaurally at 65dB SPL.

5.3.3 Task and Procedure

The procedure followed the same as previous experiments. Subjects were asked to sit in front of a computer screen and listened to the sound stimuli one by one. The stimuli was only allowed to play once by clicking the button on the GUI operated by the subjects. A modified version of GUI was showed in figure 5.2.

Before the formal tests start, a training session was conducted to ensure subject could get familiar with single syllables that used to form concurrent pairs. Single syllables were played in order of initial consonants. A 20-trial test of single syllable recognition was then run for assessing subject's ability in identifying unmixed stimuli. Correct rate over 95% was considered as passing the training tests. Only one out of sixteen subjects failed to pass at the first time because of the confusion between dental consonant “z” and retroflex consonant “zh”. Another round of training was then provided until the subject passed the test at second time.

In the formal tests, thirteen sessions of stimuli containing 57 pairs each (except for one session had 50 pairs only) were presented randomly. There were 5-10 minutes break between sessions to alleviate fatigue. The whole experiment took around 120-150 minutes to finish.

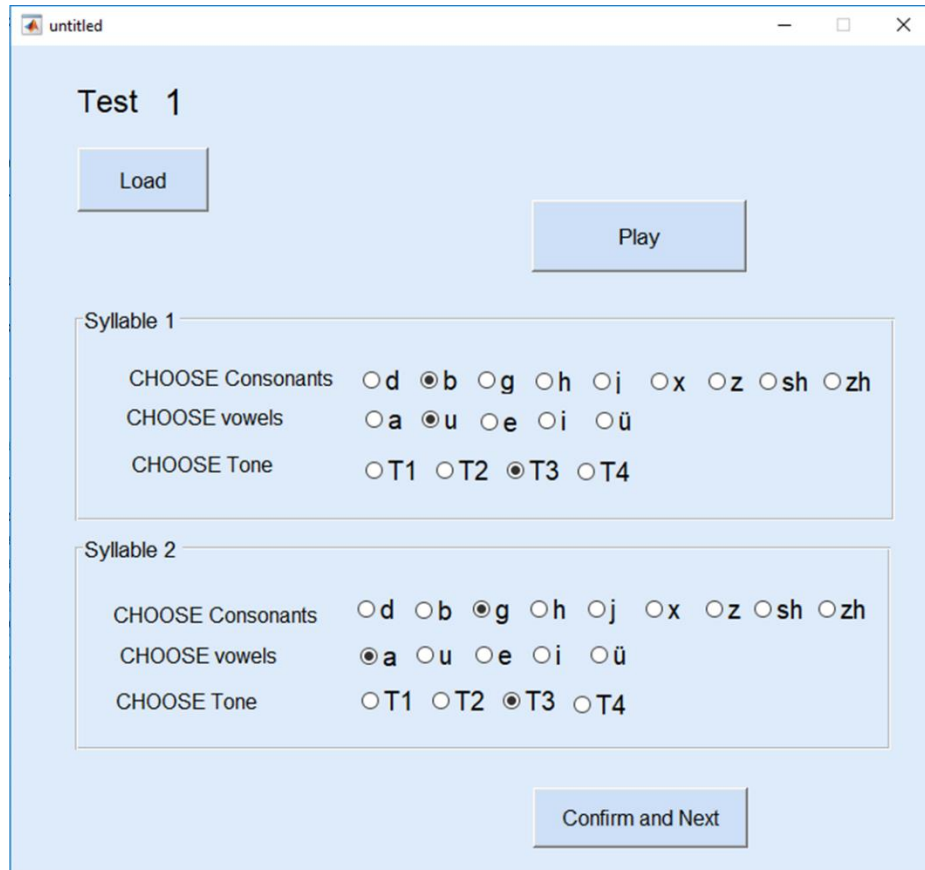


Figure 5.2 Screenshot of GUI used in experiment

5.3.4 Subjects

18 native-Mandarin listeners including ten males and eight females registered to participate in the experiment. All subjects were students from The University of Science and Technology without history of hearing impairment. Except for two subjects, sixteen out of eighteen subjects had passed the hearing test with pure-tone detection threshold below 20dB hearing level (HL) at frequency 250, 500, 1000, 2000, 4000 and 8000Hz. Two subjects failed to pass the test were rejected to continue the experiment. As a result, 16 subjects including eight males and eight females participated in the tests. They were aging between 22 and 29 with a mean of 24.6 years.

Among 16 subjects, five were from southern China. Influenced by the dialect, only one subject reported daily difficulties in differentiating dental and retroflex consonants (e.g. z and zh in this experiment). During her training session, more efforts were spent on listening to stimuli consisting of “z” and “zh” until she had confidence in identifying as well as passed

the tests of single-syllable recognition. For other southern subjects, they all passed the training session test without reporting of differentiation difficulties.

5.4 Results and analysis

5.4.1 Effect of spectral contrast on the syllable recognition correct rate

Based on findings from experiment 2 that when particular element cues available, the effect of spectral contrast provided different degrees of benefits. Thus the exploration on the function of spectral contrast was carried separately for different groups described in section 5.3.2.

5.4.1.1 Both vowels and consonants differences available

For groups with both vowel difference and consonant difference available, a scatter plot was drawn to show the change of syllable recognition correct rate as a function of normalized spectral contrasts in figure 5.3. A Pearson correlation test showed the relationship was significant ($r = 0.285$, $p = 0.010$). By digging into the data points that deviated from the trend, it was found that two specific categories contributed to most part of the aberrant results. First, when consonant “sh” mixed with consonant “z”, correct rates were largely dropped (all below 49.5) even when the spectral contrasts reached 0.58. Second, pairs that include “di” as one of the competing syllables also led to low rate of correct recognition (no more than 48.7). The reason for low accuracy would be discussed in the discussion part.

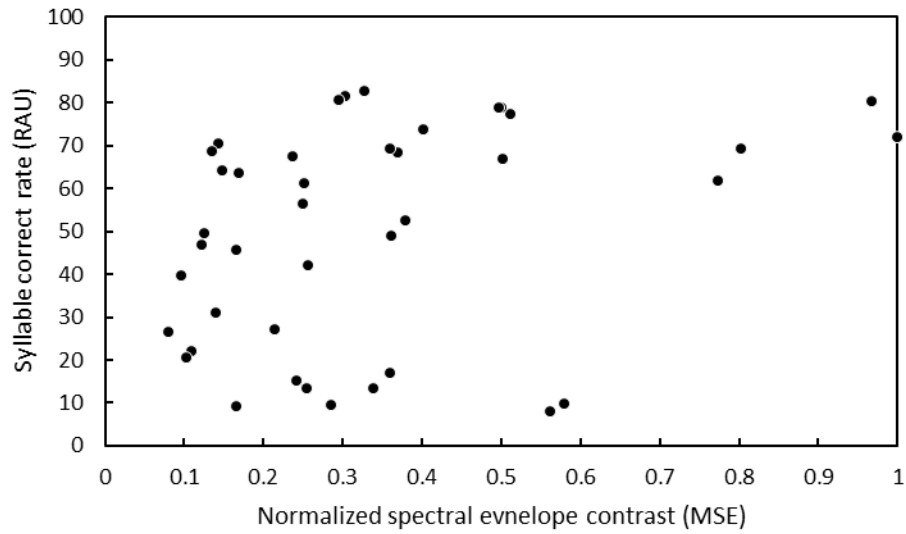


Figure 5.3 Scatter plot of syllable correct rate as a function of normalized spectral envelope contrast

These two categories of data was removed from the group data and figure 5.4 showed the replot results. Pearson correlation tests showed a significant result ($r = 0.502$, $p < 0.001$). A further regression line was fitted and it was found that compared with linear regression, the trend of the relationship between spectral contrast and correct rate was more of logarithmic scale ($r\text{-squared} = 0.38$ while $r\text{-squared} = 0.24$ for linear). This discovery was consistent with experiment 1 and 2 where listeners utilized the information of spectral contrast in a nonlinear way.

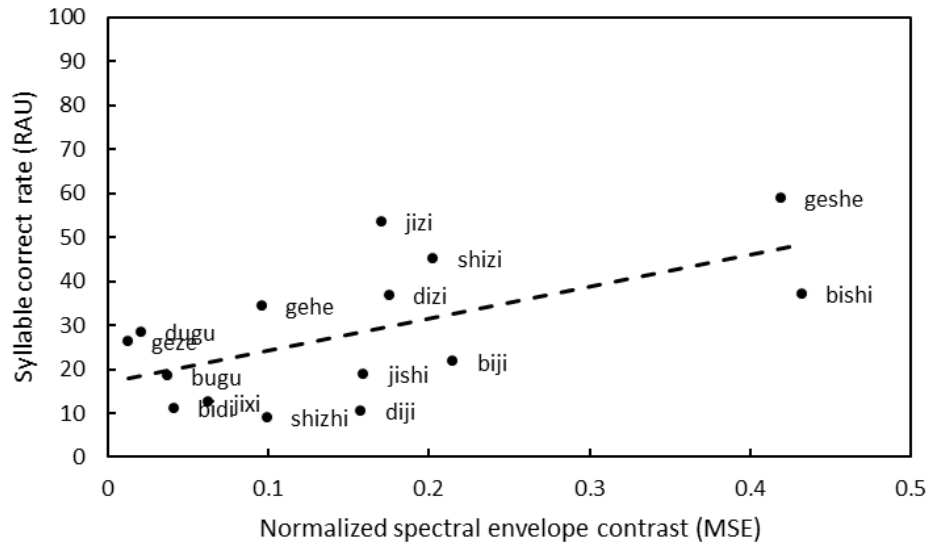


Figure 5.5 Scatter plot of syllable correct rate as a function of spectral contrast with only consonant differences available

5.4.1.3 Only vowels difference available with identical consonants

Due to the limited number of concurrent-syllable pairs, only seven kinds of composition existed when vowel differed and consonants remained identical in a concurrent pair.

The results of this group was showed as scatter plot in figure 5.6, where a positive relationship was observed between spectral contrasts and syllable correct recognition rate. As the data only contained seven points, statistical analysis was not conducted to avoid biased results.

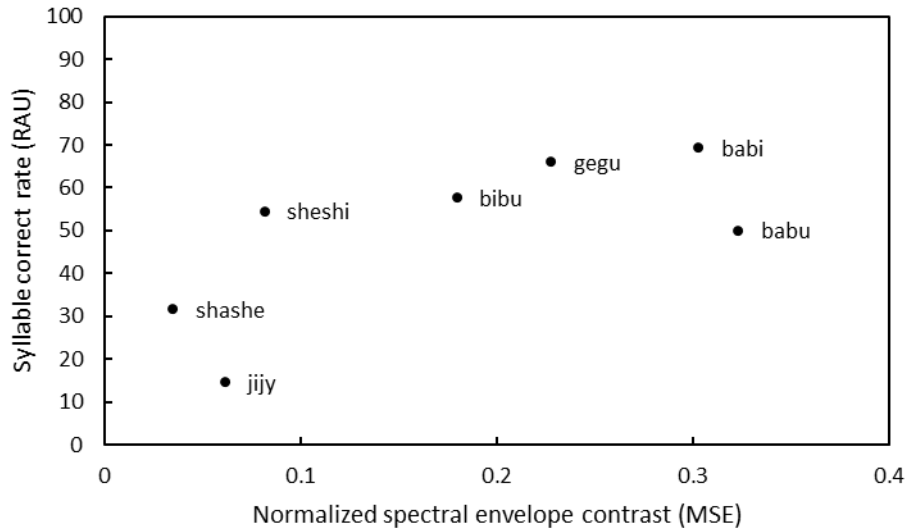


Figure 5.6 Scatter plot of syllable correct rate as a function of spectral contrast with only vowel difference available

As in the group where neither consonant nor vowel difference was available, tone provided most part of the information as a temporal cue. The spectral envelope contrast didn't take temporal information into consideration, thus this group of data was not included for analyzed.

Combined the results from three groups described, linear relation appeared when only consonant difference available. However, the highest accuracy only reached 58.8 from pair “ge-she”, while the correct rate reached 82.6 and 69.4 for two groups with vowel difference available respectively.

5.4.2 Comparison of syllables, vowels and tones correct recognition rate

Element-based recognition performance was investigated by comparing the correct rate of different measurements. The averaged correct recognition rate of consonants, vowels, tones and syllables were plotted as a bar chart showed in figure 5.7. One-way repeated-measure ANOVA showed that significant difference existed among four measurements [$F(3,45)=428.33$, $p<0.001$]. In post-hoc multiple comparison, any measure of recognition performance was significantly distinct from other three.

Among three elements (consonants, vowels and tones), consonants had the lowest accuracy while vowels were the easiest to be recognized. As a result, the error in concurrent CV syllable recognition was more possible from confusions in identifying consonants.

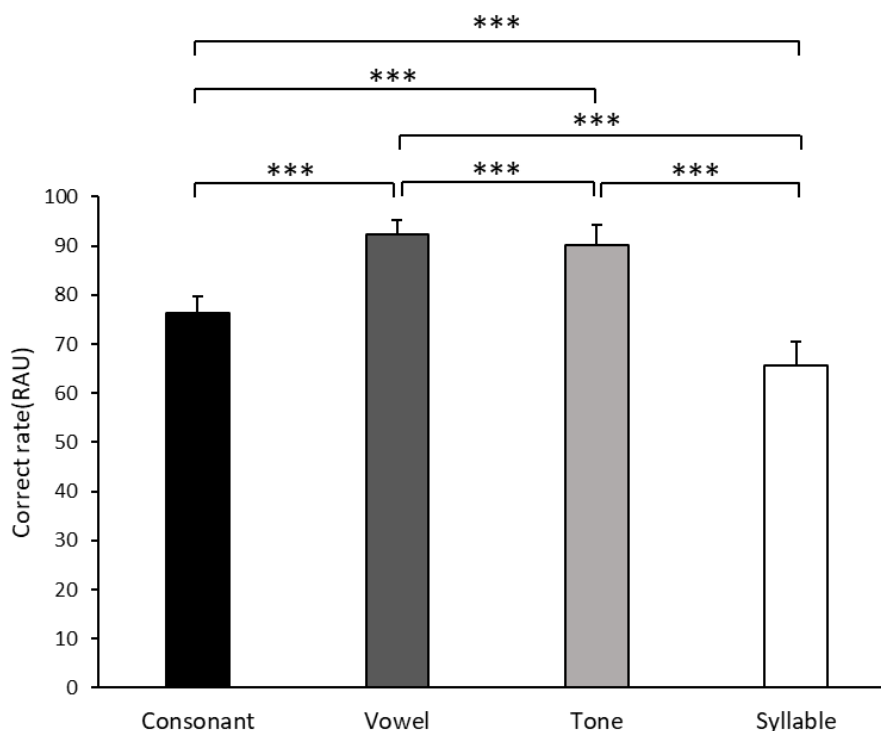


Figure 5.7 Averaged correct recognition rate of consonants, vowels, tones and syllable, *** represented for $p < 0.001$

5.4.3 Effect of tone category on the tone recognition correct rate

Tone recognition performance was evaluated in terms of tone confusion matrix showed as figure 5.8, where horizontal line represented for tone presented in the stimulus and vertical column displayed subject's response of corresponding stimulus. To understand the matrix, the value in a "T1-T1" grid gave the information of average correct rate on tone 1 recognition, while a "T1-T2" grid showed the average percentage of response that tone 1 was wrongly recognized as tone 2. The remaining grids could be explained in the same way. Background color of the grids was drawn along the color bar on the right side, a darker blue represented for lower error rate and a brighter yellow stood for higher correct rate. Tone 3

had the lowest accuracy (84.3%) among four tones, while tone 4 was the least likely one to be wrongly identified.

A one-way repeated-measures ANOVA revealed that main effect of tone category was significant [$F(3,45)=13.85$, $p<0.001$]. In order to investigate the relative relations among tone categories, post-hoc multiple comparisons with Sidak correction was conducted to show that recognition performance of Tone 3 was significantly worse than other 3 tones ($p < 0.002$), whilst no significant difference was found between any two of other tones ($p > 0.194$). This finding was consistent with experiment one, where tone 3 brought the biggest difficulty on recognizing tones of concurrent vowels.

One possible reason was that most tone information was carried by temporal change of fundamental frequency curve. Longer duration of vowels than consonants delivered the large parts of clue on tone identification. As a result, in the concurrent syllable recognition task when an initial consonant was available, similar confusions could be found among four tones.

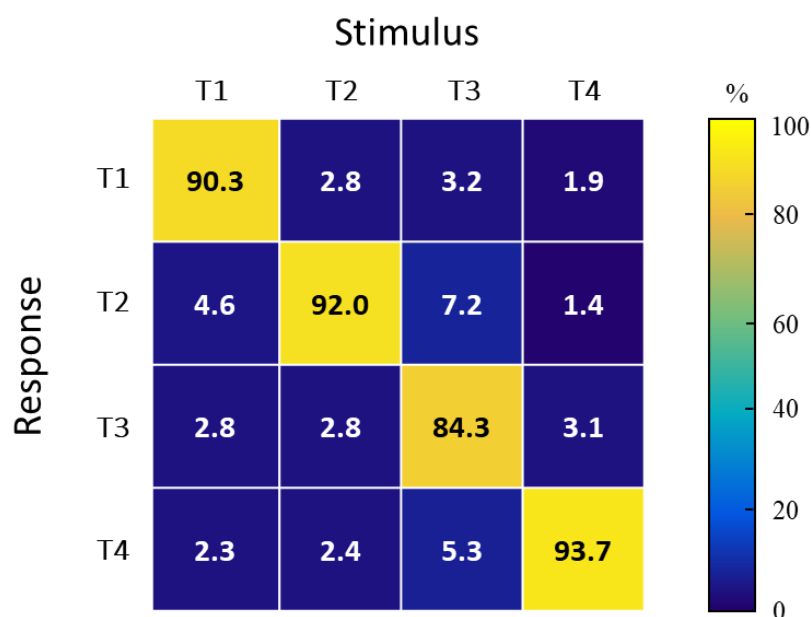


Figure 5.8 Confusion matrix of tone recognition from concurrent-syllable pair

5.4.4 Evaluate relative contributions of consonants, vowels and tones by model fitting

To investigate the relative contributions of consonants, vowels and tones on recognizing concurrent syllables, a model fitting was introduced based on past studies (Fletcher, 1995; Fu et al., 1998).

The probability of recognizing a syllable was evaluated by product of consonants, vowels and tones recognition with a corresponding weighting coefficient. As their studies were used for sentences or syllables recognition without competing sound sources, a correction coefficient was added accounting for concurrent masking of syllables in our cases. An equation based on power function was then used to describe the relationship:

$$p_s = ap_c^{w_c} p_v^{w_v} p_t^{w_t}$$

where p_s , p_c , p_v and p_t represented for correct recognition probability of syllables, consonants, vowels and tones respectively, w_c , w_v , w_t were the weightings and a was the correction coefficient. The data used for model fitting was from all syllable pairs in recognition task (no elimination of pairs due to specifically low accuracies). Fitting outputs revealed that the weights for three elements were $w_c = 0.86$, $w_v = 0.48$, $w_t = 0.74$ and correction $a = 0.93$ with r-squared = 0.93.

The results demonstrated that in a concurrent syllable recognition task, consonants contributed to the largest portion of challenge. As a result, an improvement of consonant intelligibility in concurrent syllable pairs helped enhance the recognition of individual syllable. In addition, the correction coefficient smaller than one showed a simultaneously presented competitor would deteriorate the recognition of single syllable.

5.5 Discussions and conclusions

Experiment 3 included more categories of consonants and vowels in mandarin so as to explore the concurrent CV syllable recognition performance from a more general perspective. The role of spectral contrasts was reinvestigated and the results showed conclusions from previous experiments still hold in this case except for the linear trend in specifically grouped syllable pairs.

When two syllables in a pair possessed different vowels, spectral contrasts could be utilized to a large extent to improve the recognition accuracy regardless of consonant difference. This was manifested in the relatively high correct rate in syllable group with vowel difference available. As for the group when consonants difference existed but vowels differences were absent, despite the positive effect that spectral contrasts brought, the room for enhancement of the syllable recognition was limited, which reflected by relatively low correct rate for all combinations in this group. Recognition of consonant was not that easy mainly due to the intrinsic property. The unvoiced consonants used in the experiment had shorter duration and less spectral power than vowels. As the categories of consonants increased, confusions appeared in identifying each of them.

Additionally, it was showed in section 5.4.1.1 that when syllables with “sh” and “z” were combined together, the performance was extremely degraded. One possible reason could be their close locations of energy peak on the spectrum. That is to say, as they produced large amount of energy in similar range of frequency, listeners could not use this spectral cue to do the differentiation. Another difficulty appeared in consonant “b” and “d”, which was also observed from experiment 2. From the results, when syllable consisting of “di” mixed with other syllables, listeners tended to wrongly recognized “d” as “b”. Both “b” and “d” were plosive unaspirated with similar spectral characteristic. Thus, as spectral contrasts only accounted for part of the spectral cues, this factor in concurrent syllable recognition was not as effective as in concurrent vowel recognition. Other cues beneficial for recognizing consonants should be explored to enhance the recognition performance.

Tone recognition in concurrent syllable was assessed by using confusion matrix. The results were highly consistent with the first experiment where concurrent vowel recognition was evaluated. Tone 3 identification had the significant lowest accuracy compared with other three tones. The “falling-rising” characteristics of tone 3 produced two directions of pitch change temporally, which might raise the difficulty of recognition. In natural speech, syllables with tone 3 had the longest duration, while in our cases, the cue of syllable duration was eliminated by equalizing each syllable to have the same length regardless of the tone category. This could be another factor that made tone 3 hardest to be perceived and identified.

A power function based model was introduced to examine the relative contributions of consonants, vowels and tones in recognition task. The model fitting results showed that consonants weighted more than vowels and tones in probability to recognize a syllable from concurrent presented competitor. It was for the reason that, when two CV syllables presented together, the biggest difficulty emerged from confusions between consonants. Vowels had clear formants on the spectrum as well as fundamental contour carried the information of tones. In contrast, not all consonants had a concentration of energy on the spectrum. Diverse natures of consonants formed confusions in recognition. In addition, a correction coefficient smaller than one indicated that when two syllables presented together, one could mask another thus increase the difficulty of recognition compared with condition that a syllable played alone. In order to produce higher accuracy in concurrent syllable recognition, emphasis on cues that help identify consonants was needed besides enlarging the spectral contrasts.

CHAPTER 6 COMPARISON OF RECOGNITION ACCURACY BETWEEN HUMAN AND DEEP-LEARNING MODEL

Summary

A deep-learning model was introduced to evaluate the performance on concurrent syllable recognition. As the state-of-the-art, many studies worked on improving model performance from optimizing structure or algorithms. This experiment showed a comparison between recognition accuracy from human and model, and the results was expected to give useful insights on model iteration.

Two separated comparisons were conducted on concurrent vowel recognition and concurrent syllable recognition where the syllable consisting of initial consonants and final vowels. In concurrent vowel recognition task, the model showed almost consistent way of utilizing information of spectral contrasts between two simultaneously played syllables. By evaluating from three metrics, a rise of separated file quality was significantly correlated with larger spectral contrasts. Nonetheless, the spectral information was linearly applied in model's separation.

However, in concurrent syllable recognition with consonants involved, the model showed no significant feature of separating concurrent pairs having different spectral contrasts. Possible reasons for such unexpected results were discussed.

6.1 Introduction

With the development of deep-learning technology, emerging works devoted to solve real problems. Many applications could be seen in daily life such as artificial intelligent (AI) assistant interacted with users, speech enhancement tools for filtering noise. The-state-of-the-art triggered our interests to do a comparison between human and machine so as to improve performance of model by applying human experience.

Current studies enhance the system performance by optimizing the structures or algorithms of the model. However, a conjecture was put up to be verified that, if cues that utilized by

human in speech separation and recognition could provide useful insights in improving model performance.

Based on the idea, the model and human performances on the same recognition task were compared in this experiment to evaluate if the cues beneficial for human would have similar positive effect on model.

6.2 Hypothesis

- 6.2.1 In both human and model recognition of concurrent vowels/syllables, the spectral contrasts between two syllables in a pair helped enhance each syllable intelligibility. A larger spectral contrast was related to higher recognition accuracy.
- 6.2.2 Unlike human's nonlinear perception system, the utilization of spectral information on speech separation by model was of linear scale.

6.3 Method

6.3.1 Model description

Model used in this experiment was proposed by a study that was reviewed in chapter 2 (Luo & Mesgarani, 2019). Three-stage processing was conducted to give two separated audio files, while input was the mixtures containing each concurrent-syllable pairs. The input files for testing were exactly the same as used in human experiments described in chapter 3 and 5. Stimuli from experiment 2 was not included as it comprised a subset of syllables in experiment three.

As a cooperated work, the training and implementation of Conv-TasNet were completed by my group mate Jun, HUI. Following summary and statistical analysis of model performance was done by the author.

The setting of training was summarized in table 6.1:

| Category | Setting |
|---------------|-----------|
| Optimizer | adam |
| Learing rate | 0.001 |
| Training set | THCHS-30 |
| Sampling rate | 16,000 Hz |

Table 6.1 Parameters in training model

Materials of training sets was from THCHS-30 (Wang & Zhang, 2015) that included 198, 252 words recorded by native mandarin speakers. The coverage of bi-phoneme words and tri-phoneme words were 71.5% and 14.3% respectively.

6.3.2 Metrics for comparison

When evaluate the performance of machine, a meaningful measurement would be preferred if the metrics could be comparable with human perception. The main goal of speech enhancement is to improve the intelligibility and quality of speech for a better delivery of information. As a result, three metrics that are widely used in this field were selected to assess model performance in separating two syllables presented simultaneously.

6.3.2.1 Short-time objective intelligibility (STOI)

The score of short-time objective intelligibility (STOI) is a common measurement used to predict the intelligibility of processed or complex speech (Taal et al., 2010). STOI score falls within the range between 0 and 1, while was expected to have a monotonous relationship with the average intelligibility of testing audio files. As a result, a higher STOI indicates the better speech intelligibility.

For the comparison in this experiment, a STOI difference (subtract the input score from the output score) was used as a reflection of the stimuli intelligibility improvement. Such measurements considered the syllable intelligibility change due to the model separation work so that could be comparable with human listeners.

6.3.2.2 Perceptual evaluation of speech quality (PESQ)

At first, perceptual evaluation of speech quality (PESQ) was developed for assessing speech quality of telephone networks and codecs (Rix et al., 2001). PESQ speech evaluation system is an objective method to evaluate sound quality based on auditory model and cognitive modelling. By taking the auditory characteristics of the human listeners into account, the method included human auditory perception model to enhance the assessment of speech quality. Compared to past metrics, the evaluation results from PESQ could be closer to listener's real experience. As a result, this metric has become a widely used evaluation method for various works such as communication systems and speech separation and recognition.

The mean opinion scores(MOS) is indicated by PESQ that is divided into 5 levels according to real listener's experience showed in table 6.2, a bigger value represented for higher quality of the audio stimulus.

| Level | Value range | Listener experience |
|-----------|-------------|--|
| Excellent | 4-5 | Very good, extremely clear intelligibility |
| Good | 3.5-4 | Good, clear intelligibility with slight noise |
| Fair | 2.5-3.5 | Fairly well, not clear intelligibility with some noise |
| Poor | 1.5-2.5 | Barely well, poor intelligibility, noisy |
| Bad | 1-1.5 | Very poor, hard to perceive |

Table 6.2 Evaluation level of PESQ based on real listener experience

In this experiment, PESQ would be one of the metrics that used to summarize model output so as to match human listener's performance in concurrent syllable recognition.

6.3.2.3 Signal-to-distortion ratio (SDR)

Similar as the signal-to-noise ratio, the signal-to-distortion ratio (SDR) is a metric that used to describe the relative level of desired signal to the distortion. SDR is calculated by the ratio of signal power to the distortion power thus is expressed in decibels (dB). When the ratio is higher than 1, that is 0dB, the signal is powerful than distortion. As a result, the higher the SDR, the better the quality of speech.

We also measured the SDR difference from the output and input of model as an indicator of syllable quality change after model separation.

6.4 Results and analysis

The model performance was evaluated separately for two recognition tasks corresponding to experiment one and three (described in chapter 3 and 5).

6.4.1 Comparison of performance on concurrent-vowel recognition

The first comparison was conducted on the recognition of concurrent vowels. Human's data was from experiment one. In human listeners, the spectral contrast played a significantly important role in improving correct recognition rate with a logarithmic scale. The effect was especially manifest in low contrast range.

In model output, the STOI and PESQ scores of each concurrent-syllable pair were plotted as a function of spectral contrasts in figure 6.1. Since the range of STOI score was between 0 ~ 1, a transforming of the score to fall in 0 ~ 100 by multiply 100 was conducted to mimic the correct rate for human. In each output pair, the metric measured the average quality of two separated syllables. It could be observed that both the STOI and PESQ score had a relatively flat distribution.

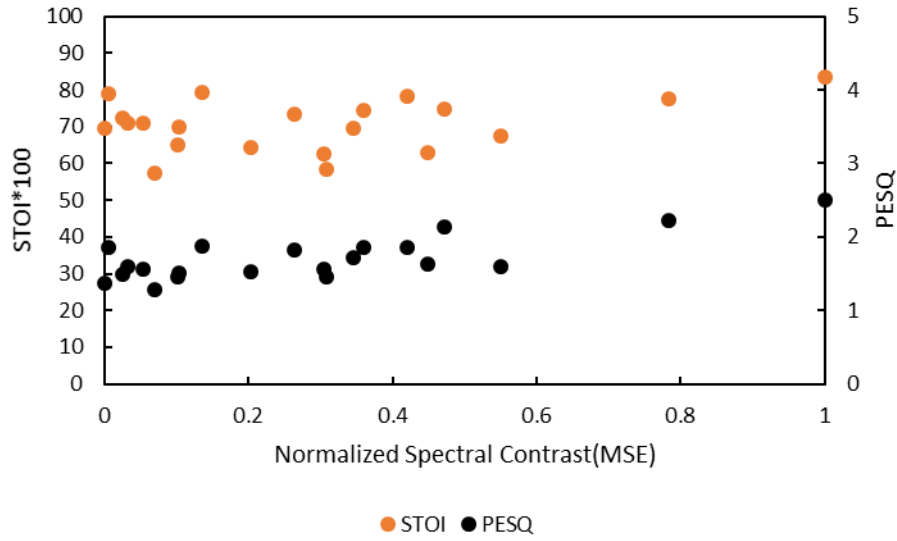


Figure 6.1 Model output of STOI and PESQ scores

However, in order to reveal the improvement of speech quality from separation of the model, adjusted scores (calculated as the difference from input to output) for three metrics were further investigated.

The STOI score difference was plotted in figure 6.2. The score was also multiplied by 100 to fall in between 0 ~ 100 to match the correct rate for human. A Pearson correlation test showed that normalized spectral contrast was significantly correlated with the STOI score improvement ($r = 0.554$, $p = 0.009$). A linear regression was conducted with a random distribution of residuals, suggesting the linearity existed between spectral contrasts and STOI score change. Though three points were negative reflecting a decrease in quality, most syllables had higher enhanced intelligibility as spectral contrasts increased. At most an improvement of 23.3 was observed when spectral contrast became the biggest.

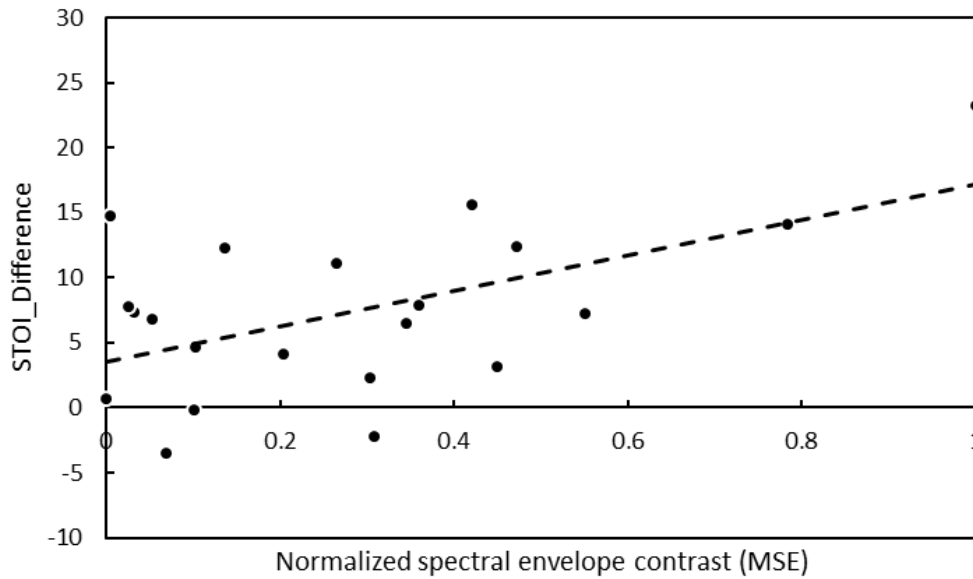


Figure 6.2 STOI score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent vowel separation

Similar results were found when plotting the PESQ score difference as a function of spectral contrasts. In figure 6.3, the score change of PESQ had significant correlation with the spectral information ($r = 0.799$, $p < 0.001$), which reflected by the linear regression results. Random pattern of residual plots suggested the linearity held. When the spectral increased to the largest for concurrent-vowel pair “au”, the improvement of PESQ exceeded 1.2, nearly a level up according to the table 6.2.

The higher correlation coefficient of PESQ than STOI revealed that the PESQ measurement was possibly more sensitive to the spectral information.

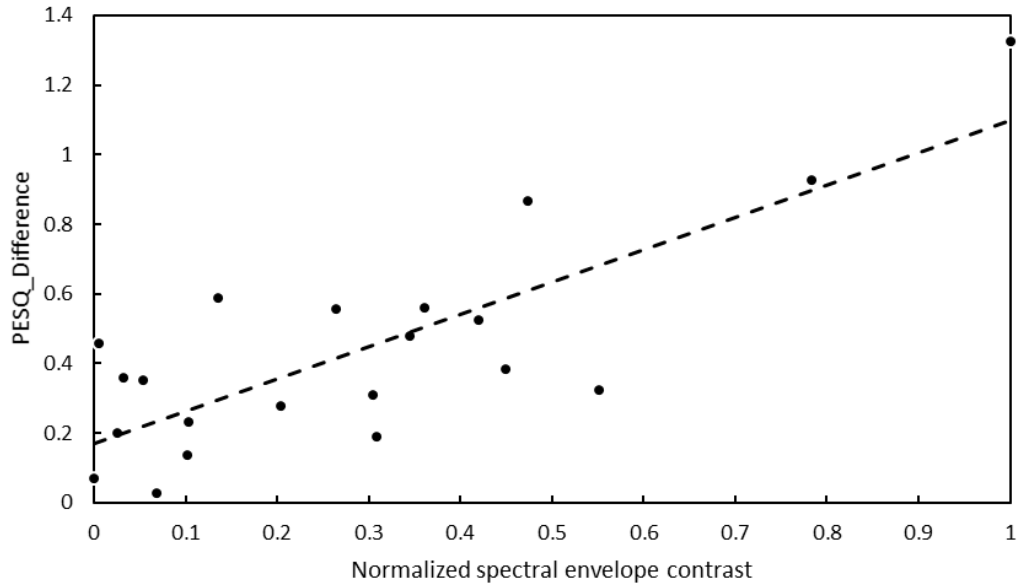


Figure 6.3 PESQ score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent vowel separation

At last, the SDR difference was plotted similarly as a function of spectral contrasts showed in figure 6.4 as an objective reference of separated vowel's intelligibility. Pearson correlation tests showed a significant relation between spectral information and SDR improvement ($r = 0.771$, $p < 0.001$). The linear regression was supported by random distribution of residuals. For pair consisting of higher spectral contrasts between two vowels, the SDR had better enhancement. A coefficient of 0.771 showed us that, from objective evaluation, the spectral contrasts also provide important information for model to do the separation tasks. A ratio of 16.6 was improved when the spectral contrast were largest for concurrent vowel pair "au".

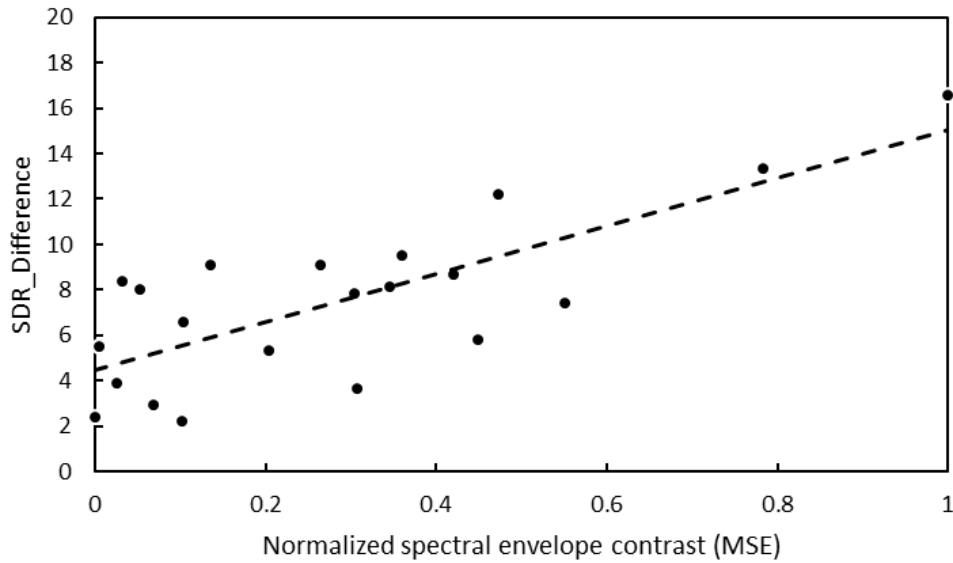


Figure 6.4 SDR score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent vowel separation

In general, model performance in concurrent vowel separation showed consistent results with human listeners despite the degree of utilization. It was showed in previous chapter that due to the nonlinearity of human perception, logarithmic relationship was found. As for model's utilization of spectral contrasts cue, higher amounts of spectral information led to a linear improvement in output vowel's quality.

6.4.2 Comparison of performance on concurrent-syllable recognition

When an initial consonant was added to the vowel to form the syllable, degraded performance was expected from both human and model on recognizing the concurrent syllable pairs. It was showed in experiment 2 and 3 that the existence of consonant aroused big difficulties for human listener to identify possibly due to their unvoiced features and relatively short durations.

The performance of model was also evaluated from three metrics calculated as the difference between the input and output ($\text{score_output} - \text{score_input}$). To be consistent with the settings in human experiments, the concurrent syllable pairs were divided into three groups : (1)

Different vowels, different consonants (DCDV); (2) Same vowels, different consonants (SVDC); (3) Same consonants, different vowels (SCDV), which was consistent with the grouping in experiment three. Model output was also evaluated separately for three groups.

6.4.2.1 Model performance for different vowels, different consonants (DCDV) group

In figure 6.5, the STOI score before and after the model separation was plotted as a function of spectral contrasts for DCDV group where both elements differences were available. No observable trend could be found from the plot, indicating the STOI score improvements were not influenced only by spectral information for concurrent-syllable pairs. A Pearson correlation tests showed no significant correlation was found ($p = 0.695$). Despite the insignificant relationship, all degraded-quality syllables (with negative STOI change) were found only for spectral contrasts value lower than 0.55.

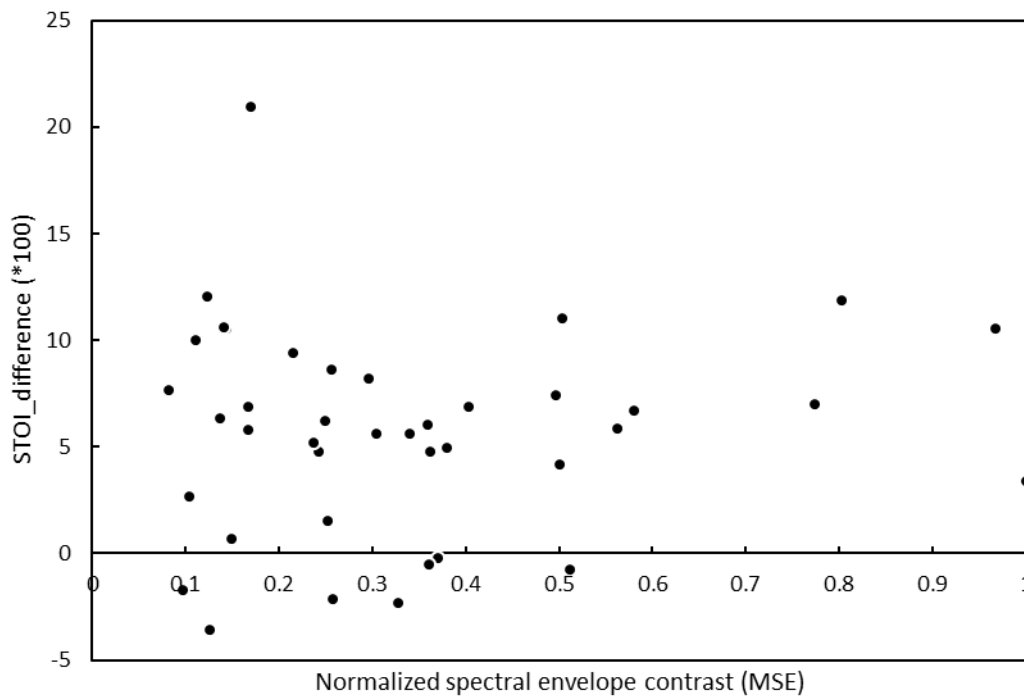


Figure 6.5 STOI score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent syllable separation (DCDV group)

In addition, the evaluation from PESQ difference was showed in figure 6.6. A relatively flat distribution of PESQ score change was found along the variation of spectral contrasts.

To our surprise, the two syllable pairs with the highest PESQ score (62.6 and 62.5 respectively) improvement had relatively low spectral contrasts (at around 0.13). But both syllable pairs had very similar features: “bi-ge” and “di-ge”. Both pairs composed of vowels “i” and “e”, while consonants only differed in one of the syllables. The similarity of “b” and “d” has been discussed in chapter 4 and 5, they had same manner but different place of articulation. Thus, the two pairs could be concluded to have the same pattern that made the separation of model easier compared to other syllable pairs. However, the reason behind it still need to be investigated.

A Pearson correlation tests was also conducted to show no significant relationship between spectral contrasts and PESQ score change in this case ($p = 0.721$). The effect of spectral contrasts was not dominant when initial consonants involved.

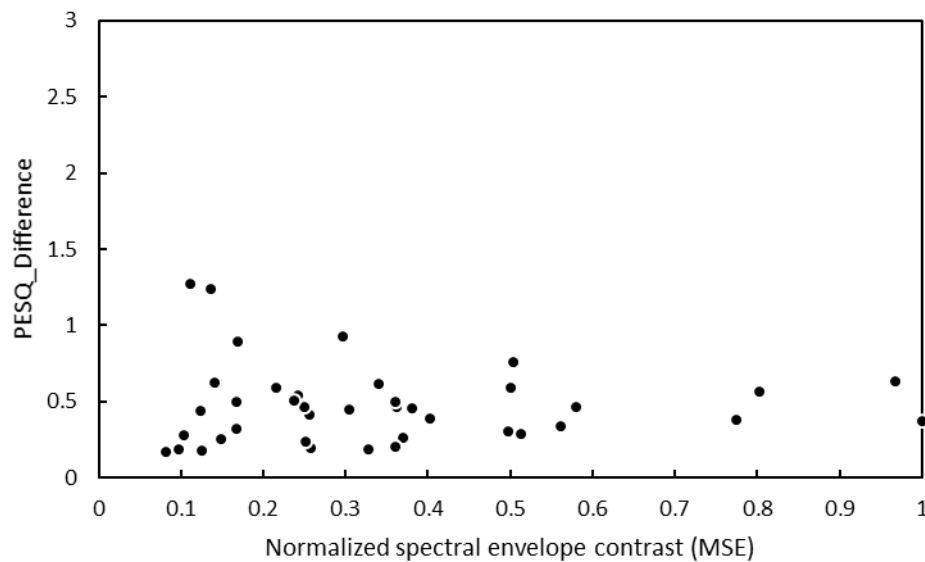


Figure 6.6 PESQ score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent syllable separation (DCDV group)

Objective evaluation of separated syllables was done through the measurements of SDR change. Figure 6.7 showed the SDR difference plotted as a function of spectral contrasts. The result was similar as STOI and PESQ change. For most syllable pairs, the improvement in SDR was not affected by increasement of spectral contrasts. Pearson correlation tests showed that there was no significant relationship between them ($p = 0.377$).

It was also surprising that from the measurement of STOI, PESQ, SDR change, the highest improvement was not consistent in three metrics.

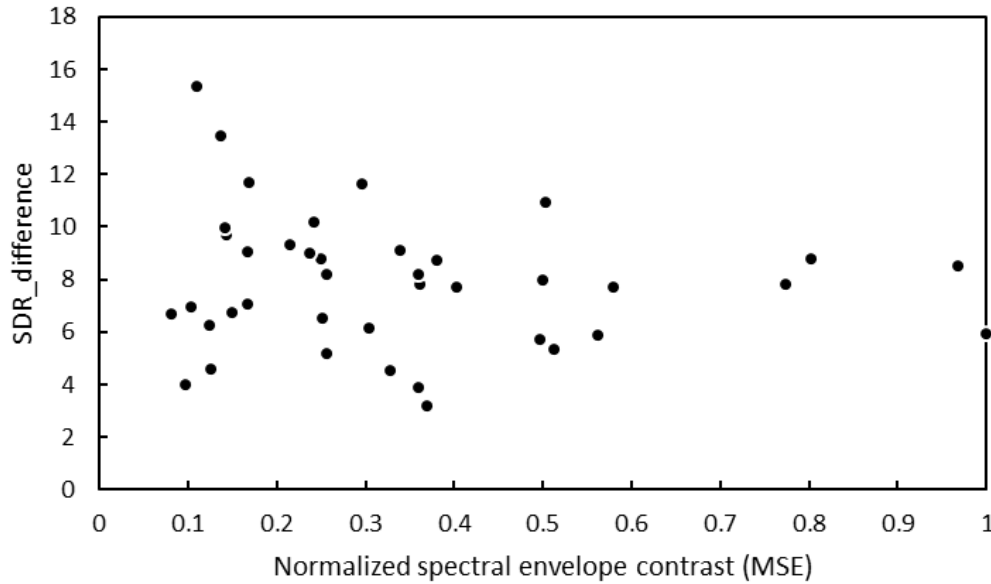


Figure 6.7 SDR score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent syllable separation (DCDV group)

6.4.2.2 Model performance for same vowels, different consonants (SVDC) group and same consonants, different vowels (SCDV) group

For both SVDC and SCDV groups, only one element difference was available. So an integrated evaluation of two groups' performance was showed in this section for a better comparison.

In the measurement of STOI, figure 6.8 showed a scatter plot for STOI difference changed with increasing spectral contrasts. It could be clearly observed that for black dots that was from SVDC group data, a linear correlation existed between the spectral contrasts and STOI improvements. Syllable pairs that had low spectral contrasts (lower than 0.05) showed degraded STOI score after the separation of model. Statistical tests revealed that the correlation was significant with a high correlation coefficient value ($r = 0.859$, $p < 0.001$). The residual plots from regression showed a random pattern, which supports the linearity.

For SCDV group, except for the “ba-bu” syllable pair which showed negative STOI change (-0.97), other data points had a roughly positive relationship between spectral contrasts and STOI difference from observation. Due to the limited selection of combination ways of consonants and vowels, statistical analysis was not done in SCDV group to avoid biased conclusions coming from big data variations.

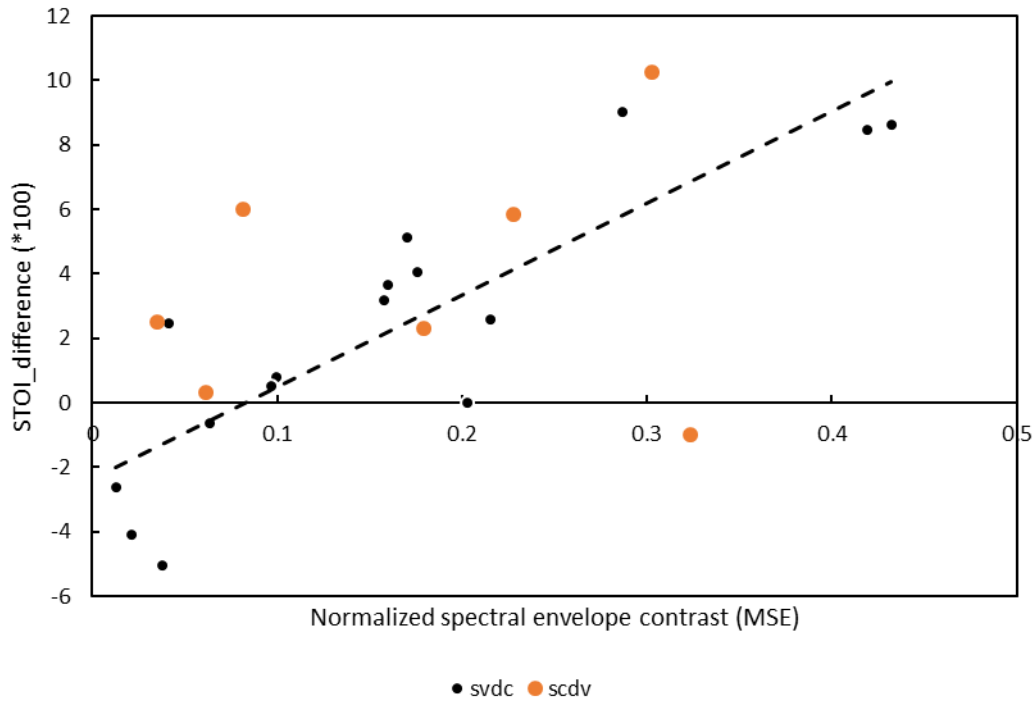


Figure 6.8 STOI score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent syllable separation (SVDC and SCDV groups)

The evaluation of PESQ score change was plotted in figure 6.9. For SVDC group, though relatively flat compared to STOI score improvement, an increasing of PESQ score change could be observed as the spectral contrasts enlarging. A Pearson correlation tests showed that the increasing trend was significant ($r = 0.537$, $p = 0.032$). Random distribution of residuals reflected the linear regression was reasonable. The three kinds of syllable pairs had highest PESQ score difference were with the top three biggest spectral contrasts.

As for the SCDV group, same as STOI score change, all data showed a linearly increasing trend between spectral contrasts and PESQ score difference except for the one who had the

largest spectral contrasts value (syllable pair “ba-bu”). A relatively highest value of PESQ was acquired (0.92) for the syllable pair possessing second-large spectral contrasts (syllable pair “ba-bi”). The low quality for “ba-bu” probably reasoned by model error.

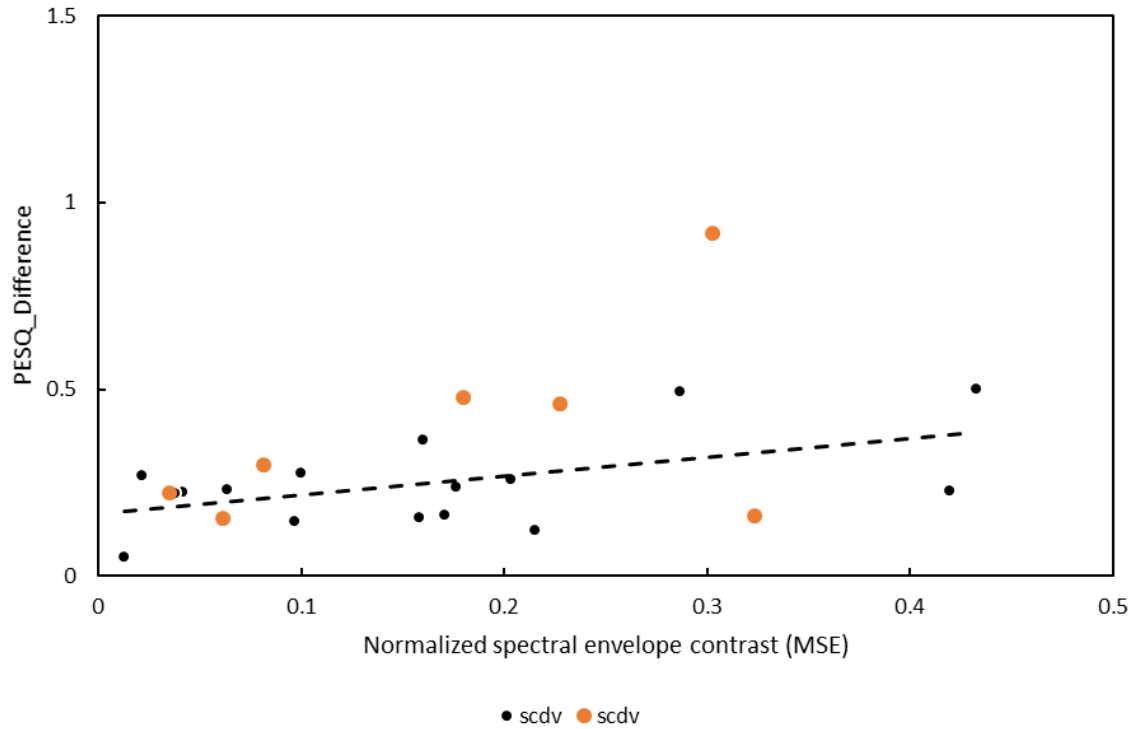


Figure 6.9 PESQ score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent syllable separation (SVDC and SCDV groups)

The SDR value change of two groups was plotted in figure 6.10 as an objective evaluation of output sound files. First, in SVDC group a consistent finding with STOI and PESQ score was found. As the spectral contrasts increased, signal become more surpass the distortion for the separated syllables. Pearson correlation tests showed the trend was significant ($r = 0.569$, $p = 0.021$). A random pattern in the residual plots from regression showed the linearity between variables.

SCDV group also showed similar trend of improvement. For pairs had more spectral contrast information, the SDR improved to a larger extent. Despite the relatively low change in pair “ba-bu”, spectral contrasts played an important role when model separated the concurrent syllables and thus improved the intelligibility for each syllable.

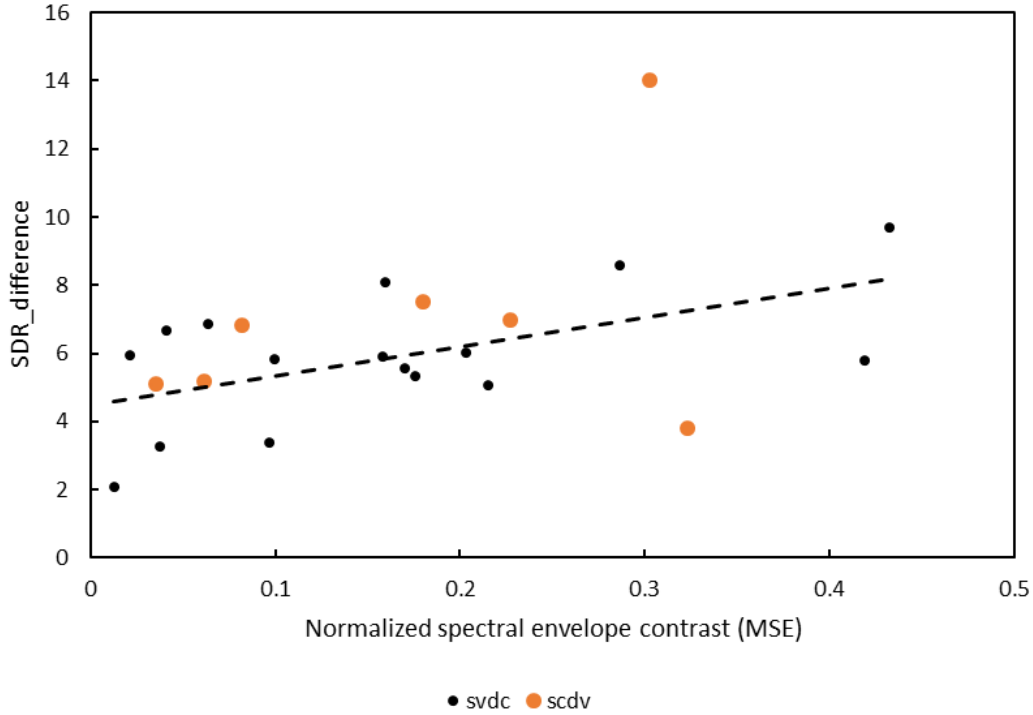


Figure 6.10 SDR score difference from model output plotted as a function of normalized spectral envelope contrast in concurrent syllable separation (SVDC and SCDV groups)

Generally, both SVDC and SCDV groups showed the effects of spectral contrasts on improvement of separated syllable quality were positive. It should be noticed that although the biggest contrast value in SCDV group was still lower than two values in SVDC group, the largest improvement from three metrics were in SCDV group were all higher than the largest one in SVDC group.

6.5 Discussions and conclusions

In this experiment, the performance of deep-learning model on separating concurrent syllables was thoroughly evaluated in order to compare with human listeners. Though limitation existed in matching the real accuracy between human and machine, the comparison had a certain degree of reference value.

Part of the results in this experiment was of expectation. In concurrent vowel recognition task, model could effectively separate two vowels with higher sound quality under the

influence of spectral contrasts information. The effect of spectral cue followed a linear scale for all three metrics, which implied the spectral contrasts in the test stimuli help model to separate two mixed signals. Consistent with human, when the spectral contrasts increased, higher accuracies were found.

However, when model was used to separate concurrent syllables where consonants were involved, the information provided by spectral contrasts was not significantly effective with both vowels and consonants differences presented. Specifically, when two syllables comprising the concurrent pair differed on vowels and consonants, recognition accuracy didn't rely on the variation of spectral contrasts. In contrast, if the consonants differences existed with identical vowel in a concurrent-syllable pair, the separation accuracy was significantly improved by enlarging the spectral contrasts. This implied that the spectral contrasts could still facilitate concurrent-syllable separation by machine in the absence of vowel difference. It was also found in experiment three that human listeners showed a linear utilization of spectral contrasts on improving recognition accuracy when only consonants difference presented. As a result, human and machine performance showed a relatively high similarity in this condition.

CHAPTER 7 SUMMARY OF RESULTS AND DISCUSSIONS

7.1 Summary of findings in the current study

This study has new findings in terms of concurrent vowel recognition in Mandarin. A logarithmic utilization of spectral contrasts were shown, which was inconsistent with past studies (Fu, Wu, et al., 2019). In addition, it was the first study that evaluated the role of spectral contrasts information in concurrent syllable recognition. Significant improvements in recognition accuracy were found to link with higher amounts of spectral contrasts. Relative contributions of consonants, vowels and tones were investigated by fitting the human data into a modified power function. Weighting coefficient of consonants was the highest, indicating that the identification of consonants accounted for biggest part of difficulties in concurrent Mandarin syllable recognition.

A comparison between human listeners and a deep-learning model on the same recognition task provided insights for model enhancements.

7.2 Concurrent vowel recognition in Mandarin

The first experiment was designed to evaluate how listeners utilize spectral contrasts as a cue in concurrent Mandarin vowel recognition. The design referred to past studies (Fu, Wu, et al., 2019) as a verification. The results from this experiment can provide preliminary conclusions into preparation for extending the stimuli range to cover more common-used cases.

Most findings in this experiment was consistent with Fu and co-workers' study except for the trend of utilizing spectral contrasts information. The effect of spectral contrasts in our case was followed a logarithmic scale ($p < 0.001$) instead of linear in previous studies. The log usage of spectral difference could be read as the presentation of spectral contrasts information was especially effective when two vowels in a pair differed less spectrally. Possible explanation could be link with the theory of Steven's power law. In the experiment, a small increasement of contrasts could provide useful clues for human listeners to do the differentiation when spectral contrast of vowels in a pair was low. An increasemetn of normalized spectral contrasts at 0.3 was sufficient for accuracy of tonal vowel recognition improved over 50% for some specific vowel categories. The effect of spectral power difference might surpass other cues for vowel pairs with lower contrasts. However, when the

spectral contrasts increased to a certain degree, such as “a” and “u”, the effect of spectral energy difference became weak on improving the recognition accuracy. Possible reasons could be spectral contrasts were so large that listeners can easily use intrinsic features of concurrent vowels to recognize.

Tone recognition was evaluated in the first experiment. Consistent with past works, tone 3 had the significantly lowest accuracy (75.2%), while other three tones had accuracy all above 90%. Tones in Mandarin was featured by the distinct pitch change temporally. The lowest accuracy might because that tone 3 was the only tone had two consecutive changes of directions in pitch contour, which made it harder to be recognized. Additionally, tone 3 had the longest duration naturally, while the cue of duration was eliminated in the experiment design. As a result, to some extent, this design weakened the features of tone 3 for recognition.

7.3 Concurrent syllable recognition in Mandarin with four consonants involved

As a new topic, the second experiment was designed as a continued study of concurrent-vowel recognition by adding an initial consonant to form the syllable. In order to had a preliminary understanding of this topic. Only four consonants that either differed on place or manner of articulation were involved. Important findings in this experiment included the role of spectral contrast was still important on the recognition of concurrent consonant + vowel (CV) syllables. Similar as experiment one, the performance of recognition also had a significant correlation ($p < 0.001$) with spectral contrasts in a logarithmic scale especially for syllable pairs without vowel difference. In order to examine the effect from categories, syllable pairs were subdivided into two groups which differed on the presentation of vowel differences. When both consonant difference and vowel difference presented, by increasing the spectral contrasts, the correct recognition rate could be largely improved. However, the room for improvement was relatively limited when only consonant difference available. It could be explained by the relative duration of consonants and vowels of the syllable despite the enhancement of accuracies by increasing spectral contrasts. From definition, the spectral contrasts considered the information of overall spectral energy difference from two vowels in a concurrent pair. Initial consonants carried less energy than followed vowels, as a result, the energy change from the consonants provided finite clues for listeners to utilize.

A detailed comparison was discussed in terms of the categories of consonants selected. In this experiment, syllables consisted of consonants that were differentiated from place and manner of articulation. Start with the most used consonant “d”, another two consonants were either differed on place or manner of articulation. The last consonants had both different place and manner of articulation with “d”. Since manner of articulation accounted for larger part of spectral features of vowels, “b” and “d” that both had plosive aspirated articulation manner were easily be wrongly recognized as each other. Though pronounced from bilabial and alveolar place respectively, similar identity produced difficulty in recognition when they were presented concurrently, especially the vowel followed was identical.

In contrast, as the consonant “sh” provided pronounced as a noise-like sound due to the fricative feature. It was the easiest one to be recognized among four consonants. Besides the spectral contrasts that was largely different from other consonants, the noise-featured initial sound of a syllable with “sh” could surpass another consonant thus improved the correct recognition rate of concurrent syllables.

In summary, the increasing similarities with consonant “d” produced progressive level of difficulties on recognizing concurrent syllable pair from “sh”, “l” to “b”. Spectral contrasts indicating the spectral power differences between two syllables were shown to be beneficial for concurrent Mandarin syllable recognition as a spectral cue.

7.4 Concurrent syllable recognition in Mandarin with extended selection of consonants

The third experiment was a continued study of experiment two. Under the same design, more categories of consonants and vowels in Mandarin were involved in order to explore the concurrent CV syllable recognition performance from a more general perspective.

The effect of spectral contrasts was reinvestigated in this experiment. The results suggested a consistent role of spectral contrasts in concurrent syllable recognition with one exception. In the group of syllables where only consonants difference available, the spectral contrasts cue functioned linearly.

If the two syllables presented concurrently had different vowels, spectral contrasts could be largely utilized to improve the recognition accuracy regardless of consonant difference. This was manifested by the relatively high accuracy of recognition in the syllable grouping by

available vowel difference. However, when only consonants difference existed with absent vowel differences, relatively low accuracies for recognitions of all combinations in this group were found. The positive effect that provided by spectral contrasts was limited.

Also, the difficulties in recognizing consonants were possibly due to their intrinsic properties. The unvoiced consonants used in the experiment had shorter duration and less spectral power than vowels. Compared to experiment two, more involvement of categories of consonants created more confusions in identifying concurrent Mandarin syllable.

A detailed comparison in terms of the consonant categories showed in two cases, listeners had most difficulties in recognition. The first was that when syllables with “sh” and “z” were combined together, the performance was extremely degraded. A possible explanation could be their close locations of energy peak on the spectrum, which was the intrinsic features of two consonants. The pronunciation of two consonants produced large amount of energy in similar range of frequency. As a result, it was hard for listeners to differentiate only with the cue of spectral contrasts. Second case appeared in the mixture of consonants “b” and “d”. This was discussed in last experiment where similar confusions were observed. The results in experiment three showed that, when syllable consisting of “di” mixed with other syllables, listeners tended to wrongly recognized “d” as “b”. Interestingly, such difficulties only appeared when two consonants combined with vowel “i”. Additionally, both “b” and “d” were plosive unaspirated with similar spectral characteristic. The information provided by spectral contrasts was too rough for differentiation. As a result, it could be inferred that spectral contrasts in concurrent syllable recognition was not as effective as in concurrent vowel recognition. Other cues beneficial for recognizing consonants needed to be enhanced in order to improve the recognition accuracies of concurrent Mandarin syllables.

In this experiment, tone recognition was also assessed by a presenting of confusion matrix. The results were highly consistent with the first experiment where concurrent vowel recognition was studied. Tone 3 identification had the significantly lowest accuracy compared with other three tones. It could be due to the “falling-rising” characteristics of tone 3, which produced two directions of pitch change temporally and raised the difficulty of recognition. In unprocessed speech, syllables with tone 3 had the longest duration, while in our cases, the cue of syllable duration was eliminated by equalizing each syllable to have the

same length regardless of the tone category. This was another possible factor that producing low accuracies in recognizing tone 3.

In further analysis, a power function based model was introduced to examine the relative contributions of consonants, vowels and tones in recognition task. Accuracies were summarized in terms of these three elements separately. By model fitting, it was shown that consonants weighted more than vowels and tones in probability to recognize a syllable from concurrent presented competitor. Possible reason for explaining this could be the biggest difficulty emerged from confusions between consonants when two CV syllables presented together. In contrast, clear spectral formants and relatively higher energy made vowels easier to be recognized. From the number of categories, consonants have the most categories leading to a higher chance of error. Additionally, results showed the correction coefficient used to modify the model was smaller than one, indicating that competing syllables could mask another when presented concurrently. As a result, besides providing a larger spectral contrasts, it was important to improve recognition intelligibility of consonants for producing higher accuracy in concurrent Mandarin syllable recognition.

7.5 Concurrent syllable separation by deep-learning model and a comparison with human

The last experiment evaluated the performance of deep-learning model on separating concurrent Mandarin syllables and compared the accuracies with human listeners. Despite the error in matching the real accuracy between human and machine, the comparison had some extents of value.

According to the model output, part of the analysis was of expectation. When the model was used in separating concurrent Mandarin vowels, effect of spectral contrasts between concurrent vowels was significant ($p < 0.009$ for all three metrics) in linear scale. Larger contrasts effectively helped model produce higher sound quality for separated vowels. This was consistent with human, when the spectral contrasts increased, higher accuracies were found.

However, it was surprising that the utilization of spectral contrasts was not significantly effective for machine in separating concurrent Mandarin syllables where both vowels and consonants differences presented. Specifically, in concurrent pairs where two syllables differed on vowels and consonants, recognition accuracy of machine showed no advantage of the variation in spectral contrasts. Only in the condition when consonants differences presented with identical vowels, the separation performance was significantly enhanced by increasing of spectral contrasts. This was a reflection that machine could take advantage of the spectral contrasts in concurrent Mandarin syllable separation in the absence of vowel difference. It should be noticed that in experiment three, human listeners was found to show a linear utilization of spectral contrasts on enhancing recognition performance in the same condition that only consonants differences presented as machine. Based on this, human and machine performance showed relatively high consistency in this case.

7.6 Summary

In conclusion, a table was used to summarize the effects of spectral contrasts in different experiment conditions:

Table 7.1 Summarization on effects of spectral contrasts

| Task | Human performance | Deep-learning model |
|---------------------------------|---|--|
| Concurrent vowel recognition | Positive correlation Logarithmic scale | Positive correlation Linear scale |
| Concurrent syllable recognition | Positive correlation Logarithmic scale when both vowels and consonants differences available; Linear scale when only consonants differences available | No significant correlation when both consonants and vowels differences available; Linearly positive when only consonants difference available |

CHAPTER 8 CONCLUSIONS, LIMITATIONS AND FUTURE WORKS

8.1 Conclusions

8.1.1 Effects of spectral contrasts on concurrent-vowel recognition in Mandarin

In concurrent vowel recognition tasks, the spectral contrasts between the two concurrent paired vowels significantly affected the recognition accuracies. This suggests that listeners utilized the differences in vowel spectral contrasts as spectral cues to facilitate vowel separation and recognition. A larger contrast increased the recognition accuracies. Unlike past literature which suggested a linear relationship between recognition accuracies and spectral contrasts (Fu, Wu, et al., 2019). Present results suggested an increase in recognition accuracies with logarithmic spectral contrasts. Whilst this finding is new and not consistent with previous literature, it is consistent with the general findings of Steve's power law on psychophysics.

8.1.2 Effects of spectral contrasts on concurrent-syllable (consonant + vowel) recognition in Mandarin

Mandarin words often composed of an initial consonant followed by a vowel. Role of spectral contrasts on concurrent syllable recognition was evaluated with concurrent paired “consonant-vowel (CV)”-pattern syllables.

Four consonants differed either in place of articulation or manner of articulation were tested. Results showed that spectral contrasts of CV syllables significantly enhanced recognition. Similar to the findings with concurrent vowels, a higher accuracy of performance could be observed with increasing spectral contrasts in a logarithmic scale.

When both vowel and consonant differences were present, accuracies in concurrent CV syllable recognition were high (the highest reached 87.2%), while performance degraded (the highest was 65.7%) when only consonant difference was present (vowels in a syllable pair were the same).

Further testings with more categories of consonants and vowels confirmed the significant influence of spectral contrasts to concurrent syllable recognition. Nonetheless, the effect was weakened due to the limited reflection of consonant differences in spectral contrasts value, which indicated that the recognition of consonants relied on more cues such as location of energy peak in the spectrum, temporal envelope cues.

8.1.3 Relative contributions of consonant, vowels, and tones on concurrent syllable recognition with Mandarin

Based on the findings of experiments two and three, recognition of concurrent consonants showed significantly lower accuracy compared with recognition concurrent vowel and tone. As most consonants in Mandarin were unvoiced (all consonants used in the experiment were unvoiced), a possible explanation was that low sound energy, absent from unvoiced consonants, could be utilized by listeners to differentiate and recognize concurrent syllables.

In addition to the experiments, a model using power function was fitted to quantitatively describe the relationship among consonant, vowel, tone and syllable recognition performance. Output of the model showed that the weighting coefficients of consonant, vowel and tone were 0.86, 0.48, 0.74 respectively, which revealed that consonants accounted for the biggest variance in syllable recognition accuracies. Thus, improvement of consonant intelligibility when competing syllable existed helped listener correctly recognize the target information to a greater extent.

8.1.4 Comparison of recognition accuracy between human listeners and deep-learning model

The concurrent syllable recognition results acquired from human listeners were compared with the outputs of a deep-learning model trained to separate speech. Assessment of model performance was from three metrics including short-time objective intelligibility (STOI), perceptual evaluation of speech quality (PESQ) and SDR (signal to distortion ratio).

Results of comparison revealed that in concurrent vowel recognition, model presented an improvement of recognition with increasing spectral contrasts as evidenced in all three performance metric. Statistics showed the recognition performance of model was linearly

correlated ($p < 0.009$ for all three metrics) with spectral contrasts of concurrent vowels, which was different from nonlinear (logarithmic) correlations revealed in human listeners' data.

However, similar effect of spectral contrasts was found in concurrent “consonant + vowel” (CV) syllable recognition for model when only consonants differences available. Despite human could benefit from cues provided by spectral difference, model showed flat performance in stimuli pairs with both consonants and vowels differences regardless of the amount of spectral contrasts. This phenomenon might have originated from the large amount of information when both differences available. The spectral contrasts became redundant for model to recognize. In contrast, when only consonants differences available, spectral power changes could be captured and utilized to improve separation intelligibility by deep-learning model. As a result, enlarging the spectral contrasts between competing speech sources when they have very similar vowel features, a preemphasis of consonants in training might be two possible ways to improve model accuracy in concurrent mandarin speech separation. Future studies could be done to evaluate the performance of revised model.

8.2 Limitations and Future Work

In our study, the effects of mean F0 difference, duration and sound intensity were controlled. Nevertheless, the independent variable, spectral contrast, was a rough measurement of spectral information. The spectral cue could be further subdivided considering the peak of formants, the feature of spectrum etc. Temporal cues, which were proved to be important in concurrent vowel recognition was also not

In terms of the concurrent Mandarin syllable recognition, the selection of consonants and vowels in stimuli didn't cover the whole categories in Mandarin due to that the large number of combinations were difficult for listeners to recognize in one-time experiment. Future work could be done to optimize the testing procedure so that more accurate findings could be explored.

The logarithmic effect of spectral contrasts in our study was consistent with past works. Since the perception and recognition of concurrent speech was a complex result from collective effect of numerous systems involving auditory information processing and language

processing etc., designs that examine the mechanism for human processing the spectral contrasts information are needed as evidence to understand and explain this finding.

The comparison between human and deep-learning models provided possible suggestions on improving machine performance. Further studies are needed to exploit the iteration of model or modifying training method correspondingly to verify the effectiveness of such adaptations.

REFERENCES AND BIBLIOGRAPHY

- Arehart, K. H., King, C. A., & McLean-Mudgett, K. S. (1997). Role of fundamental frequency differences in the perceptual separation of competing vowel sounds by listeners with normal hearing and listeners with hearing loss. *Journal of Speech, Language, and Hearing Research*, 40(6), 1434-1444.
- Arehart, K. H., Rossi-Katz, J., & Swensson-Prutsman, J. (2005). Double-Vowel Perception in Listeners With Cochlear Hearing Loss. *Journal of Speech, Language, and Hearing Research*
- Arehart, K. H., Souza, P. E., Muralimanohar, R. K., & Miller, C. W. (2011). Effects of age on concurrent vowel perception in acoustic and simulated electroacoustic hearing. *Journal of Speech, Language, and Hearing Research*
- Assmann, P. F., & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *The Journal of the Acoustical Society of America*, 88(2), 680-697.
- Boersma, P. (2006). Praat: doing phonetics by computer. <http://www.praat.org/>
- Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 109(3), 1101-1109.
- Chen, F., Wong, L. L., & Hu, Y. (2014). Effects of lexical tone contour on Mandarin sentence intelligibility. *Journal of Speech, Language, and Hearing Research*
- Chen, F., Wong, M. L., Zhu, S., & Wong, L. L. (2015). Relative contributions of vowels and consonants in recognizing isolated Mandarin words. *Journal of Phonetics*, 52, 26-34.
- Chintanpalli, A., Ahlstrom, J. B., & Dubno, J. R. (2016). Effects of age and hearing loss on concurrent vowel identification. *The Journal of the Acoustical Society of America*, 140(6), 4142-4153.
- Chintanpalli, A., & Heinz, M. G. (2013). The use of confusion patterns to evaluate the neural basis for concurrent vowel identification. *The Journal of the Acoustical Society of America*, 134(4), 2988-3000.
- Cole, R. A., Yan, Y., Mak, B., Fanty, M., & Bailey, T. (1996). The contribution of consonants versus vowels to word recognition in fluent speech. 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings,
- Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, 119(3), 1562-1573.
- Cruttenden, A. (2014). *Gimson's pronunciation of English*. Routledge.

- Culling, J. F., & Summerfield, Q. (1995). The role of frequency modulation in the perceptual segregation of concurrent vowels. *The Journal of the Acoustical Society of America*, 98(2), 837-846.
- Darwin, C. J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time. *The Quarterly Journal of Experimental Psychology Section A*, 33(2), 185-207.
- Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 114(5), 2913-2922.
- De Jong, K. (2003). Temporal constraints and characterising syllable structuring. In *Phonetic interpretation. Papers in laboratory phonology VI* (pp. 253-268). Cambridge University Press Cambridge.
- Dow, F. D. (1972). *An outline of Mandarin phonetics*. Australian National University Press.
- Duanmu, S. (2007). *The phonology of standard Chinese*. OUP Oxford.
- Fletcher, H. (1995). The speaking mechanism. *Speech and Hearing in Communication*, 16.
- Fogerty, D., & Humes, L. E. (2010). Perceptual contributions to monosyllabic word intelligibility: Segmental, lexical, and noise replacement factors. *The Journal of the Acoustical Society of America*, 128(5), 3114-3125.
- Fogerty, D., & Humes, L. E. (2012). The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. *The Journal of the Acoustical Society of America*, 131(2), 1490-1501.
- Fogerty, D., & Kewley-Port, D. (2009). Perceptual contributions of the consonant-vowel boundary to sentence intelligibility. *The Journal of the Acoustical Society of America*, 126(2), 847-857.
- Fu, Q.-J., & Zeng, F.-G. (2000). Identification of temporal envelope cues in Chinese tone recognition. *Asia Pacific Journal of Speech, Language and Hearing*, 5(1), 45-57.
- Fu, Q.-J., Zeng, F.-G., Shannon, R. V., & Soli, S. D. (1998). Importance of tonal envelope cues in Chinese speech recognition. *The Journal of the Acoustical Society of America*, 104(1), 505-510.
- Fu, Z., Wu, X., & Chen, J. (2019). Effects of Spectral and Temporal Cues to Mandarin Concurrent-Vowels Identification for Normal-Hearing and Hearing-Impaired Listeners. *INTERSPEECH*,
- Fu, Z., Yang, H., Chen, F., Wu, X., & Chen, J. (2019). Brainstem encoding of frequency-modulated sweeps is relevant to Mandarin concurrent-vowels identification for normal-hearing and hearing-impaired listeners. *Hearing research*, 380, 123-136.
- Fu, Z., Yang, H., Wu, X., & Chen, J. (2018). Acoustic cues utilized by normal-hearing and hearing-impaired listeners are different for mandarin concurrent-vowels identification. *Acta Acustica united with Acustica*, 104(5), 792-795.

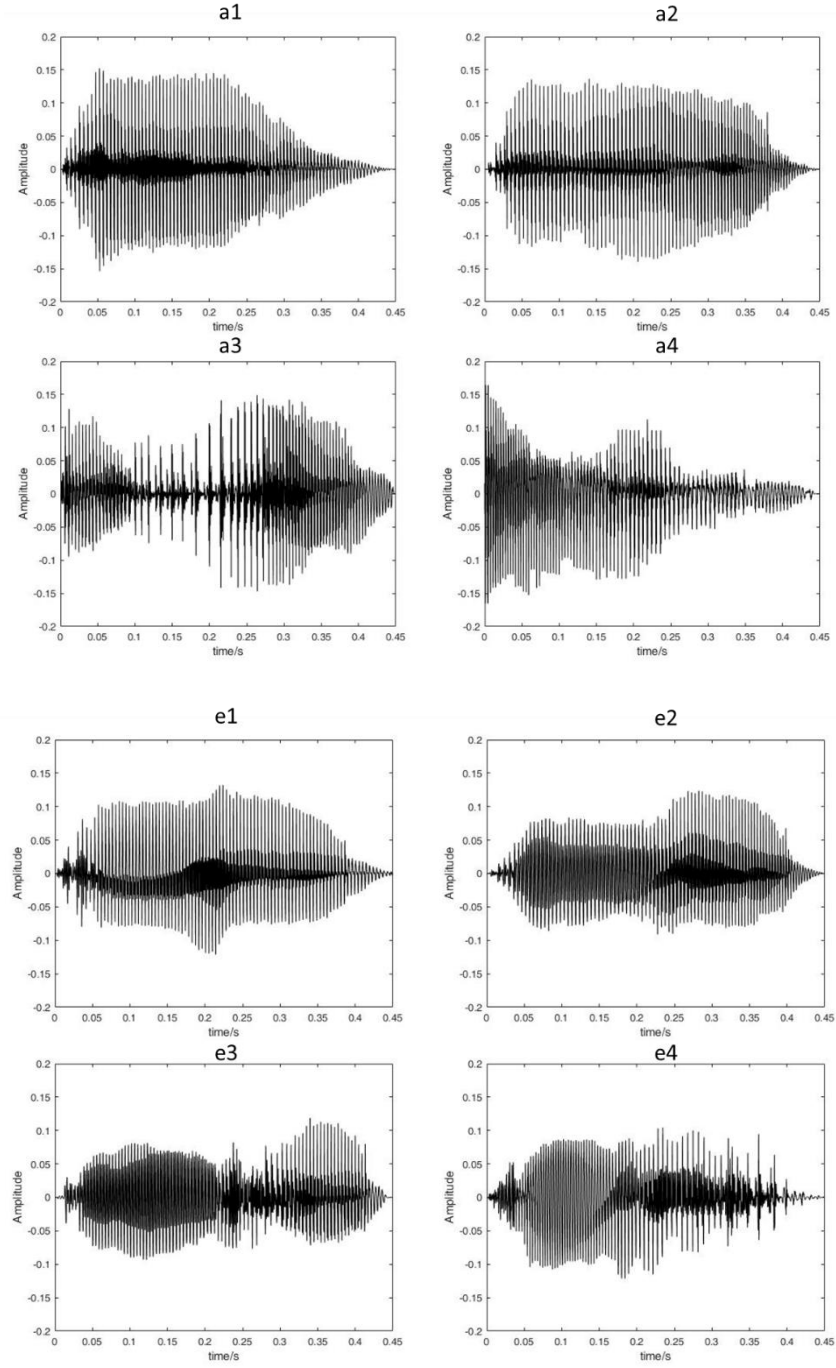
- Hedrick, M. S., & Madix, S. G. (2009). Effect of vowel identity and onset asynchrony on concurrent vowel identification. *Journal of Speech, Language, and Hearing Research*.
- Hee Lee, J., & Humes, L. E. (2012). Effect of fundamental-frequency and sentence-onset differences on speech-identification performance of young and older adults in a competing-talker background. *The Journal of the Acoustical Society of America*, 132(3), 1700-1717.
- Kewley-Port, D., Burkle, T. Z., & Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 122(4), 2365-2375.
- Ladefoged, P., & Johnson, K. (2014). *A course in phonetics*. Nelson Education.
- Li, X., Ji, H., & Wang, T. (2013). 《通用规范汉字表》使用手册. 人民出版社.
<https://books.google.com.hk/books?id=MckCoQEACAAJ>
- Luo, X., & Fu, Q.-J. (2009). Concurrent-vowel and tone recognitions in acoustic and simulated electric hearing. *The Journal of the Acoustical Society of America*, 125(5), 3223-3233.
- Luo, Y., & Mesgarani, N. (2018). Tasnet: time-domain audio separation network for real-time, single-channel speech separation. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),
- Luo, Y., & Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8), 1256-1266.
- McAdams, S. (1989). Segregation of concurrent sounds. I: Effects of frequency modulation coherence. *The Journal of the Acoustical Society of America*, 86(6), 2148-2159.
- Meddis, R., & Hewitt, M. J. (1992). Modeling the identification of concurrent vowels with different fundamental frequencies. *The Journal of the Acoustical Society of America*, 91(1), 233-245.
- Micheyl, C., & Oxenham, A. J. (2010). Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings. *Hearing research*, 266(1-2), 36-51.
- Qin, M. K., & Oxenham, A. J. (2005). Effects of envelope-vocoder processing on F0 discrimination and concurrent-vowel identification. *Ear hearing*, 26(5), 451-460.
- Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221),

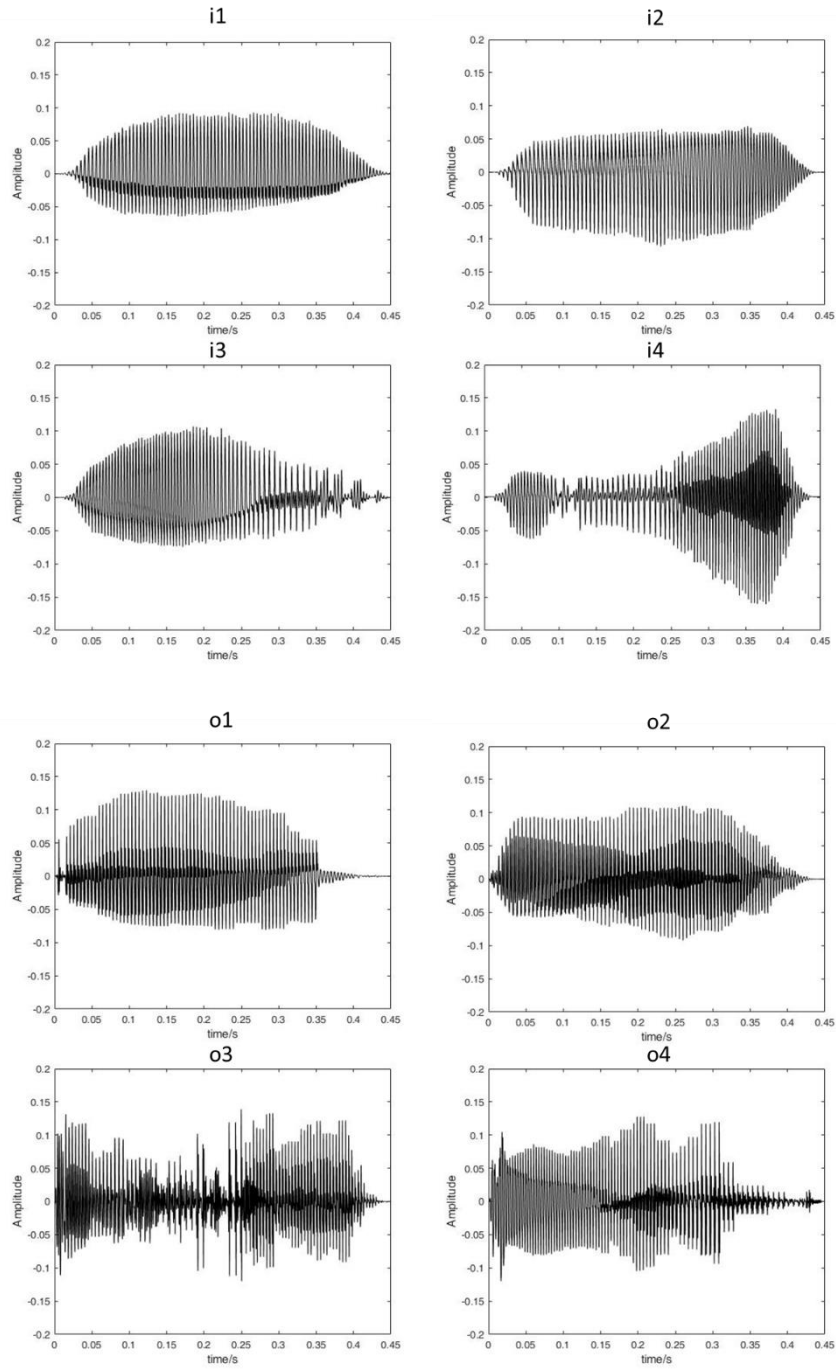
- Roberts, B., & Holmes, S. D. (2006). Grouping and the pitch of a mistuned fundamental component: Effects of applying simultaneous multiple mistunings to the other harmonics. *Hearing research*, 222(1-2), 79-88.
- Sagart, L. (1999). The origin of Chinese tones.
- Shackleton, T. M., & Meddis, R. (1992). The role of interaural time difference and fundamental frequency difference in the identification of concurrent vowel pairs. *The Journal of the Acoustical Society of America*, 91(6), 3579-3581.
- Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270(5234), 303-304.
- Snyder, J. S., & Alain, C. (2005). Age-related changes in neural activity associated with concurrent vowel segregation. *Cognitive Brain Research*, 24(3), 492-499.
- Song, Z. (2013). *MATLAB 在语音信号分析与合成中的应用*. 北京航空航天大学出版社. <https://books.google.com.hk/books?id=kqLzsgEACAAJ>
- Stoller, D., Ewert, S., & Dixon, S. (2018). Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1803.03185*.
- Studebaker, G. A. (1985). A "rationalized" arcsine transform. *Journal of Speech, Language, and Hearing Research*, 28(3), 455-462.
- Summers, V., & Leek, M. R. (1998). F0 processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss. *Journal of Speech, Language, and Hearing Research*, 41(6), 1294-1306.
- Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. 2010 IEEE international conference on acoustics, speech and signal processing,
- Tsu-Lin, M. (1970). Tones and prosody in Middle Chinese and the origin of the rising tone. *Harvard Journal of Asiatic Studies*, 30, 86-110.
- Vestergaard, M. D., Fyson, N. R., & Patterson, R. D. (2009). The interaction of vocal characteristics and audibility in the recognition of concurrent syllables. *The Journal of the Acoustical Society of America*, 125(2), 1114-1124.
- Vongpaisal, T., & Pichora-Fuller, M. K. (2007). Effect of age on F0 difference limen and concurrent vowel identification. *J Journal of Speech, Language, and Hearing Research*
- Wang, D., & Zhang, X. (2015). Thchs-30: A free chinese speech corpus. *arXiv preprint arXiv:1501.01882*.
- Xu, L., & Pfingst, B. E. (2003). Relative importance of temporal envelope and fine structure in lexical-tone perception (L). *The Journal of the Acoustical Society of America*, 114(6), 3024-3027.

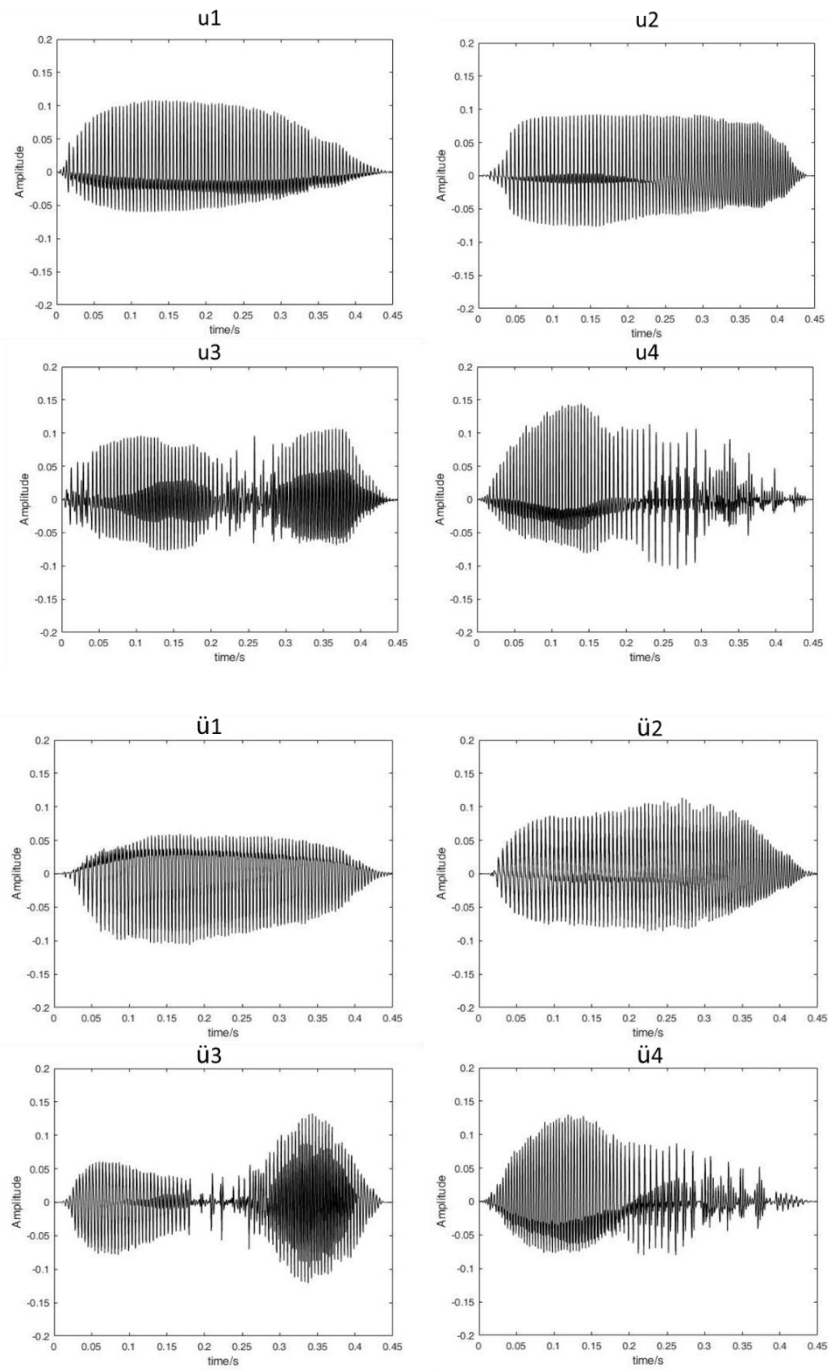
- Xu, L., Thompson, C. S., & Pfingst, B. E. (2005). Relative contributions of spectral and temporal cues for phoneme recognition. *The Journal of the Acoustical Society of America*, 117(5), 3255-3267.
- Yu, D., Kolbæk, M., Tan, Z.-H., & Jensen, J. (2017). Permutation invariant training of deep models for speaker-independent multi-talker speech separation. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),
- 教育委员会, 中., 信息工程学院, 中., & 北京恩维特声像技术中心. (2013). *中级音响师速成实用教程*. 人民邮电出版社.
- <https://books.google.com.hk/books?id=XiXioAEACAAJ>

Appendix

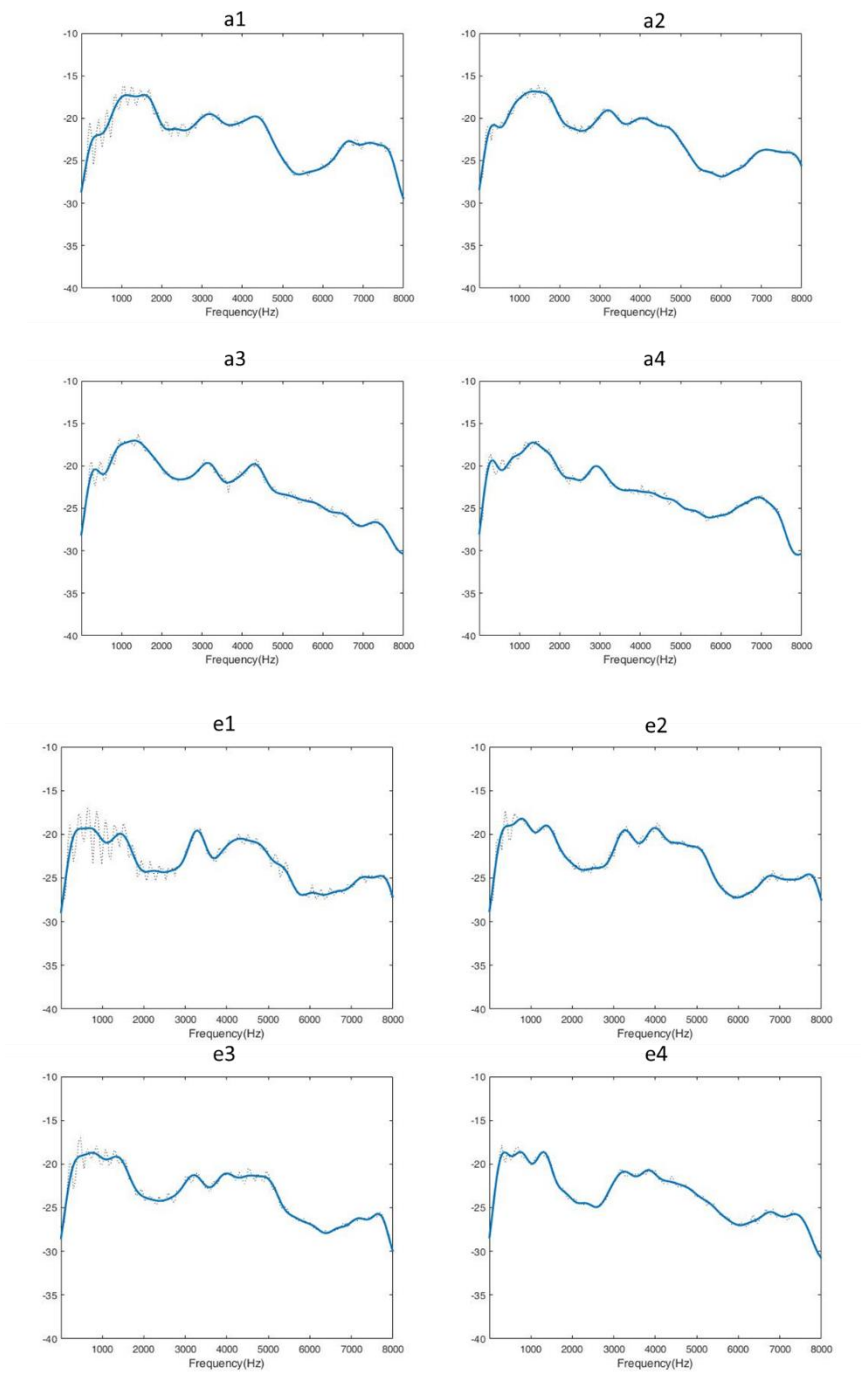
Appendix 3.1 Waveplot of single stimuli used in experiment one

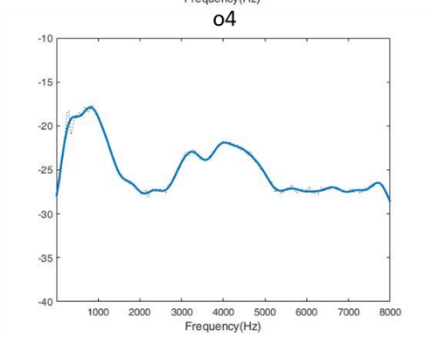
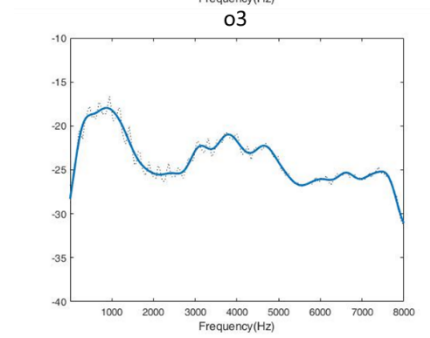
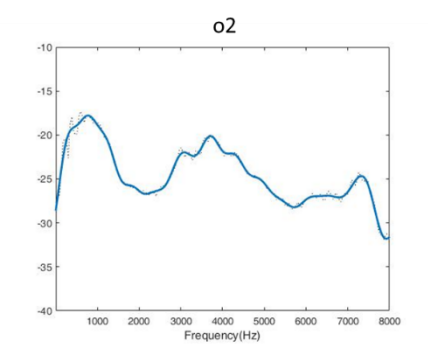
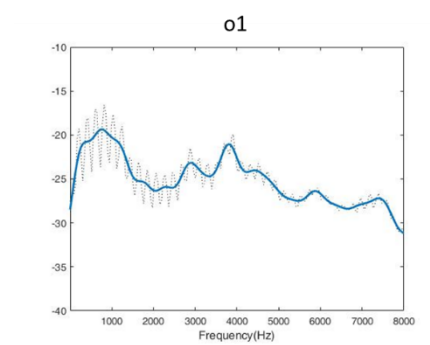
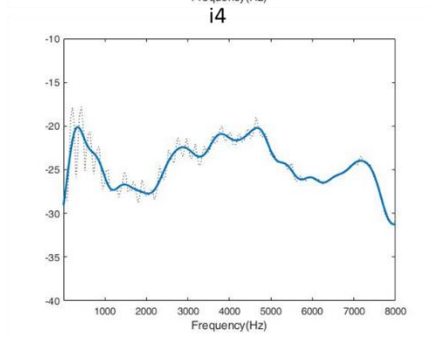
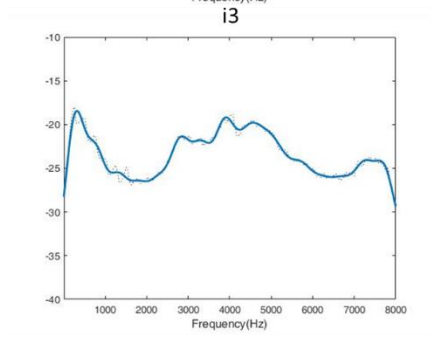
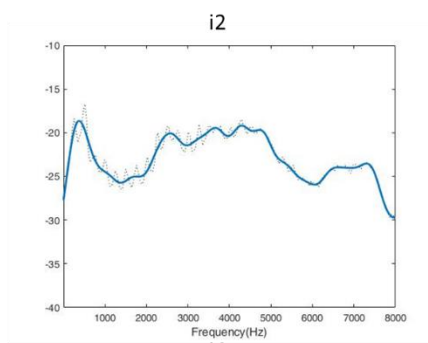
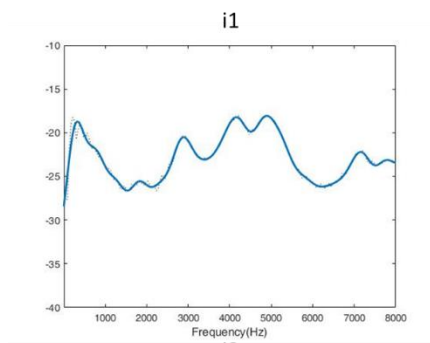


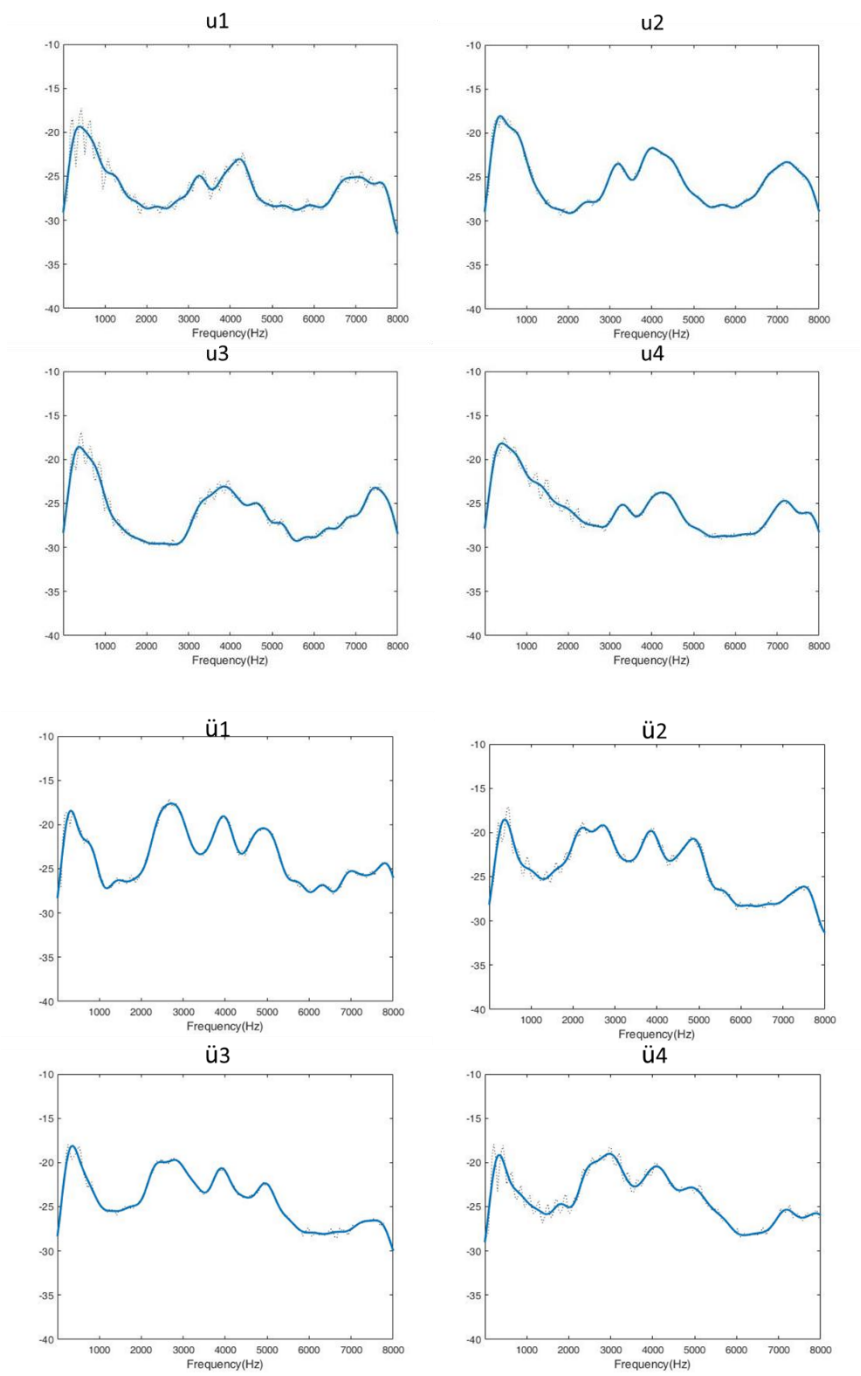




Appendix 3.2 Spectrum and spectral envelope of single stimuli used in experiment one







Appendix 4.1 Usage frequency table of consonants in mandarin

| | words | Percentage | Frequency | Percentage | Cumulated frequency | Cumulated percentage |
|----|-------|------------|-----------|------------|---------------------|----------------------|
| d | 253 | 4.4146 | 153636 | 7.7481 | 436975 | 22.037 |
| sh | 268 | 4.6763 | 153259 | 7.7291 | 590234 | 29.766 |
| j | 461 | 8.044 | 136705 | 6.8943 | 726939 | 36.66 |
| x | 352 | 6.142 | 126714 | 6.3904 | 853653 | 43.051 |
| zh | 350 | 6.1071 | 113581 | 5.7281 | 967234 | 48.779 |
| l | 398 | 6.9447 | 106013 | 5.3464 | 1073247 | 54.125 |
| b | 255 | 4.4495 | 94703 | 4.776 | 1167950 | 58.901 |
| g | 237 | 4.1354 | 93177 | 4.6991 | 1261127 | 63.6 |
| h | 280 | 4.8857 | 88569 | 4.4667 | 1349696 | 68.067 |
| t | 235 | 4.1005 | 79405 | 4.0045 | 1429101 | 72.072 |
| m | 239 | 4.1703 | 75260 | 3.7955 | 1504361 | 75.867 |
| q | 256 | 4.4669 | 70680 | 3.5645 | 1575041 | 79.432 |
| ch | 238 | 4.1529 | 66819 | 3.3698 | 1641860 | 82.801 |
| z | 136 | 2.3731 | 66130 | 3.335 | 1707990 | 86.136 |
| f | 182 | 3.1757 | 54653 | 2.7562 | 1762643 | 88.893 |
| r | 84 | 1.4657 | 47593 | 2.4002 | 1810236 | 91.293 |
| n | 130 | 2.2684 | 44937 | 2.2662 | 1855173 | 93.559 |
| k | 154 | 2.6871 | 42515 | 2.1441 | 1897688 | 95.703 |
| s | 137 | 2.3905 | 32656 | 1.6469 | 1930344 | 97.35 |
| p | 191 | 3.3328 | 27324 | 1.378 | 1957668 | 98.728 |
| c | 110 | 1.9194 | 25214 | 1.2716 | 1982882 | 100 |

Appendix 4.2 Normalized spectral contrasts of stimuli in experiment 2

| combination | Normalized spectral contrast |
|-------------|---------------------------------|
| shabu | 1 |
| bushu | 0.801007 |
| shadu | 0.694346 |
| shalu | 0.687474 |
| lashu | 0.631875 |
| lushu | 0.618368 |
| dushu | 0.615229 |
| dashu | 0.504158 |
| bashu | 0.478663 |
| dabu | 0.439197 |
| babu | 0.399378 |
| basha | 0.39067 |
| labu | 0.381555 |
| dalu | 0.345716 |
| lasha | 0.340021 |
| balu | 0.315315 |
| dasha | 0.314537 |
| ladu | 0.311945 |
| lalu | 0.296772 |
| dadu | 0.263511 |
| badu | 0.234439 |
| shashu | 0.157141 |

| | |
|--------|----------|
| bulu | 0.03376 |
| budu | 0.031037 |
| bubu | 0.016431 |
| dulu | 0.014744 |
| baba | 0.009401 |
| dudu | 0.008755 |
| lulu | 0.007139 |
| shasha | 0.006626 |
| bada | 0.006379 |
| shushu | 0.006054 |
| bala | 0.005449 |
| dada | 0.005174 |
| dala | 0.002791 |
| lala | 0 |

Appendix 4.3 Post-hoc analysis results of interactions between factors

Sidak Pairwise Comparisons: consonant

Grouping Information Using the Sidak Method and 95% Confidence

| consonant | N | Mean | Grouping |
|-----------|-----|---------|----------|
| sh | 128 | 84.6207 | A |
| b | 128 | 66.7452 | B |
| l | 128 | 65.7090 | B |
| d | 128 | 63.2810 | B |

Means that do not share a letter are significantly different.

Sidak Simultaneous Tests for Differences of Means

| Difference of consonant Levels | Difference of Means | SE of Difference | Simultaneous 95% CI | T-Value | Adjusted P-Value |
|--------------------------------|---------------------|------------------|---------------------|---------|------------------|
| d - b | -3.46 | 1.87 | (-8.41, 1.49) | -1.85 | 0.332 |
| l - b | -1.04 | 1.87 | (-5.99, 3.91) | -0.55 | 0.995 |
| sh - b | 17.88 | 1.87 | (12.93, 22.83) | 9.54 | 0.000 |
| l - d | 2.43 | 1.87 | (-2.52, 7.38) | 1.30 | 0.729 |
| sh - d | 21.34 | 1.87 | (16.39, 26.29) | 11.39 | 0.000 |
| sh - l | 18.91 | 1.87 | (13.96, 23.86) | 10.09 | 0.000 |

Individual confidence level = 99.15%

Sidak Pairwise Comparisons: vowel

Grouping Information Using the Sidak Method and 95% Confidence

| vowel | N | Mean | Grouping |
|-------|-----|---------|----------|
| a | 256 | 76.1390 | A |
| u | 256 | 64.0390 | B |

Means that do not share a letter are significantly different.

Sidak Simultaneous Tests for Differences of Means

| Difference of vowel Levels | Difference of Means | SE of Difference | Simultaneous 95% CI | T-Value | Adjusted P-Value |
|----------------------------|---------------------|------------------|---------------------|---------|------------------|
| u - a | -12.10 | 1.32 | (-14.70, -9.50) | -9.13 | 0.000 |

Individual confidence level = 95.00%

Sidak Pairwise Comparisons: tone

Grouping Information Using the Sidak Method and 95% Confidence

| tone | N | Mean | Grouping |
|------|-----|---------|----------|
| 4 | 128 | 77.1273 | A |
| 2 | 128 | 72.5445 | A |
| 1 | 128 | 67.5121 | B |
| 3 | 128 | 63.1720 | B |

Means that do not share a letter are significantly different.

Sidak Simultaneous Tests for Differences of Means

| Difference of tone Levels | Difference of Means | SE of Difference | Simultaneous 95% CI | T-Value | Adjusted P-Value |
|---------------------------|---------------------|------------------|---------------------|---------|------------------|
| 2 - 1 | 5.03 | 1.87 | (0.08, 9.98) | 2.69 | 0.044 |
| 3 - 1 | -4.34 | 1.87 | (-9.29, 0.61) | -2.32 | 0.119 |
| 4 - 1 | 9.62 | 1.87 | (4.67, 14.56) | 5.13 | 0.000 |
| 3 - 2 | -9.37 | 1.87 | (-14.32, -4.42) | -5.00 | 0.000 |
| 4 - 2 | 4.58 | 1.87 | (-0.37, 9.53) | 2.45 | 0.086 |
| 4 - 3 | 13.96 | 1.87 | (9.01, 18.90) | 7.45 | 0.000 |

Individual confidence level = 99.15%

Sidak Pairwise Comparisons: consonant*vowel

Grouping Information Using the Sidak Method and 95% Confidence

| consonant*vowel | N | Mean | Grouping |
|-----------------|----|---------|----------|
| sh u | 64 | 85.1564 | A |
| sh a | 64 | 84.0849 | A |
| d a | 64 | 77.2927 | A B |
| l a | 64 | 74.7066 | B C |
| b a | 64 | 68.4716 | C D |
| b u | 64 | 65.0189 | D |
| l u | 64 | 56.7113 | E |
| d u | 64 | 49.2693 | E |

Means that do not share a letter are significantly different.

Sidak Simultaneous Tests for Differences of Means

| Difference of consonant*vowel Levels | Difference of Means | SE of Difference | Simultaneous 95% CI | T-Value | Adjusted P-Value |
|--|------------------------|---------------------|------------------------|---------|---------------------|
| (b u) - (b a) | -3.45 | 2.65 | (-11.76, 4.85) | -1.30 | 0.998 |
| (d a) - (b a) | 8.82 | 2.65 | (0.52, 17.12) | 3.33 | 0.026 |
| (d u) - (b a) | -19.20 | 2.65 | (-27.51, -10.90) | -7.25 | 0.000 |
| (l a) - (b a) | 6.24 | 2.65 | (-2.07, 14.54) | 2.35 | 0.416 |
| (l u) - (b a) | -11.76 | 2.65 | (-20.06, -3.46) | -4.44 | 0.000 |
| (sh a) - (b a) | 15.61 | 2.65 | (7.31, 23.92) | 5.89 | 0.000 |
| (sh u) - (b a) | 16.68 | 2.65 | (8.38, 24.99) | 6.30 | 0.000 |
| (d a) - (b u) | 12.27 | 2.65 | (3.97, 20.58) | 4.63 | 0.000 |
| (d u) - (b u) | -15.75 | 2.65 | (-24.05, -7.45) | -5.94 | 0.000 |
| (l a) - (b u) | 9.69 | 2.65 | (1.38, 17.99) | 3.66 | 0.008 |
| (l u) - (b u) | -8.31 | 2.65 | (-16.61, -0.00) | -3.14 | 0.050 |
| (sh a) - (b u) | 19.07 | 2.65 | (10.76, 27.37) | 7.20 | 0.000 |
| (sh u) - (b u) | 20.14 | 2.65 | (11.83, 28.44) | 7.60 | 0.000 |
| (d u) - (d a) | -28.02 | 2.65 | (-36.33, -19.72) | -10.58 | 0.000 |
| (l a) - (d a) | -2.59 | 2.65 | (-10.89, 5.72) | -0.98 | 1.000 |
| (l u) - (d a) | -20.58 | 2.65 | (-28.89, -12.28) | -7.77 | 0.000 |
| (sh a) - (d a) | 6.79 | 2.65 | (-1.51, 15.10) | 2.56 | 0.259 |
| (sh u) - (d a) | 7.86 | 2.65 | (-0.44, 16.17) | 2.97 | 0.084 |
| (l a) - (d u) | 25.44 | 2.65 | (17.13, 33.74) | 9.60 | 0.000 |
| (l u) - (d u) | 7.44 | 2.65 | (-0.86, 15.75) | 2.81 | 0.135 |
| (sh a) - (d u) | 34.82 | 2.65 | (26.51, 43.12) | 13.14 | 0.000 |
| (sh u) - (d u) | 35.89 | 2.65 | (27.58, 44.19) | 13.55 | 0.000 |
| (l u) - (l a) | -18.00 | 2.65 | (-26.30, -9.69) | -6.79 | 0.000 |
| (sh a) - (l a) | 9.38 | 2.65 | (1.07, 17.68) | 3.54 | 0.012 |
| (sh u) - (l a) | 10.45 | 2.65 | (2.15, 18.75) | 3.94 | 0.003 |
| (sh a) - (l u) | 27.37 | 2.65 | (19.07, 35.68) | 10.33 | 0.000 |
| (sh u) - (l u) | 28.45 | 2.65 | (20.14, 36.75) | 10.74 | 0.000 |
| (sh u) - (sh a) | 1.07 | 2.65 | (-7.23, 9.38) | 0.40 | 1.000 |

Individual confidence level = 99.82%

Sidak Pairwise Comparisons: vowel*tone

Grouping Information Using the Sidak Method and 95% Confidence

| vowel*tone | N | Mean | Grouping |
|------------|----|---------|----------|
| a 4 | 64 | 85.0068 | A |
| a 2 | 64 | 82.5145 | A |
| a 1 | 64 | 70.7667 | B |
| u 4 | 64 | 69.2477 | B |
| a 3 | 64 | 66.2678 | B C |
| u 1 | 64 | 64.2575 | B C |
| u 2 | 64 | 62.5744 | B C |
| u 3 | 64 | 60.0763 | C |

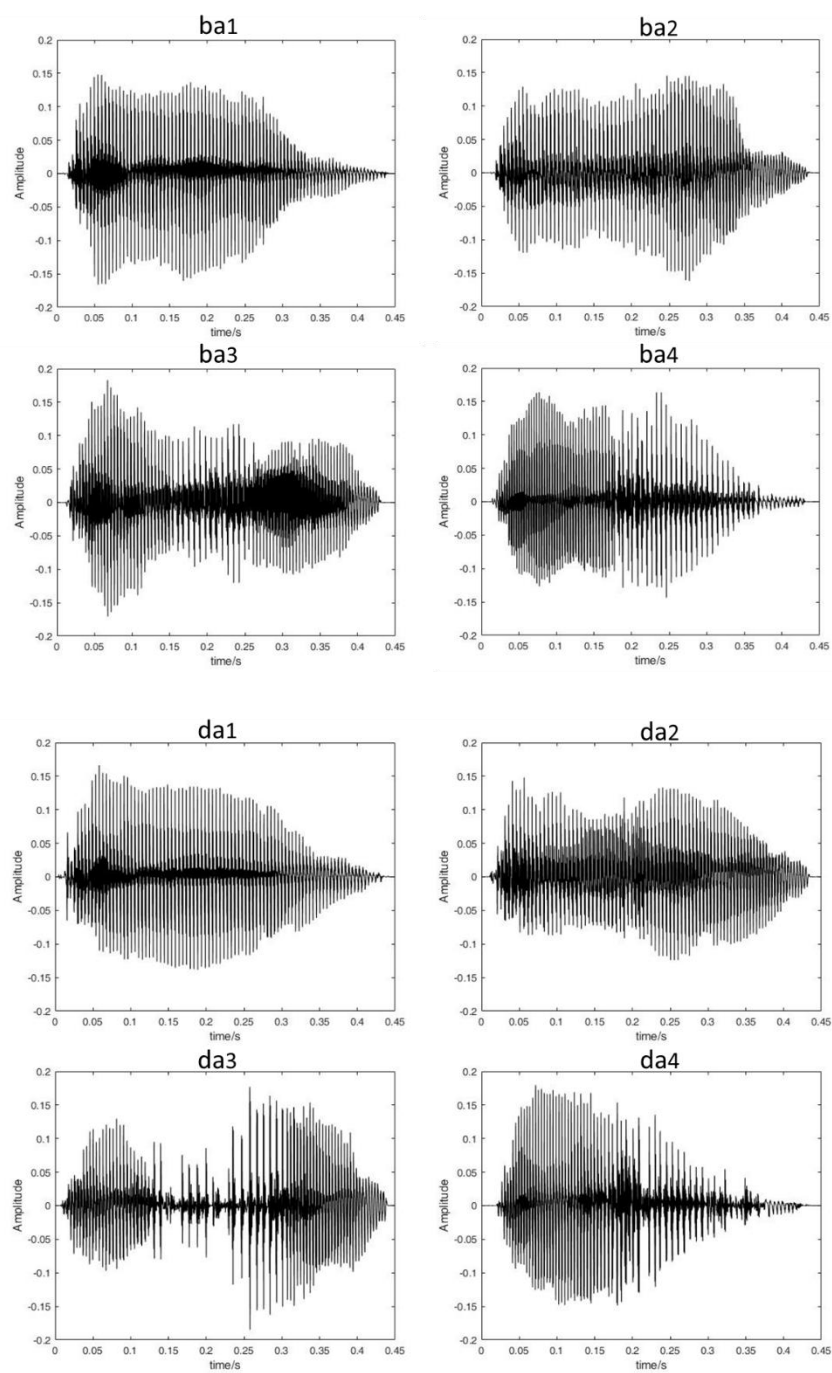
Means that do not share a letter are significantly different.

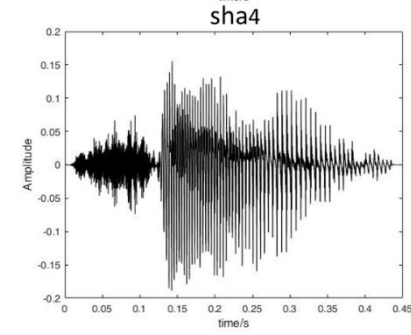
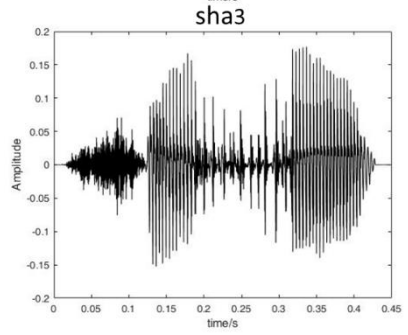
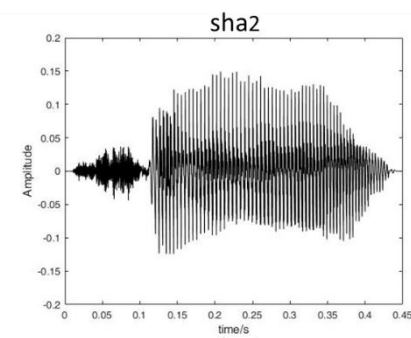
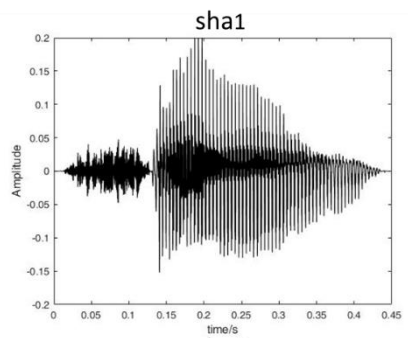
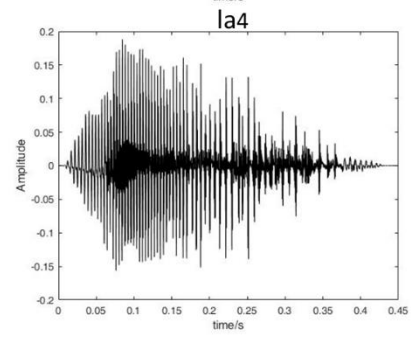
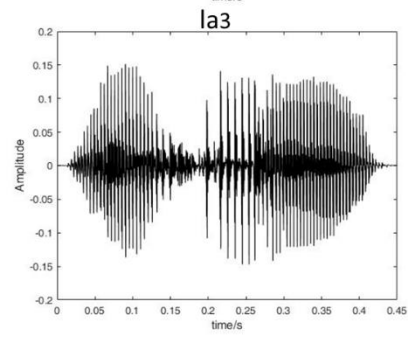
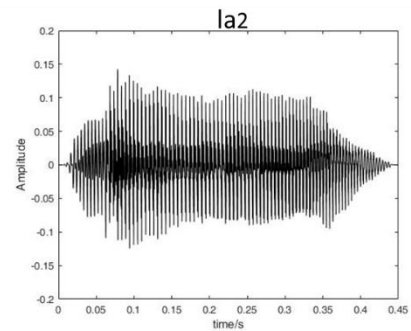
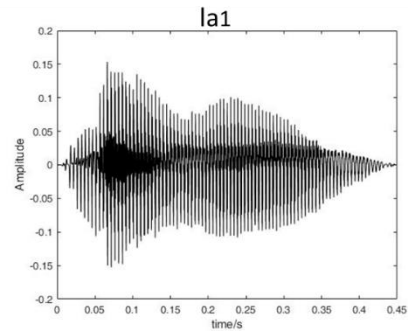
Sidak Simultaneous Tests for Differences of Means

| Difference of vowel*tone Levels | Difference of Means | SE of Difference | Simultaneous 95% CI | T-Value | Adjusted P-Value |
|---------------------------------|---------------------|------------------|---------------------|---------|------------------|
| (a 2) - (a 1) | 11.75 | 2.65 | (3.44, 20.05) | 4.43 | 0.000 |
| (a 3) - (a 1) | -4.50 | 2.65 | (-12.80, 3.80) | -1.70 | 0.929 |
| (a 4) - (a 1) | 14.24 | 2.65 | (5.94, 22.54) | 5.37 | 0.000 |
| (u 1) - (a 1) | -6.51 | 2.65 | (-14.81, 1.79) | -2.46 | 0.333 |
| (u 2) - (a 1) | -8.19 | 2.65 | (-16.50, 0.11) | -3.09 | 0.057 |
| (u 3) - (a 1) | -10.69 | 2.65 | (-18.99, -2.39) | -4.04 | 0.002 |
| (u 4) - (a 1) | -1.52 | 2.65 | (-9.82, 6.78) | -0.57 | 1.000 |
| (a 3) - (a 2) | -16.25 | 2.65 | (-24.55, -7.94) | -6.13 | 0.000 |
| (a 4) - (a 2) | 2.49 | 2.65 | (-5.81, 10.80) | 0.94 | 1.000 |
| (u 1) - (a 2) | -18.26 | 2.65 | (-26.56, -9.95) | -6.89 | 0.000 |
| (u 2) - (a 2) | -19.94 | 2.65 | (-28.24, -11.64) | -7.53 | 0.000 |
| (u 3) - (a 2) | -22.44 | 2.65 | (-30.74, -14.13) | -8.47 | 0.000 |
| (u 4) - (a 2) | -13.27 | 2.65 | (-21.57, -4.96) | -5.01 | 0.000 |
| (a 4) - (a 3) | 18.74 | 2.65 | (10.44, 27.04) | 7.07 | 0.000 |
| (u 1) - (a 3) | -2.01 | 2.65 | (-10.31, 6.29) | -0.76 | 1.000 |
| (u 2) - (a 3) | -3.69 | 2.65 | (-12.00, 4.61) | -1.39 | 0.993 |
| (u 3) - (a 3) | -6.19 | 2.65 | (-14.50, 2.11) | -2.34 | 0.430 |
| (u 4) - (a 3) | 2.98 | 2.65 | (-5.32, 11.28) | 1.12 | 1.000 |
| (u 1) - (a 4) | -20.75 | 2.65 | (-29.05, -12.45) | -7.83 | 0.000 |
| (u 2) - (a 4) | -22.43 | 2.65 | (-30.74, -14.13) | -8.47 | 0.000 |
| (u 3) - (a 4) | -24.93 | 2.65 | (-33.23, -16.63) | -9.41 | 0.000 |
| (u 4) - (a 4) | -15.76 | 2.65 | (-24.06, -7.46) | -5.95 | 0.000 |
| (u 2) - (u 1) | -1.68 | 2.65 | (-9.99, 6.62) | -0.64 | 1.000 |
| (u 3) - (u 1) | -4.18 | 2.65 | (-12.48, 4.12) | -1.58 | 0.968 |
| (u 4) - (u 1) | 4.99 | 2.65 | (-3.31, 13.29) | 1.88 | 0.824 |
| (u 3) - (u 2) | -2.50 | 2.65 | (-10.80, 5.81) | -0.94 | 1.000 |
| (u 4) - (u 2) | 6.67 | 2.65 | (-1.63, 14.98) | 2.52 | 0.289 |
| (u 4) - (u 3) | 9.17 | 2.65 | (0.87, 17.48) | 3.46 | 0.016 |

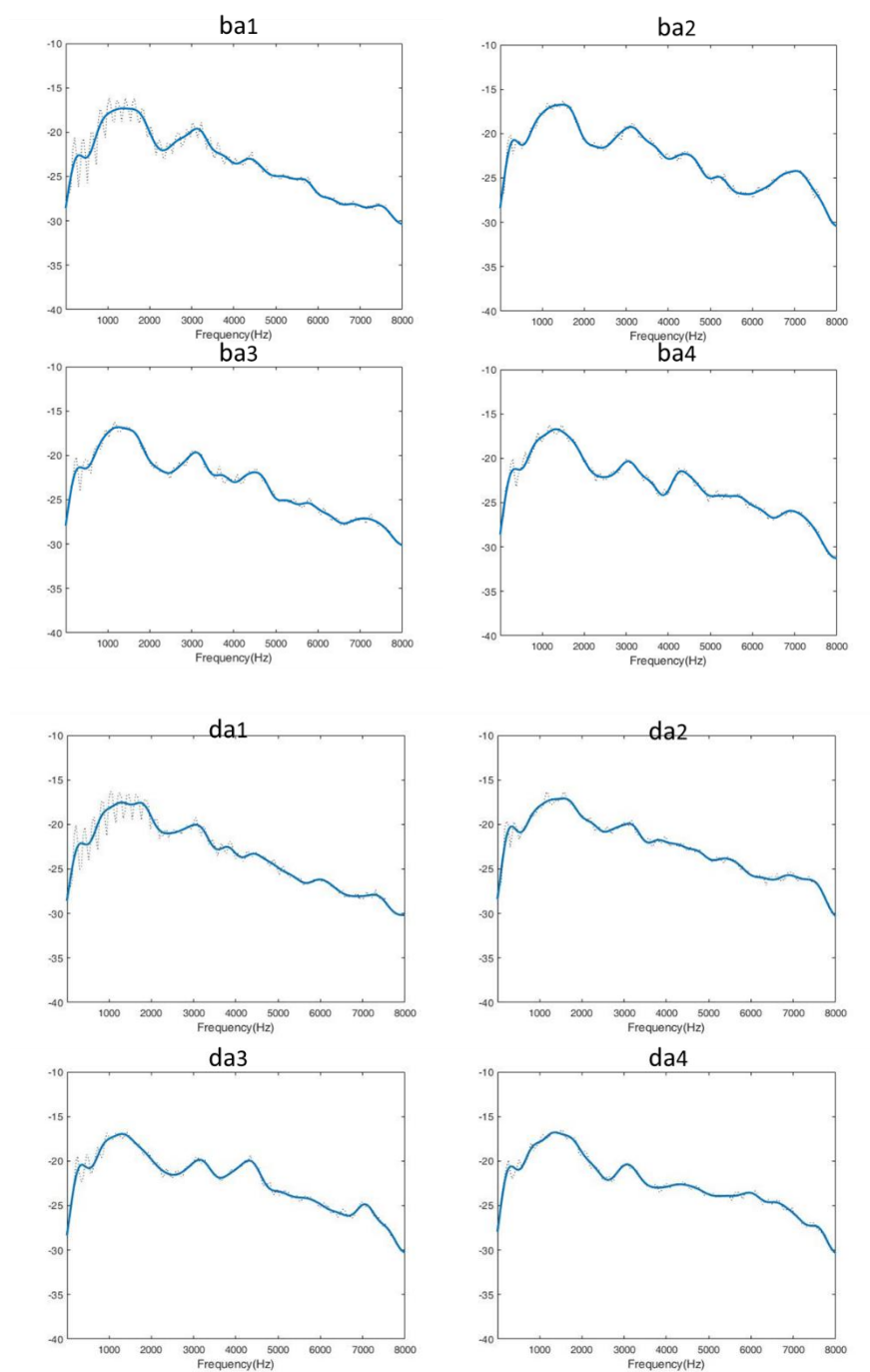
Individual confidence level = 99.82%

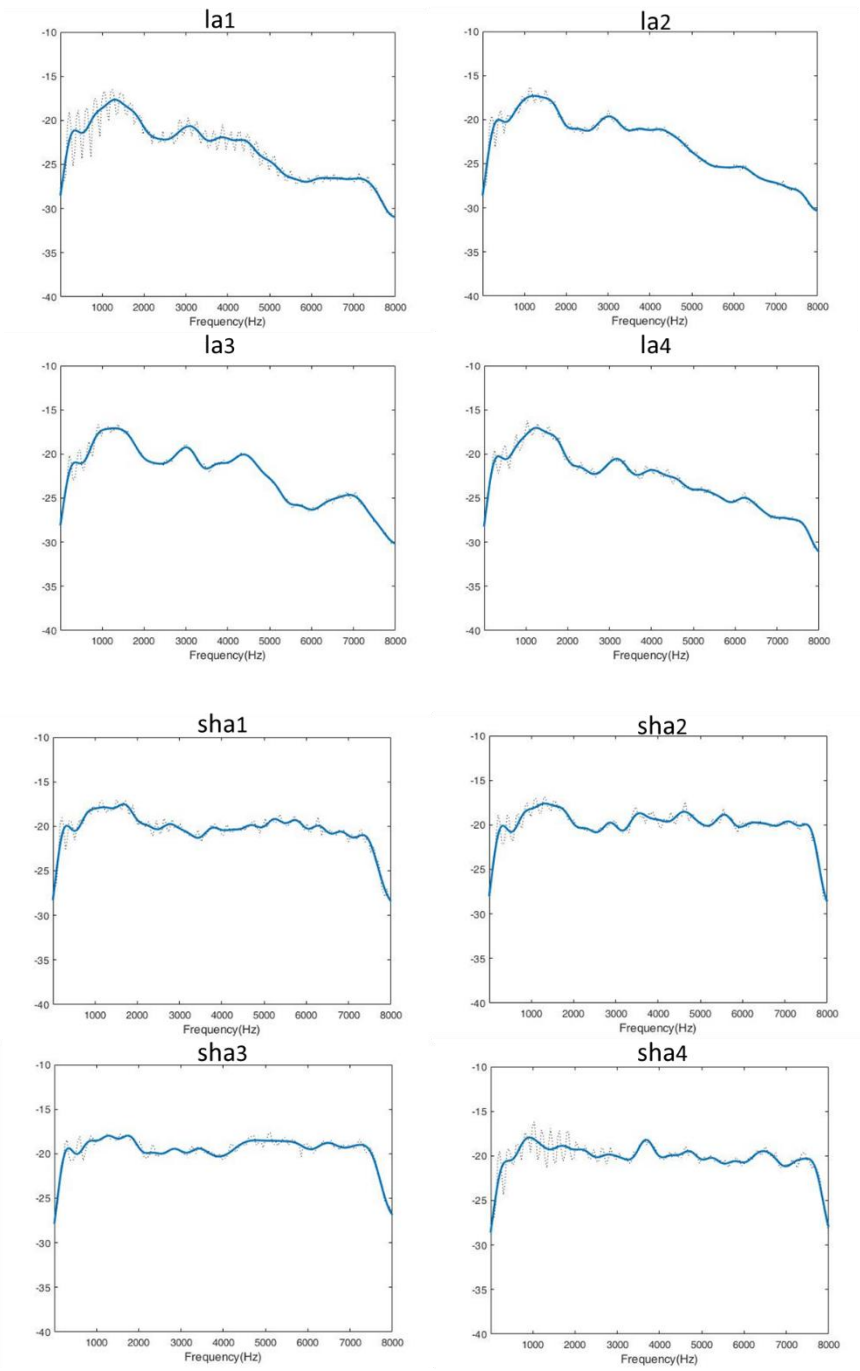
Appendix 4.4 Waveplot of single stimuli used in experiment two





Appendix 4.5 Spectrum and spectral envelope of single stimuli used in experiment two





Appendix 5.1 Combination methods of consonants and vowels

In the tables, possible combination was labeled with number “1”:

| | b | p | m | f | d | t | n | l |
|----|---|---|---|---|---|---|---|---|
| a1 | 1 | 1 | 1 | 1 | 1 | 1 | | 1 |
| a2 | 1 | 1 | 1 | 1 | 1 | | 1 | |
| a3 | 1 | | 1 | 1 | 1 | 1 | 1 | 1 |
| a4 | 1 | 1 | 1 | | 1 | 1 | 1 | 1 |
| o1 | 1 | 1 | 1 | | | | | |
| o2 | 1 | 1 | 1 | 1 | | | | |
| o3 | | | | | | | | |
| o4 | | 1 | 1 | | | | | |
| e1 | | | | | | | | |
| e2 | | | | | 1 | | | |
| e3 | | | | | | | | |
| e4 | | | | | | 1 | | 1 |
| i1 | 1 | 1 | | | 1 | 1 | | |
| i2 | 1 | 1 | 1 | | 1 | 1 | 1 | 1 |
| i3 | 1 | 1 | 1 | | 1 | 1 | 1 | 1 |
| i4 | 1 | 1 | 1 | | 1 | 1 | 1 | 1 |
| u1 | | 1 | | 1 | 1 | 1 | | |
| u2 | | 1 | | 1 | 1 | 1 | 1 | 1 |
| u3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| u4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| y1 | | | | | | | | |
| y2 | | | | | | | | 1 |
| y3 | | | | | | | 1 | 1 |
| y4 | | | | | | | | 1 |

Continued table:

| | g | k | h | j | q | x |
|----|---|---|---|---|---|---|
| a1 | | | 1 | 1 | | |
| a2 | | | | | | |
| a3 | | | | | | |
| a4 | | | | | | |
| o1 | | | | | | |
| o2 | | | | | | |
| o3 | | | | | | |
| o4 | | | | | | |
| e1 | | 1 | 1 | 1 | | |
| e2 | | 1 | 1 | 1 | | |
| e3 | | | 1 | | | |
| e4 | | 1 | 1 | 1 | | |
| i1 | | | | | 1 | 1 |
| i2 | | | | | 1 | 1 |
| i3 | | | | | 1 | 1 |
| i4 | | | | | 1 | 1 |
| u1 | | 1 | 1 | 1 | | |
| u2 | | | | 1 | | |
| u3 | | 1 | 1 | 1 | | |
| u4 | | 1 | 1 | 1 | | |
| y1 | | | | | 1 | 1 |
| y2 | | | | | 1 | 1 |
| y3 | | | | | 1 | 1 |
| y4 | | | | | 1 | 1 |

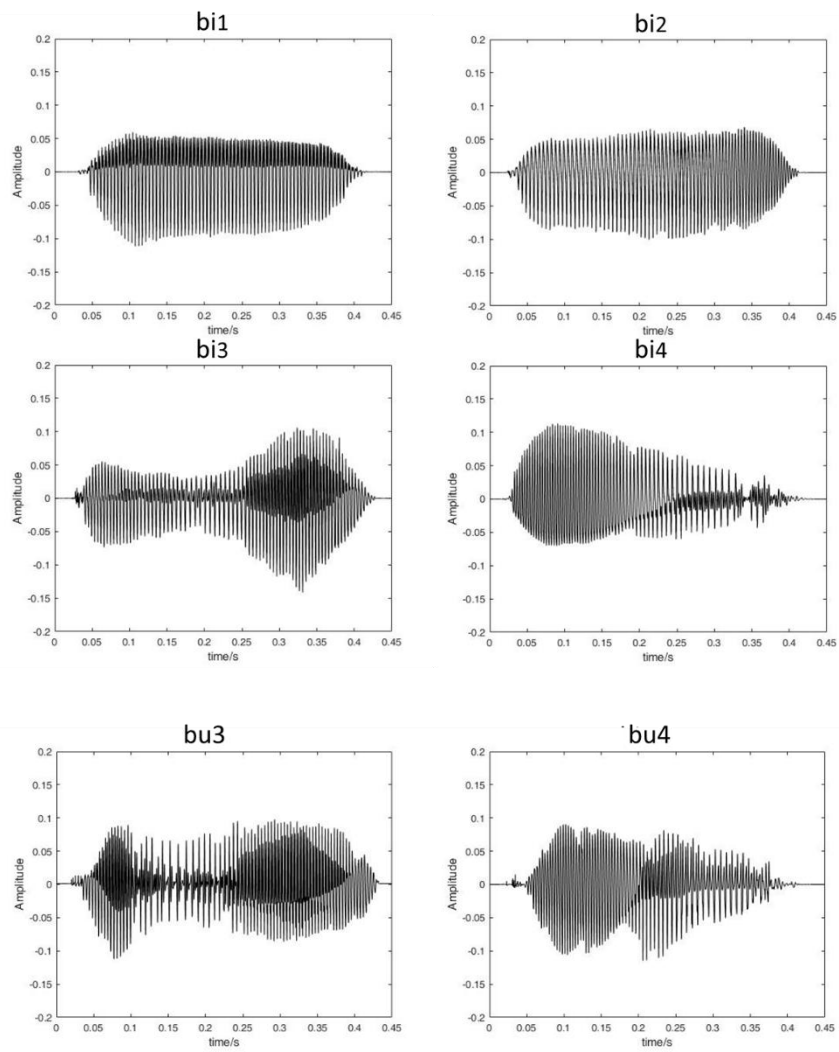
Continued second table:

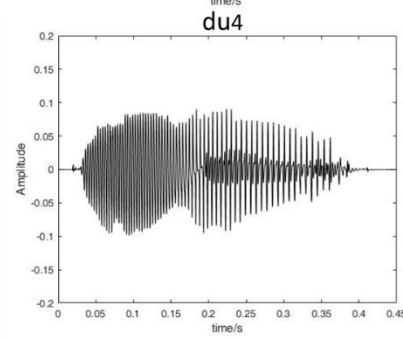
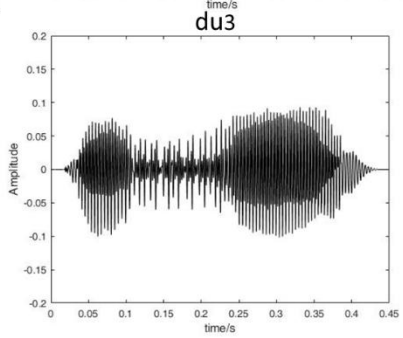
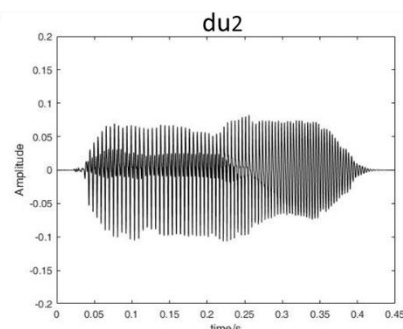
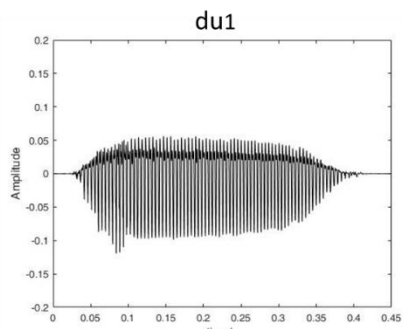
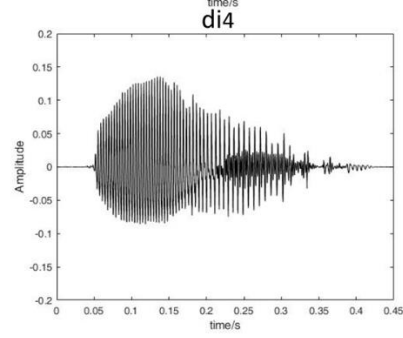
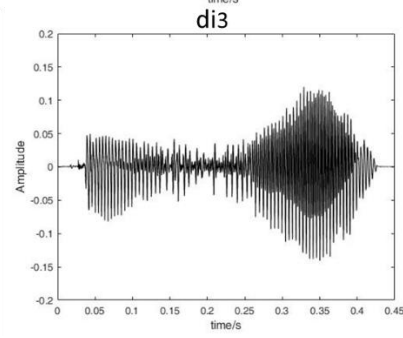
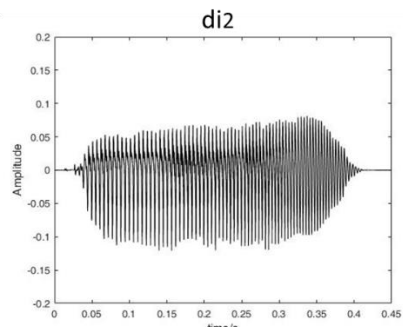
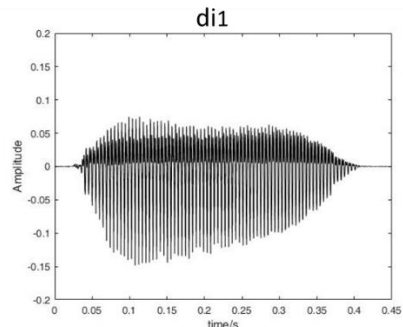
| | zh | ch | sh | r | z | c | s |
|----|----|----|----|---|---|---|---|
| a1 | 1 | 1 | 1 | | | 1 | 1 |
| a2 | 1 | 1 | | | 1 | | |
| a3 | 1 | | 1 | | | | 1 |
| a4 | 1 | 1 | | | | | |
| o1 | | | | | | | |
| o2 | | | | | | | |
| o3 | | | | | | | |
| o4 | | | | | | | |
| e1 | 1 | 1 | | | | | |
| e2 | 1 | | 1 | | 1 | | |
| e3 | 1 | 1 | 1 | 1 | | | |
| e4 | 1 | 1 | 1 | 1 | | 1 | 1 |
| i1 | 1 | 1 | 1 | | 1 | | 1 |
| i2 | 1 | 1 | 1 | | | 1 | |
| i3 | 1 | 1 | 1 | | 1 | 1 | 1 |
| i4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| u1 | 1 | 1 | 1 | | 1 | 1 | 1 |
| u2 | 1 | 1 | 1 | 1 | 1 | | 1 |
| u3 | 1 | 1 | 1 | 1 | 1 | | |
| u4 | | 1 | | 1 | | 1 | 1 |
| y1 | | | | | | | |
| y2 | | | | | | | |
| y3 | | | | | | | |
| y4 | | | | | | | |

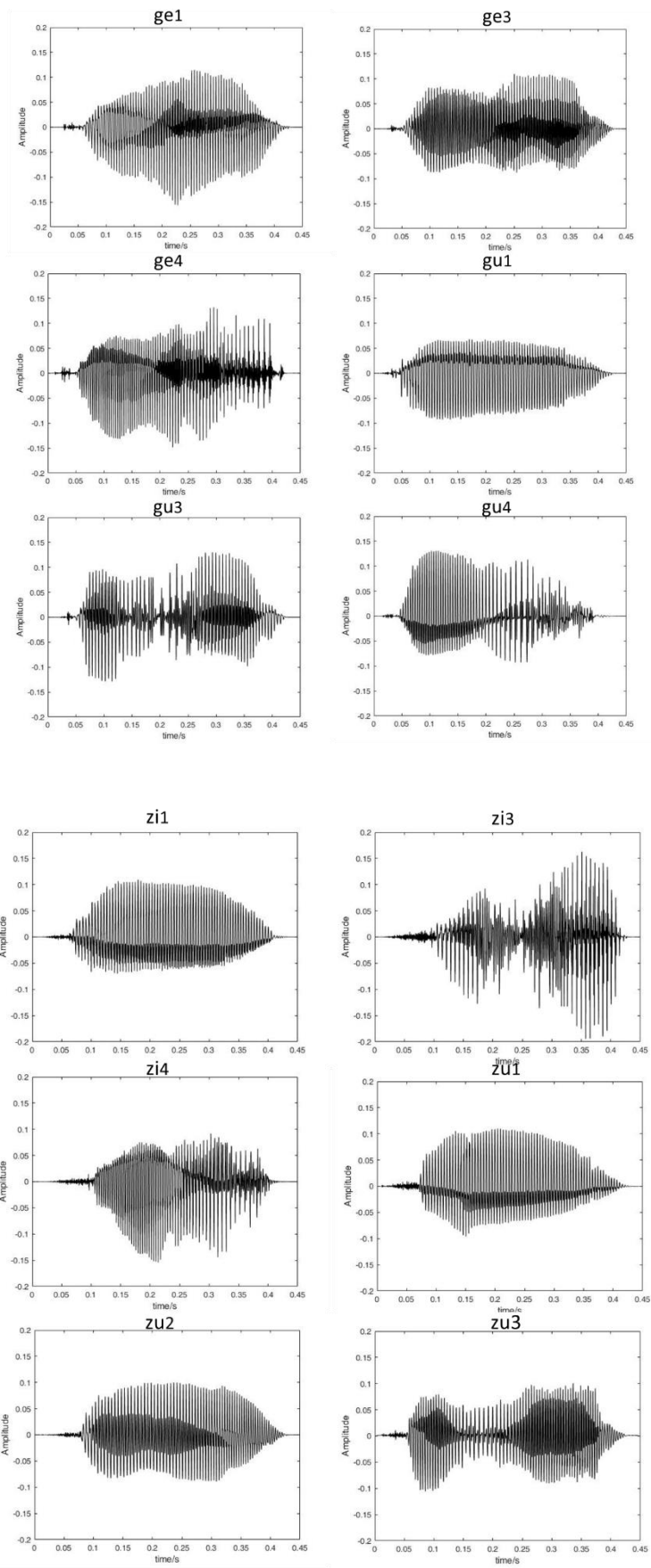
Appendix 5.2 Usage frequency table of vowels in mandarin

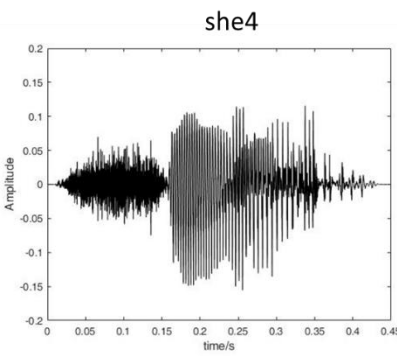
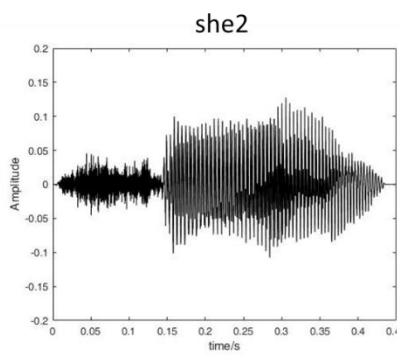
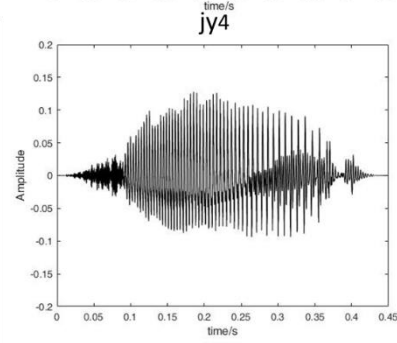
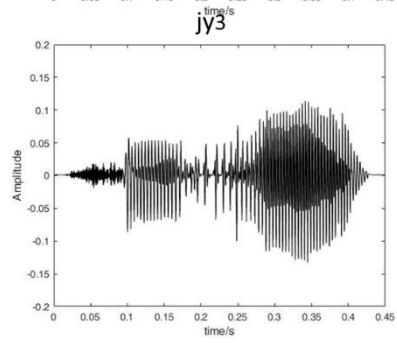
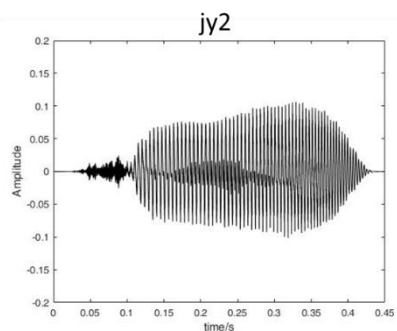
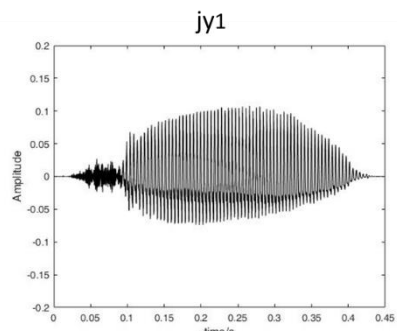
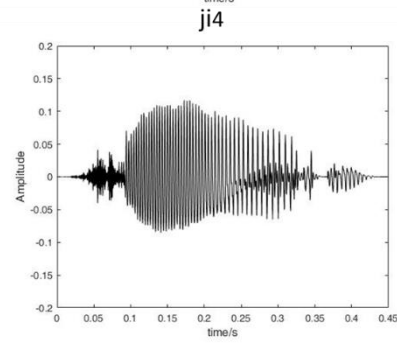
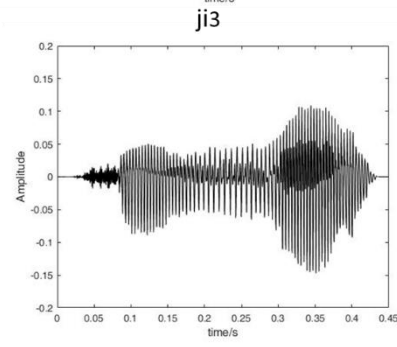
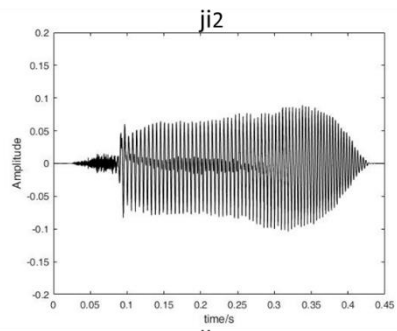
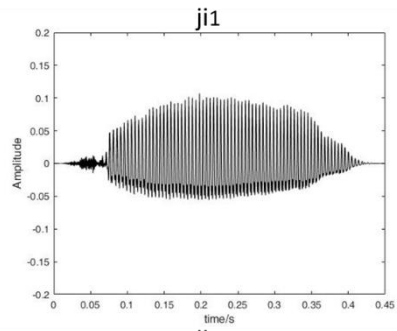
| Vowel | Frequency |
|-----------|-----------|
| i | 97072 |
| e | 85374 |
| u | 69901 |
| ian | 47125 |
| ai | 42051 |
| ing | 39894 |
| ong | 38850 |
| an | 38228 |
| uo | 37476 |
| ui | 33698 |
| a | 32212 |
| eng | 31452 |
| eng | 30658 |
| ang | 30510 |
| ao | 29445 |
| iu | 27481 |
| in | 26605 |
| yu | 26336 |
| ie | 22618 |
| iao | 22084 |
| ou | 21683 |
| iang | 19177 |
| uan | 17899 |
| ei | 17100 |
| uan | 12774 |
| ia | 12371 |
| uang | 9773 |
| ue | 9762 |
| un | 9357 |
| ua | 6889 |
| er | 6489 |
| yun | 5744 |
| iong | 4086 |
| o | 3777 |
| uai | 3649 |
| Total | 1051159 |

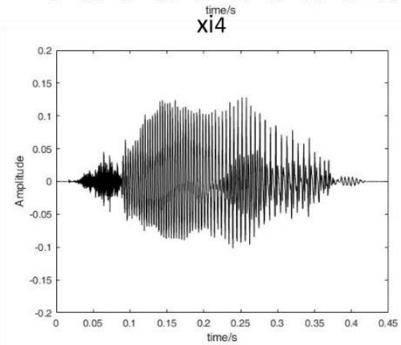
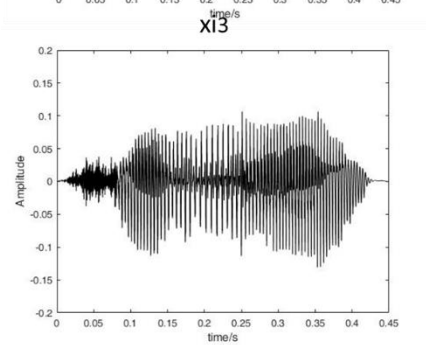
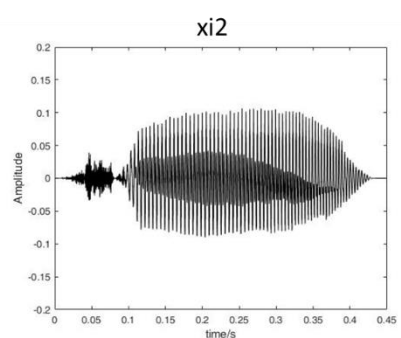
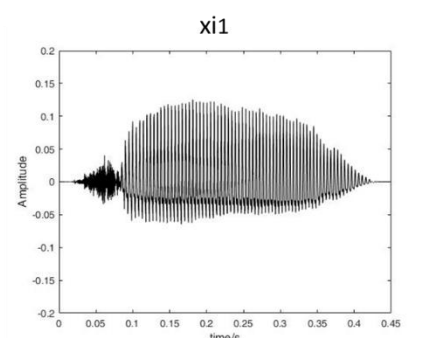
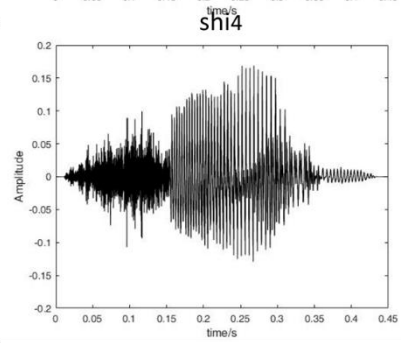
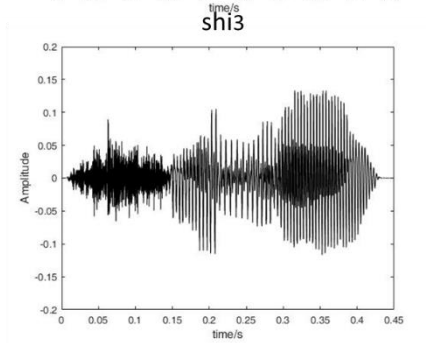
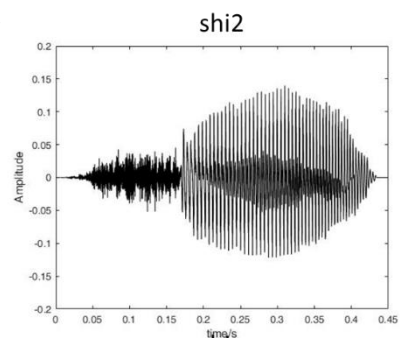
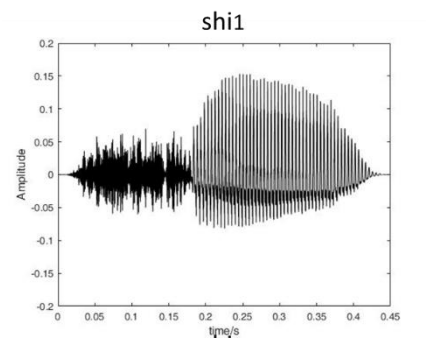
Appendix 5.3 Waveplot of single stimuli used in experiment three

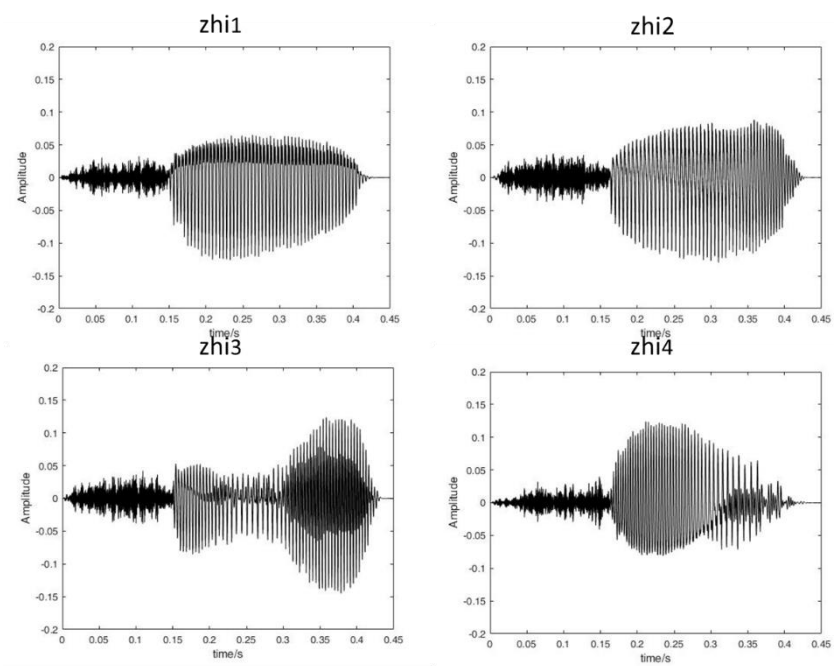












Appendix 5.4 Spectrum and spectral envelope of single stimuli used in experiment three

