The relationship between intrinsic f0, intrinsic pitch and lexical tone in Hong Kong Cantonese

Jessica Siddins

München 2016

The relationship between intrinsic f0, intrinsic pitch and lexical tone in Hong Kong Cantonese

Inauguraldissertation
zur Erlangung des Doktorgrades der Philosophie
an der Ludwig-Maximilians-Universität München

vorgelegt von
Jessica Siddins
aus St. George

2017

Erstgutachter: Prof. Dr. Jonathan Harrington
Zweitgutachter: Prof. Dr. Philip Hoole
Datum der mündlichen Prüfung: 7. Februar 2017

Contents

Abstract **ix**

Zusammenfassung **xi**

1 Introduction **1**
1.1 Theoretical background . . . 1
1.1.1 Prosody . . . 1
1.1.2 The relationship between speech production and speech perception 5
1.1.3 Sound change . . . 8
1.2 The present study: intrinsic f0, intrinsic pitch and lexical tone in Cantonese 12
1.2.1 Motivations . . . 13
1.2.2 Aims of this study . . . 15
1.2.3 Hong Kong Cantonese . . . 16

2 Experiment 1: Intrinsic f0 in Cantonese monophthong production **25**
2.1 Hypotheses . . . 25
2.2 Method . . . 26
2.2.1 Participants . . . 26
2.2.2 Stimuli . . . 26
2.2.3 Procedure . . . 27
2.2.4 Post-processing . . . 27
2.2.5 Analysis . . . 28
2.3 Results . . . 29
2.3.1 f0 analysis . . . 29
2.3.2 Machine classification . . . 31
2.4 Discussion . . . 33

3 Experiment 2: Intrinsic pitch in perception of Cantonese monophthongs **37**
3.1 Hypotheses . . . 38
3.2 Method . . . 39
3.2.1 Participants . . . 39
3.2.2 Stimuli . . . 40
3.2.3 Procedure . . . 41

3.2.4 Analysis . . . 42
3.3 Results . . . 43
3.4 Discussion . . . 46

4 Linking production and perception (Experiments 1 and 2) 51
4.1 Overview of results . . . 51
4.2 Are production and perception aligned? . . . 52
4.3 Individual variation . . . 56

5 Experiment 3: Intrinsic pitch in perception of Cantonese diphthongs 59
5.1 Hypotheses . . . 60
5.2 Method . . . 60
5.2.1 Participants . . . 61
5.2.2 Stimuli . . . 61
5.2.3 Procedure . . . 64
5.2.4 Analysis . . . 65
5.3 Results . . . 66
5.4 Discussion . . . 69

6 Experiment 4: Intrinsic f0 in Cantonese diphthong production 77
6.1 Hypotheses . . . 77
6.2 Method . . . 78
6.2.1 Participants . . . 78
6.2.2 Stimuli . . . 78
6.2.3 Procedure . . . 80
6.2.4 Post-processing . . . 80
6.2.5 Analysis . . . 81
6.3 Results . . . 82
6.3.1 Slopes . . . 82
6.3.2 Curvature . . . 85
6.4 Discussion . . . 87
6.4.1 Slopes . . . 88
6.4.2 Curvature . . . 90

7 General Discussion 91
7.1 Intrinsic f0: automatic or controlled? . . . 94
7.2 Intrinsic pitch: automatic or controlled? . . . 97
7.2.1 Virtual pitch model . . . 97
7.2.2 Gestural theories of speech perception . . . 98
7.2.3 Auditory enhancement . . . 99
7.2.4 Intrinsic pitch in tone languages . . . 101
7.3 Microprosody, macroprosody and the phonetics-phonology interface . . . 101
7.3.1 Sound change . . . 101

7.4 Tone processing 102
7.5 Further work 102

Appendices **105**

A Translations of stimuli **107**
A.1 Experiment 1 107
A.2 Experiment 4 107

B By-listener pre-examination of perception data in Chapter 3 **109**

C Spectrograms of perception stimuli from Chapter 5 **115**

D Example spectrograms from production data in Chapter 6 **127**
D.1 Speaker 1 128
D.2 Speaker 2 129
D.3 Speaker 3 130
D.4 Speaker 5 131
D.5 Speaker 6 132
D.6 Speaker 7 133
D.7 Speaker 8 134
D.8 Speaker 9 135
D.9 Speaker 10 136

E Original f0 contours from production data in Chapter 6 **137**

Disclosure of pre-published data **141**

References **143**

Acknowledgements **151**

Abstract

Changes in fundamental frequency result from two fundamentally different processes: phonological or linguistic determinants of f0 (macroprosody), which are used consciously by a speaker to achieve various intonation patterns or lexical tones, and phonetic determinants of f0 (microprosody), which are an automatic, unintended consequence of coarticulation with neighbouring or concurrent speech sounds. Vowel-intrinsic f0 refers to a phonetic phenomenon by which f0 decreases with vowel openness, while vowel-intrinsic pitch describes the increase in perceived pitch with vowel openness. This paradoxical relationship is believed to be necessary so that vowel-intrinsic (phonetic) f0 does not affect linguistic (phonological) f0 use in tone and intonation languages. However, there is evidence that perceptual normalisation is at best only partial. Thus arises the question as to how the listener processes the "leftover" intrinsic f0. We predict that, in tone languages, any portion of intrinsic (phonetic) f0 not fully compensated for in perception must be processed as phonological f0 and thus interpreted as tone. Therefore, the aim of this dissertation is to establish the degree to which intrinsic f0 interferes with phonological f0 in the form of lexical tone with broader implications for tone processing, sound change and speech technology.

In a first set of experiments, we investigated intrinsic f0 and intrinsic pitch on monophthongs in Hong Kong Cantonese. In terms of intrinsic f0, results revealed an f0-raising effect of close vowels on f0 of the high (55) level tone, but not the low (22) or mid (33) level tones in production. However, machine classification of f0 based on training with tone information only showed an effect of vowel openness on the correct classification of the low (22) and mid (33) level tones and not the high (55) level tone. This latter effect was reflected in perception, where we found a pitch-lowering effect of close vowels on the low (22) vs. mid (33) category boundary but not on the mid (33) vs. high (55) category boundary. Data from speaker-listeners who participated in both experiments revealed that intrinsic f0 and intrinsic pitch were matched for the lower tone contrast, but that intrinsic pitch lagged behind intrinsic f0 for the higher tone contrast. We concluded that this

relationship reflects compensation for coarticulation where a phonological contrast would otherwise be at risk of misperception.

In the second set of experiments, we extended a previous study by investigating whether perception of intrinsic pitch is dynamic in Hong Kong Cantonese and falls or rises depending on diphthong pattern. Contrary to our results for monophthongs, this was not the case. In order to establish whether this was part of a mismatch between production and perception, we additionally investigated intrinsic f0 in production of Cantonese diphthongs. However, there was no falling or rising intrinsic f0, which explained the absence of an intrinsic pitch effect in perception. Furthermore, both the diphthong production and perception studies revealed somewhat paradoxical results involving an effect in the opposite direction to our hypothesis in one experimental condition. We provide a number of possible explanations for these results, both theoretical and methodological.

Overall, we were able to confirm the presence of intrinsic f0 and intrinsic pitch in a complex tone system. However, the effects were not consistent across all conditions. Intrinsic pitch applied only to the condition in which intrinsic f0 has the most potential for disrupting tone contrasts, which we interpret as evidence of compensation for coarticulation. The results are discussed in terms of their implications for tone processing, the phonetics-phonology interface and speech technology.

Zusammenfassung

Veränderungen in der Grundfrequenz (f0) der Stimme sind das Ergebnis zweier grundsätzlich verschiedener Vorgänge. Die phonologischen oder linguistischen Determinanten von f0 (Makroprosodie) werden vom Sprecher bewusst eingesetzt, um verschiedene Intonationsmuster oder lexikalische Töne zu erzeugen. Im Gegensatz dazu stellen die phonetischen Determinanten von f0 (Mikroprosodie) eine automatische, unbewusste Folge der Koartikulation mit den Sprachlauten der unmittelbaren Umgebung dar. Die vokalintrinsische f0 bezieht sich auf ein phonetisches Phänomen, wodurch f0 mit steigender Vokalhöhe zunimmt, während die vokalintrinsische Tonhöhe eine perzeptuelle Normalisierung für diesen Effekt beschreibt, bei dem die empfundene Tonhöhe mit fallender Vokalhöhe abnimmt. Dieser paradoxe Zusammenhang scheint notwendig zu sein um die vokalintrinsische (phonetische) f0 daran zu hindern, die linguistische (phonologische) Verwendung von f0 in Ton- und Intonationssprachen zu beeinflussen. Allerdings lässt sich nachweisen, dass diese perzeptuelle Normalisierung bestenfalls nur unvollständig ist.

Dadurch ergibt sich die Frage, auf welche Art und Weise der Hörer die restliche intrinsische f0 verarbeitet, die in der Perzeption nicht normalisiert wird. Aus unserer Sicht lässt sich vorhersagen, dass in Tonsprachen jegliche Anteile von intrinsischer f0, die nicht vollständig bei der Perzeption kompensiert werden, als phonologische f0 verarbeitet werden müssen und somit als Ton interpretiert werden. Aus diesem Grund ist es das Ziel dieser Dissertation zu untersuchen, wie viel die intrinsische f0 die phonologische f0 (d.h. den lexikalischen Ton) beeinflusst, wobei sich weitreichende Implikationen für Tonverarbeitung, Lautwandel und Sprachtechnologie ergeben.

Um mögliche Interaktionen zwischen intrinsischer f0, intrinsischer Tonhöhe und lexikalischem Ton zu untersuchen wurde das Beispiel des Hongkong Kantonesischen herangezogen. Hongkong Kantonesisch ist eine komplexe Tonsprache, welche sowohl ein ausgeprägtes Vokalinventar als auch ein komplexes Tonsystem aufweist. Das Tonsystem besteht aus drei Registertönen (d.h. Tönen mit gleichbleibender Tonhöhe) sowie zwei steigenden und einem fallenden Konturton (d.h. Tönen, die eine Tonhöhenveränderung im Verlauf des Tons

aufweisen). Das Vokalinventar beinhaltet mehrere Vokale unterschiedlicher Öffnungsgrade oder Zungenhöhen, in Form von Monophthongen und Diphthongen. Somit war es möglich, den Effekt von gleichbleibender intrinsischer f0 und intrinsischer Tonhöhe bei Monophthongen sowie sich verändernde intrinsische f0 und intrinsische Tonhöhe bei Diphthongen zu untersuchen. Während es schon Hinweise aus dem Deutschen (Niebuhr, 2004) auf sich verändernde intrinsische Tonhöhe bei Diphthongen gibt, wurde dies noch nicht in Tonsprachen untersucht. Außerdem wurde die intrinsische f0 auf Diphthongen noch nicht getestet. Dynamische Effekte hätten Folgen für die Tonverarbeitung in komplexen Tonsystemen wie dem Kantonesischen und wären ein weiterer Beleg für die Annahme, dass den Effekten automatische biomechanische oder auditorische Prozesse zu Grunde liegen. Die intrinsische f0 wurde in Sprachproduktionsexperimenten erforscht, während die intrinsische Tonhöhe mittels Sprachperzeptionsexperimenten untersucht wurde.

In Kapitel 2 wird ein Sprachproduktionsexperiment dargestellt, bei dem Sprecher des Kantonesischen Minimalpaarreihen von realen Wörtern produzierten. Diese umfassten offene und geschlossene Vokale auf drei Registertönen. Die Vokalhöhe machte sich bei f0 des hohen (55) Registertons bemerkbar, indem geschlossene Vokale eine höhere f0 aufwiesen als offene, nicht jedoch bei den anderen Tönen. Eine maschinelle Klassifizierung nach einer Trainingsphase, die ausschließlich auf Toninformationen basiert war, zeigte einen interessanten Effekt. So beeinflusste die Vokalhöhe hier die korrekte Klassifizierung der tiefen (22) und mittleren (33) Registertöne, aber nicht die des hohen (55) Registertons. In Übereinstimmung mit früheren Studien war der Effekt der intrinsische f0 in den niedrigeren Bereichen der f0 eines Sprechers also vermindert (f0 Analyse), aber dennoch soweit ausgeprägt, dass sie den Kontrast zwischen den beiden niedrigeren Töne abschwächte (maschinelle Klassifizierung). Dies erklärten wir durch die größere Nähe der beiden tieferen Registertöne zueinander. Im Gegensatz zu dem weit markanteren Kontrast zwischen dem mittleren (33) und dem hohen (55) Registerton wird der Kontrast zwischen dem tiefen (22) und dem mittleren (33) Registerton schon bei sehr geringem Einfluss von Vokalhöhe gefährdet.

Kapitel 3 beschreibt eine Studie zur Tonverarbeitung, in welcher Hörern manipulierte Stimuli in einem *forced choice* Experiment präsentiert wurden. Das Vorhaben dieser Studie war festzustellen, ob die Hörer einen intrinsischen Tonhöhenunterschied wahrnehmen und ob, bei Bestätigung, die Größe des Effekts festgestellt werden kann. Es wurde ein f0 Kontinuum von niedrig nach hoch (d.h. von dem tiefen (22) Registerton bis hin zum hohen (55) Registerton über den mittleren (33) Registerton) hergestellt und exakt das gleiche

Kontinuum über Wörter mit offenen und geschlossenen Vokalen gelegt. Die Hörer empfingen auditiv einzelne Schritte des Kontinuums in randomisierter Reihenfolge und wählten die Wörter in orthographischer Form aus, deren Töne dem, was sie hörten, am nächsten kamen. Es ergab sich ein intrinsischer Tonhöheneffekt an der Grenze zwischen dem tiefen (22) und mittleren (33) Registerton, jedoch nicht an der Grenze zwischen dem mittleren (33) und dem hohen (55) Registerton. Dieser Umstand deutet darauf hin, dass Hörer die intrinsische f0 wenigstens teilweise normalisieren, und zwar dort, wo die intrinsische f0 den Tonkontrast am meisten gefährdet.

Für die genauere Untersuchung der Beziehungen zwischen der Produktion (intrinsische f0) und Perzeption (intrinsische Tonhöhe) stützt sich das Kapitel 4 auf die Daten aus den Kapiteln 2 und 3. Hier wird das Ergebnis der Produktion dem Ergebnis der Perzeption jedes Teilnehmers gegenübergestellt. Insgesamt scheinen dabei die Effekte der intrinsischen f0 und der intrinsischen Tonhöhe beim Kontrast zwischen den tiefen (22) und mittleren (33) Registertönen gleich groß zu sein, während bei dem Kontrast zwischen den mittleren (33) und den hohen (55) Registertönen die intrinsische f0 größer war als die intrinsische Tonhöhe. Allerdings zeigte sich bei näherer Untersuchung der Einzelergebnisse, dass die Teilnehmer erheblich variierten und dem Gesamtmuster nicht entsprachen. Somit hat es nicht den Anschein, dass die Teilnehmer, die eine ausgeprägte intrinsische f0 in der Produktion zeigten, auch in der Perzeption einen großen intrinsischen Tonhöhenunterschied wahrnahmen. Es scheint eher der Fall zu sein, dass Produktion und Perzeption bei individueller Betrachtung nicht korrelieren.

In Kapitel 5 findet sich eine Erweiterung einer Studie von Niebuhr (2004). Es wurde untersucht, ob sich die intrinsische Tonhöhe, wie in Kapitel 3 beschrieben, auf Diphthonge dynamisch auswirkt, und wie sich diese Betrachtung auf eine Tonsprache ausdehnt. Niebuhr (2004) fand einen senkenden Einfluss von steigenden Diphthongen sowie einen steigernden Effekt von fallenden Diphthongen auf die wahrgenommene Tonhöhe. Ein f0 Kontinuum von einer sich senkenden f0 zu einer ansteigenden f0 (d.h. vom fallenden (21) Konturton zum leicht steigenden (23) Konturton über den tiefen (22) Registerton im Kantonesischen) wurde generiert und über fallende und steigende Diphthonge gelegt. Wieder handelte es sich bei allen möglichen Vokal- und Tonkombinationen um reale Wörter. Die Hörer nahmen an einem *forced choice* Experiment mit drei Alternativen teil und wurden instruiert, das Wort mit dem Ton auszuwählen, welches dem präsentierten Stimulus am nächsten kam. Sollten die Hörer für den phonetischen Anteil von f0 (d.h. die intrinsische f0) in Diphthongen kompensieren, würde die Kategoriengrenze zwischen dem fallenden (21) Konturton

und dem tiefen (22) Registerton sowie auch zwischen dem tiefen (22) Registerton und dem leicht steigenden (23) Konturton früher für fallende als für steigende Diphthonge erwartet werden. Dies würde auch die Ergebnisse von Niebuhr (2004) bestätigen. Allerdings entsprachen die Ergebnisse nicht dieser Hypothese. Stattdessen schienen die Hörer an der Kategoriengrenze zwischen dem fallenden (21) Konturton und dem tiefen (22) Registerton die fallenden Diphthonge als stärker abfallend zu empfinden als steigende Diphthonge. Auf die Grenze zwischen dem tiefen (22) Registerton und dem leicht steigenden (23) Konturton hatte der Diphthong keinen Einfluss.

Kapitel 6 erweitert die Studie in Kapitel 2 auf die Produktion von Diphthongen und Konturtönen, um sie den Perzeptionsergebnissen im Kapitel 5 gegenüber zu stellen. Angenommen der Zusammenhang zwischen intrinsischer f0 und Vokalhöhe sei automatisch und inhärent, müsste die f0 auf steigenden Diphthongen ansteigen und auf fallenden Diphthongen sinken. Dieser Umstand könnte möglicherweise die kantonesischen Tonkontraste beeinträchtigen. Zu erwarten wäre zum Beispiel eine sehr ähnliche f0 Kontur des leicht steigenden (23) Tons auf steigenden Diphthongen im Vergleich zu dem stark steigenden (25) Ton auf fallenden Diphthongen. Die Sprecher produzierten Minimalpaarreihen bestehend aus steigenden und fallenden Diphthongen sowie einen fallendsteigenden Triphthong und einen Monophthong in der Kontrollbedingung auf den leicht steigenden (23) und fallenden (21) Konturtönen und dem tiefen (22) Registerton. Unsere Hypothese ließ sich nicht nachweisen. Im Gegenteil, das einzige statistisch signifikante Ergebnis war ein die f0 absenkender Effekt bei steigenden Diphthongen in Kombination mit dem fallenden (21) Ton.

Zusammenfassend war es möglich, den Effekt intrinsischer f0 und intrinsischer Tonhöhe in einem komplexen Tonsystem zu bestätigen. Allerdings zeigten sich die Effekte nicht über alle unterschiedlichen Bedingungen konsistent. Bemerkenswert war das Ausbleiben eines dynamischen, sich verändernden Effekts von intrinsischer f0 auf Diphthonge. Möglicherweise trat der Effekt hier nicht ein, weil die intrinsische f0 in den tieferen f0-Bereichen generell stark eingeschränkt wird. Andererseits ist es möglich, dass die unterschiedliche Artikulation sowie die daraus entstehenden spektralen Eigenschaften der Diphthonge im Vergleich zu den Monophthongen dazu führen, dass Diphthonge für die intrinsische f0 generell nicht anfällig sind. Der Effekt der empfundenen intrinsischen Tonhöhe tritt nur dort auf, wo die intrinsische f0 den Tonkontrast in der Sprachproduktion abschwächte. Dies wurde als Nachweis für die bewusste Kompensation von Koartikulation interpretiert. Zusammen mit den Ergebnissen vorheriger Studien scheint die intrinsische Tonhöhe somit kein automatis-

ches auditorisches Phänomenon zu sein (und wäre insofern nicht „intrinsisch“).

Chapter 1

Introduction

1.1 Theoretical background

1.1.1 Prosody

It is no easy task to draw a line between the segmental and suprasegmental domains of speech, and as we will see, the two domains interact. A speech analysis focussing on the segmental level might try to separate vowels from consonants or voiced from unvoiced tokens. It might measure voice onset time of an obstruent or the exact place of articulation of sound. An analysis at the suprasegmental level may involve equally diverse concepts. On the one hand, there are microscopically small phenomena such as jitter or shimmer in aging speech, while on the other hand there is the sing-song melody of an entire utterance or even a family of languages.

Speech prosody is the suprasegmental domain of phonetics. Its acoustic correlates in the speech signal are fundamental frequency, intensity and duration, which are most commonly measured in Herz (Hz), decibels (dB) and milliseconds (ms), respectively. Psychoacoustically, while duration is relatively straightforward, fundamental frequency is perceived as pitch and intensity as loudness, but these perceptual parameters are much more difficult to gauge.

In this study, we focus on fundamental frequency and pitch. The difference between the two is an important one that is not always acknowledged in the literature. It is especially important to differentiate between the two in this study, in which we attempt to compare them.

Fundamental frequency and pitch

Both phenomena relate to the glottal vibrations that cause voicing. These vibrations can be measured and the resulting signal depicted as a serious of repeating, quasi-periodic waveforms. Fundamental frequency (known as f0) is a physical measure of the number of waveform repetitions per second resulting from glottal vibration and is measured in Herz (Hz). Pitch, however, is a psychoacoustic phenomenon that is not as easy to measure, as it

represents a listener's perception of f0 rather than the f0 itself (Stoll, 1984). All else being equal, perceived pitch tends to be positively correlated with f0, in that an increase in f0 causes an increase in pitch and vice versa[1], but this relationship is by no means linear or even straightforward. Thus, f0 is a physical and pitch a psychological phenomenon, and both have different functions in speech communication. f0 is a product of the larynx, while pitch has its origins in the inner ear and the brain.

Microprosody vs. macroprosody

Prosodic phenomena are often split into two subtypes.

Macroprosody involves the modulation of f0, intensity and duration for communicative purposes. Languages such as Italian make a phonological distinction between long and short consonants (so-called geminates), while languages such as English do not. Speech in a noisy environment such as a café might be produced louder, in order to be understood, than in a quiet context such as a museum. f0 might be modulated to express surprise or anger. Macroprosody also signals tone (see separate section on tone below) and intonation (e.g. question vs. statement) and therefore is largely language-specific, although certain patterns are more common than others.

Microprosody is inherently different. It results automatically from speech mechanisms that cannot be controlled by the speaker and is thus inherent in speech. While microprosodic phenomena may or may not be (consciously) audible, they are not willingly modulated and accordingly tend to be perceptually less salient or obvious in everyday speech. For example, it is not common knowledge outside of speech science that open vowels are associated with an intrinsically slightly longer duration than close vowels due to the slightly longer time it takes for jaw to open and tongue position to lower for open vowels. Other phenomena such as jitter and shimmer, already mentioned above, refer to uncontrolled microvariation of f0 and intensity, respectively, and vary on a continuum from largely imperceptible in healthy adult speech to perceptually very prominent and at times even destructive to successful communication in pathological speech.

Essentially, the difference is between phonetic and linguistic variation of speech prosody. Microprosody is a result of intrinsic physiological constraints or aerodynamic perturbations, while macroprosody is actively controlled by the speaker. We might logically conclude, then, that macroprosody is relevant for speech communication, while microprosody is not. However, previous studies have indicated that this is not the case.

Microprosody has been known to encroach on the domain of macroprosody. For example, according to Jun (1996), phrase-initial strengthening (an intonation pattern) in Korean originated from f0 perturbation of the phrase-initial onset obstruent. Going further back, tonogenesis (the evolution of tones in a non-tone language) in South East Asia has largely been attributed to the phonologisation of segmental effects on microprosody. Specifically, the voicing contrast in onset obstruents is commonly associated with microprosodic f0 perturbation: voiced obstruents lower the f0 onset of the following vowel, while voiceless

[1] See Gelfand (1998) for an excellent discussion of the complicated relationship between f0 and pitch.

obstruents raised it (the underlying mechanisms are discussed in the next section). Over time, the phonetic effect of obstruent voicing on the f0 of following vowel was mistaken as the phonological cue and the rising and falling tones appeared in place of the voicing contrast. Evidently, microprosody and macroprosody interact.

Intrinsic f0 vs. intrinsic pitch

One way the segmental and suprasegmental domains inevitably overlap is during coarticulation. Just as overlapping articulatory gestures or auditory contrasts influence one another, coarticulatory effects arising from speech segments can cross the divide into the suprasegmental domain. Intrinsic f0 is a particularly well-documented case of segmental features influencing speech prosody.

Intrinsic f0 may refer to a number of microprosodic phenomena. The above-mentioned microprosodic f0 variation at vowel onset due to onset obstruent voicing is one such case. The term "intrinsic" is used to denote the uncontrolled, phonetic nature of the effect in question (i.e. in the above case the origin of the f0 variation is simply a phonetic side effect of the onset obstruent voicing), although as we will see, this account is still debated to some extent in the literature. Vowel-intrinsic f0 refers to the universal tendency[2] for f0 to decrease with vowel openness[3]: all else being equal, close vowels are associated with a higher f0 than open vowels (e.g. Hombert, 1977b). This is the type of intrinsic f0 that is the focus of this study. A number of theories, both aerodynamic and physiological, have been proposed to explain this phenomenon and its universal tendencies. The *source-tract coupling theory* (Flanagan, 1965; Lieberman, 1970; Atkinson, 1973) proposes an interaction between vocal tract configuration and vocal folds, whereby the closer F1 is to f0, the stronger its ability to raise f0. F1 is closest to f0 when it is low, such as in close vowels. Hence, the F1 of close vowels but not open vowels, are supposed to raise f0. More support is found in the literature for various versions of the so-called *tongue-pull theories*. The vertical tension hypothesis (Ladefoged, 1964; Lehiste, 1970) predicts that the raising of the tongue for close vowels pulls the larynx up via the hyoid bone, thereby increasing the vertical pull on the vocal folds and increasing f0. Instead, the horizontal tension hypothesis (Honda, 1983) argues that contraction of the posterior genioglossus muscle for close vowels pulls the hyoid bone forward, resulting in horizontal pull on the vocal folds and subsequent f0 raising. Yet other theories show evidence that multiple mechanisms play a role, including possible active suppression or enhancement of intrinsic f0 differences to facilitate speech perception (e.g. Silverman, 1987; Connell, 2002; Hoole & Honda, 2011). This study does not aim to evaluate directly which (if any) of these theories best explains the phenomenon. Instead, the focus lies on documenting both the presence and the size of the effect where it co-occurs with macroprosodic (i.e. tonal) perturbation of f0.

[2]See Whalen and Levitt (1995), but see also Connell's findings for Mambila (Connell, 2002).

[3]Normally, we would phrase this differently. f0 is generally said to increase with vowel *height*, but as we will be dealing extensively not only with vowel height but also f0 height and tone height, we will refer to vowel *openness* for clarity. Thus, rather than associating *low* vowels with lower f0, we will refer to *open* vowels as having lower f0.

Vowel-intrinsic pitch refers to a phenomenon in pitch perception roughly equivalent to the intrinsic f0 effect found in vowel production[4]. Interestingly, there is a somewhat paradoxical relationship between intrinsic f0 and intrinsic pitch. At the macroprosodic level (e.g. tone and intonation), f0 patterns with and is correlated with pitch, so a lexical tone with a high f0 is perceived as having high pitch. Typically, in spite of higher intrinsic f0 on close vowels and lower intrinsic f0 on open vowels, open and close vowels are perceived as having the same (perceived) pitch. Accordingly, Hombert (1977b) found that when the f0 of open and close vowels was manipulated to be exactly equal (keeping all other factors constant), open vowels were perceived as having higher pitch than close vowels, even though open vowels are known to induce a lower intrinsic f0 than close vowels in normal speech. However, the "intrinsic" nature of vowel-intrinsic pitch is yet to explained. In fact, the intrinsic pitch phenomenon is considerably less well-documented and less robust compared with the intrinsic f0 effect, and it has been suggested that the effect is language-specific rather than universal (Pape & Mooshammer, 2008; Pape, 2009) (in which case the label "intrinsic" is misleading). Regardless of the mechanisms behind intrinsic pitch, for the purposes of this investigation it is important only that it refers to a perceptual effect that is negatively correlated with the vowel-intrinsic f0 effect found in perception.

Intrinsic f0 therefore refers to physical differences in f0 as a result of coarticulation (that is, unintended, phonetic variation[5].). Consequently, in our study we use this term to describe the effect of differing vowels on f0 as measured in Hz, typically in speech production experiments.

As the psychoacoustic correlate of intrinsic f0, intrinsic pitch instead refers to the way f0 is perceived depending on phonetic environment. Crucially, we use this term to describe the effect of differing vowels on pitch, typically in speech perception experiments.

Thus, throughout the remainder of this work, we will use the terms "intrinsic f0" and "intrinsic pitch" to refer to vowel-intrinsic effects on f0 in speech production and pitch in speech perception, respectively, unless noted otherwise. The investigation does not consider any other types of intrinsic f0.

Tone

We have already discussed the difference between f0 and pitch. A brief discussion of yet another f0-related term is in order, because it is one of the critical variables in our study. Tone refers to the contrastive use of f0 to distinguish between lexical meanings or grammatical forms of a word[6]. As such, it has phonological and morphological functions

[4]It is important to note here that, just as the distinction between f0 and pitch is often neglected, so is the distinction between intrinsic f0 and intrinsic pitch. We follow authors such as Hombert (1977b), Fowler and Brown (1997), Niebuhr (2004) and Pape (2009) in distinguishing clearly between the two. In this study, intrinsic f0 refers to the effect of vowel openness on f0 in production and intrinsic pitch to the psychoacoustic effect of vowel openness on pitch perception.

[5]For an alternative account of intrinsic f0, see Diehl and Kluender (1989); Kingston (1992, 2007).

[6]The use of the term "word" here is somewhat of a simplification. In general, tones are spread across the syllable, although the exact definition of the tone-bearing unit is still a matter of debate. Duanmu (1990, Section 3.5) lists the two main possibilities as either the voiced portion of the entire syllable or the

and is a macroprosodic phenomenon. There are different types of tones and tone languages. In this work, we focus on "true" tone languages that distinguish between register and/or contour tones, such as the tone languages of Africa or South-East Asia[7]. Tone is not a static phenomenon. Just as a speech segments such as vowels have a temporal dimension, so do tones, which are characterised not only by their height but also by their movement in time. Register tones are characterised by a relatively constant f0 over the course of the tone, while contour tones are characterised by their f0 slope (e.g. "falling" or "rising") or even what we will refer to as their f0 curvature (e.g. "falling-rising" or "dipping"). Gandour (1979) refers to these three dimensions as height, direction and contour, respectively. All "true" tone languages make use of register tones, while contour tones are less widespread (Maddieson, 1978).

Each tone language has its own language-specific tone inventory, and the tones of any one language are described in relation to each other rather than as a type of absolute measure. For example, the high and low tones in a tone system comprising five different tones are likely to be further apart (in f0 and pitch) than and thus different from the high and low tones in a two-tone system (Maddieson, 1978). Similarly, what is described simply as a "falling" tone in Mandarin is not necessarily similar to a "falling" tone in Cantonese. More exact descriptions can be attempted using Chao tone numerals, introduced by Yuen Ren Chao to designate relative pitch height on a scale from 1 (lowest) to 5 (highest). According to this system, the canonical form of the Mandarin falling tone might be described as 51 (that is, pitch begins at the upper limit of the tone space and sinks to the lower limit), whereas the Cantonese falling tone might be described as 21 (pitch begins at a low level and sinks to an even lower level).

For an even more exact description, tone can be measured on the basis of f0 (Hz). Note, however, that the exact f0 values will vary both within and between speakers due to speech and speaker physiology and factors such as coarticulation. An f0 contour that might be representative of the average female speaker's low tone will certainly be higher than the average male speaker's low tone and possibly higher even than his high tone. It is the tone space within an individual speaker that is relevant. Importantly, listeners are well aware of this and automatically adjust for speaker-specific characteristics (e.g. Wong & Diehl, 1998).

Thus, we see that f0 is a physical measure, pitch its psychoacoustic correlate, and tone a linguistic concept. All three are related but clearly separate identities.

1.1.2 The relationship between speech production and speech perception

Presumably, speech production and perception share the same goal: successful communication.

voiced portion of the syllable rime only (f0 modulation is by definition restricted to voiced segments).

[7]but our study might be equally important for "pitch accent" tone languages that distinguish between tones in certain contexts only, such as Swedish or Japanese.

Speech production

At least since the work of Menzerath and de Lacerda (1933), it has been evident that speech production does not merely involve a string of constant, unvarying units simply concatenated together to compose larger and larger segments such as syllables, words and phrases. As the speech signal in the form of oscillograms and spectrograms shows, the "boundaries" we perceive between various segments, syllables, words or phrases in speech perception have no clear visible correlates. Rather, speech production is inherently characterised by variability, overlap and contextual variation in which no one segment is produced the same way twice.

Speech perception

Accordingly, an unresolved question is how the signal is successfully decoded in speech perception. Some theories of speech perception, such as motor theory and direct realism (A. M. Liberman & Mattingly, 1985; Fowler, 1986) argue that rather than perceiving abstract entities such as phonemes, we perceive directly the articulatory gestures carried out during speech production. As such, they argue, the lack of invariance in the speech signal is not a problem for speech perception, which is uniquely different from other types of auditory processing because it is speech-specific. In a sense, production aligns itself according to perception. Coarticulation is perceived directly as it occurs and the listener is able to attribute it automatically to its gestural source. One implication of this theory is that speech perception is very tightly linked to and indeed inseparable from production.

Other theories of speech perception propose that there is in fact nothing unique about speech perception. Instead, proponents of theories such as the spectral contrast model (Diehl & Kluender, 1989) or the virtual pitch model (Terhardt, 1974) argue that speech is processed like any other auditory stimulus. Support for this model comes from evidence that other species are able to process speech similarly to humans (Lotto, Kluender, & Holt, 1997). Models such as the auditory enhancement theory (Kingston, 1992; Kingston et al., 2014) as well as Lindblom's H & H and adaptive dispersion theories (Liljencrants & Lindblom, 1972; Lindblom, 1990) indicate that rather than speech perception adapting itself to variation in speech production, it is instead the speaker that takes the specific speech situation and requirements of the listener into account during speech production.

Compensation for coarticulation

One account of why perception is generally quite accurate and unproblematic despite the huge amount of intrinsic and extrinsic variation in production is that listeners are able to normalise perceptually for these factors. In the case of phonetic variation, there is strong support for the idea that listeners are able to compensate for coarticulation in the signal. In a now-classic perception study, Mann and Repp (1980) synthesised a continuum between [ʃ] and [s]. The ambiguous fricative preceded rounded [u] or unrounded [a]. Listeners were instructed to choose one of the two fricatives according to what they perceived. Naturally, perception was categorical, but the category boundary differed depending on the following

vowel: preceding [u], listeners' responses were biased toward [s]. The authors interpreted this as evidence that listeners attributed some of the spectral qualities of the fricatives to the influence of vowel context[8]. This study and others indicate that listeners take account of coarticulatory processes during speech perception. In a sense, perceptual normalisation might factor out the variability in production, explaining why speech communication is generally successful in the face of so much variation.

However, the concept of compensation for coarticulation by no means implies that coarticulation in production is completely factored out in perception. As noted above, Hombert (1977b) found a perceptual intrinsic pitch effect that is negatively correlated with the intrinsic f0 effect in production. He ruled out an explanation for the perceptual effect on the basis of spectral characteristics of vowel openness and instead suggested that intrinsic pitch reflects perceptual normalisation for the intrinsic f0 effect found in production:

> Since vowel quality and intrinsic f0 are always produced simultaneously, it is possible that our auditory system subjects the speech signal to some form of normalization having the effect of raising the pitch of low vowels (or lowering the pitch of high vowels) (Hombert, 1977b, p. 15).

Yet Hombert notes that the difference in intrinsic pitch between the open and close vowels in his data is smaller than the intrinsic f0 differences reported prior to his study. This is just one of many accounts of what is believed to be compensation for coarticulation not being well-aligned with the coarticulation itself. For example, in a study of spoken and sung vowels, Fowler and Brown (1997) found that the size of intrinsic pitch in perception was seldom equal to the size of the intrinsic f0 effect in production. Similarly, incomplete compensation for coarticulation has also been found in vowel-on-vowel coarticulation (Beddor, Harnsberger, & Lindemann, 2002; Fowler, 2005) and intrinsic vowel duration (Lindblom, 1967).

In addition, compensation for coarticulation has also been shown to be language-specific. For example, in a cross-linguistic study of vowel-on-vowel coarticulation, Beddor et al. (2002) found that compensation for coarticulation depended on native language experience.

Complete compensation for coarticulation would provide an excellent account of how speech perception is so robust despite the relatively messy input from speech production. Yet the reports of incomplete and language-specific compensation are somewhat problematic for this account. In recent years, models of phonetically biased sound change have argued that misperception of coarticulation could trigger sound changes. Thus, it may be that where compensation is complete, production and perception are aligned and there is no possibility of phonetically biased sound change. In contrast, where compensation is incomplete, production and perception are misaligned and what the speaker says is not

[8]That is, in the [u] context they attributed the low third formant transitions not to the labialisation they might have otherwise perceived as part of [ʃ] but to the vowel rounding. In the [a] context, the absence of any other explanation for low third formant transitions into the vowel led the listeners to perceive more [ʃ].

exactly what the listener hears. When it is the listener's turn to speak, his or her speech might be minimally influenced by the (misparsed) speech he or she has just heard.

1.1.3 Sound change

The sound patterns of all languages tend to change over time. Sound change patterns are manifold (see Campbell, 2013, for an overview). They can involve allophonic or phonemic changes; they can be phonetically or lexically conditioned; they can apply to some parts of the community and not others.

Sound change can begin with a wide variety of phonological processes. Neutralisation, for example, has led to a sound change known as final obstruent devoicing in many languages. Assimilation, on the other hand, has led to vowel harmony in languages such as Hungarian and a distinction between nasal vs. oral vowels in some Romance languages. Metathesis is also seen in sound changes, for example in the change from Latin <pa**r**abo**l**a> to Spanish <pa**l**ab**r**a>.

Mergers are a type of sound change in which a phonemic contrast is lost (such as in final obstruent devoicing). For example, an existing phoneme merges with an another existing phoneme (A, B > B), or two existing phonemes merge into a new phoneme (A, B > C). *Splits* show the opposite pattern: one phoneme splits into two (e.g. A > A, B or A > B, C), such as in the development of contrastive nasal and oral vowels from oral vowels in nasalised and non-nasalised contexts. Diachronically, splits have been shown to follow mergers, whereas the reverse pattern has not been documented.

Chain shifts are interconnected sound changes, similar to the idea of one sound change setting off a chain reaction of events. Grimm's Law is one such chain shift. It describes the process by which voiceless plosives became voiceless fricatives, voiced plosives became voiceless plosives, and voiced aspirated plosives became voiced unaspirated plosives during the development of Proto-Germanic from Proto-Indo-European. Chain shifts can be separated into two sub-types. In push chains, one phoneme is said to move towards another and push it out of its place. This second phoneme may then push yet another phoneme out of its spot, and so on. In pull chains, a phoneme moves away from its original spot, freeing up room for another phoneme to take its place. Yet another phoneme may move into the gap left by the second phoneme, and so on.

In our work, we are particularly interested in phonetically motivated sound change. Phonetically motivated sound changes are those resulting from coarticulatory patterns, although the mechanisms by which such patterns cause sound change are unclear. Some of the currently popular theories of how phonetically conditioned sound changes occur are described below.

Sound change according to John Ohala

This model is based on the idea of compensation for coarticulation as discussed above in Section 1.1.2. During successful speech communication (i.e. most of the time), Ohala describes the following hypothetical situation:

> A speaker intends to utter the sound sequence /ut/. [...] This utterance may be distorted if it is subject to coarticulation into something like [yt]. This version is transmitted to the listener who applies his "reconstructive" rules, which [...] crucially depend on his having correctly perceived the environment causing the distortion, in this case the [t]. The listener therefore reconstructs the intended signal /ut/ (Ohala, 1981, pp. 182-183).

Ohala then goes on to pose a second hypothetical situation in which the listener fails to compensate for coarticulation in which the speaker again intends to say /ut/ but again produces something like [yt]. If the /t/ is not perceptually salient enough (e.g. due to coarticulation or noise), then the listener will be unaware of the conditioning environment for the [y] and hence be unable to compensate for the fronting effects of the dental stop on the vowel. As a result, the listener should both "hear" and "interpret" the intended signal as /y/. For Ohala, this would constitute what he considers a phonetically motivated mini sound change in the mind of one speaker-listener[9]. A number of historical sound changes are believed to have come about by this kind of process, some of which are described below in the discussion of tonal sound change. Thus, the listener's ability to normalise perceptually for phonetic effects in the speech signal plays a crucial role in this model of sound change.

Ohala notes that such misperceptions are rare and mini sound changes even rarer. Normally, the listener has access to other contextual knowledge to accurately recover the intended signal. Instead, Ohala suggests that it is inexperienced language users such as language learners (either L2 speakers or children still acquiring L1) who are most likely to make errors in perception.

Not just historical changes such as tonogenesis but also synchronic change has been found to largely support this theory (Harrington, Kleber, & Reubold, 2008; Beddor, 2009; Kuang & Cui, 2016).

Sound change according to Lindblom and colleagues

According to Lindblom's H & H theory mentioned above, much within-speaker variation comes about because the speaker adapts his or her speech according to the demands of the speech situation and the aim of sufficient discriminability or contrast (1990, p. 403). Speech may be hyperarticulated in noisy situations, for new information or for uncommon words, while it might hypoarticulated when the information is old and there is no background noise. The speaker constantly balances the needs of the listener with his or her own wish to economise where possible. Similary, Lindblom, Guion, Hura, Moon, and Willerman (1995) propose that speech perception is characterised by two "modes" of perception. The "what" mode is aided by context and native language knowledge and the focus is on the content, while the "how" mode is not aided by these types of knowledge and instead the focus is on the signal. Lindblom et al. (1995) suggest that both modes can be active

[9]Ohala additionally points out that in this model, this scenario can only explain the initiation of sound changes and not their spread to other members of the speech community.

simultaneously, but it is the "how" mode that provides the variants that might become new pronunciations in a sound change. The main difference from Ohala's model is that misperception or incomplete compensation on the part of the listener is not the primary trigger for sound change, although it may well occur. Instead, Lindblom et al. (1995) argue that listeners, when they become speakers, may choose to adopt novel pronunciations they have heard for various reasons (i.e. they might choose not to compensate for coarticulation).

Tonal sound change

Just as sound changes can occur on the segmental level, as with well-known processes such as /u/-fronting (Harrington et al., 2008) or vowel nasalisation (Beddor, 2009), they can also occur at the suprasegmental level. As we have seen, segmental coarticulation may even affect suprasegmental parameters and phonologise at the suprasegmental level.

Tonal sound changes function similarly to segmental sound changes. Tonogenesis refers to the emergence of tone in a language that was previously non-tonal (Matisoff, 1973). Where a tone system already exists, tone splits refer to the development of two contrasting tones from one pre-existing tone, as in the separation of tones according to vowel length in Yue dialects (Dong, 2014), while tone mergers refer to the neutralisation of a tone contrast such as the distinction between the high level and high-falling tones in Cantonese (Bauer & Benedict, 1997; Matthews & Yip, 2013). Chain shifts can "push" or "pull" tones along in a language's tone space and thereby affect their phonetic characteristics (Hsieh, 2005).

Some types of segmental perturbation have long been suspected to have prompted tonogenesis. According to Hombert, Ohala, and Ewan (1979) and references cited therein, obstruent voicing led to an intrinsic difference in the f0 onset on the following vowel in both South East Asian and African tone languages, in that voiceless onset obstruents generally gave way to tones with higher f0 onsets. In addition, the same authors cite others in linking breathy voiced consonants, implosives and glottal stops with tonogenesis in a number of languages from different families. In particular, tonogenesis as a result of the f0-raising effect of glottal stops and the f0-lowering effect of glottal fricatives on the preceding vowel is variously attested for languages throughout East and South East Asia (Hombert, 1975; Hombert et al., 1979). Hombert et al. (1979, p. 38) propose that, in order for tonogenesis to be phonetically motivated, the following maxims need to apply:

1. the sound change pattern should be attested cross-linguistically;

2. the underlying phonetic mechanism should be an intrinsic side effect of speech (i.e. a natural coarticulatory or aerodynamic effect) and apply to all speakers, independent of their native language; and

3. the phonetic mechanism should be perceptible.

While the intrinsic mechanisms underlying some proposed types of phonetically motivated tonogenesis are not clear, there is overwhelming agreement in the literature that

intrinsic f0 due to onset consonant voicing fulfils all three of the above criteria. Diachronically, similar types of tonogenesis have occurred in languages spoken on different continents. Phonetically, the authors present a number of possible triggers, perhaps the strongest being the so-called vertical tension hypothesis based on intrinsically different larynx height for voiced and voiceless stops. Perceptually, they show that intrinsic f0 differences due to onset voicing in non-tonal languages are indeed audible. Their account of phonetically motivated tonal sound change is in line with that of Ohala's (1981, 1993) misperception account:

> [the] pronunciation intended by the speaker may get distorted by the time it is perceived by the listener - either by the action of articulatory constraints which affect the way the sounds are uttered, or by the action of auditory constraints which affect the way the sounds are analysed by the listener's ear. Since the listener does not have independent access to the mind of the speaker, and thus may be unable to determine what parts of the received signal were intended and what were not, he may intentionally reproduce and probably exaggerate these distortions when he repeats the same utterance. Thus an intrinsic perturbation will come to be used extrinsically (Hombert et al., 1979, p. 37).

With regard to vowel-intrinsic f0 due to vowel openness, however, many have argued that this type of intrinsic f0 is unlikely to lead to tonal sound change, because there is little evidence of this type of change diachronically (Hombert, 1975; Maddieson, 1976; Hombert, 1977b; Hombert et al., 1979). Vowel openness has been linked with tonogenesis and other types of tonal sound change in Hausa (Pilszczikowa-Chodak, 1972), Ngizim (Schuh, 1971, cited in Hombert et al., 1979), Maninka (Spears, 1968, cited in Hombert et al., 1979), Fuzhou (Chen & Norman, 1965, cited in Hombert et al., 1979), Omei (Cheung, 1973, cited in Hombert (1977b)), Lahu (Matisoff, 1973, cited in Hombert et al., 1979), Passamaquoddy (Nicholas & Francis, 1988, cited in Whalen, Gick, & LeSourd, 1999), Kanazawa (Nitta, 2001) and Tupuri (Odden, 2010). However, most of these links have been criticised, and not without good reason. For some languages, it appears more likely that the tone influences the vowel and not the reverse, while for others, tone height and vowel openness seem to stand in complementary distribution to each other.

The main explanation as to why consonant-induced intrinsic f0 has been known to trigger tonal sound change but vowel-intrinsic f0 has not is based on the differing types of resulting perturbation as well as different perceptual processing mechanisms for each. As noted above in Section 1.1.2, Hombert (1977b) suggests that perceptual normalisation for vowel-intrinsic f0 (in the form of intrinsic pitch) results in the effect being largely inaudible to listeners, while consonant-induced intrinsic f0 is not compensated and thus is able to be (mistakenly) parsed as tone[10]. In addition, some authors have argued that because

[10]There is evidence, however, that consonant-induced intrinsic f0 is indeed audible to listeners. f0 height at vowel onset is known to be a cue to onset consonant voicing, especially when primary cues such as VOT or glottal pulsing during the closure phase are masked or neutralised (e.g. Francis, Ciocca, Wong, & Chan, 2006; Winn, Chatterjee, & Idsardi, 2013).

vowel-intrinsic f0 is an inherent characteristic of the vowel itself and cannot be separated from it, it is perceptually less salient than the effect of consonant-induced intrinsic f0 on the following vowel, which is described as "dynamic" because the effect is separated from its (consonantal) trigger and begins strong and eventually fades out during the course of the vowel (Hombert, 1977b; Hombert et al., 1979; Maddieson, 1976). Either of these explanations would rule out Hombert et al.'s (1979) third maxim that the intrinsic effects be perceptible. On the other hand, Reinholt Petersen (1986) argues that vowel-intrinsic f0 and consonant-induced intrinsic f0 are processed similarly and the effects of both are well above perceptual thresholds. Instead, he argues that the vowel-intrinsic pitch effect found in perception is not compensation for intrinsic f0 but rather fits within the virtual pitch model (Terhardt, 1974; Stoll, 1984).

1.2 The present study: intrinsic f0, intrinsic pitch and lexical tone in Cantonese

Already, we have encountered several paradoxes. Hombert (1977a) and Hombert et al. (1979) argued that vowel-intrinsic f0 is not perceptually salient enough to lead to tonal sound changes in the way that f0 perturbation due to onset voicing has. And yet in his own data, Hombert (1977b, 1977a) shows that vowel-intrinsic f0 is perceptible, and in addition, as Reinholt Petersen (1986) points out, the effect is well above the perceptibility thresholds documented by Klatt (1973) and similar in size to the effects found at vowel onset as a result of obstruent voicing.

Furthermore, despite suggesting that intrinsic pitch is irrelevant for sound change because it is not perceptually salient, Hombert (1977a) and Hombert et al. (1979) nevertheless find that listeners do hear intrinsic pitch and even find it salient enough to use it as perceptual normalisation for intrinsic f0. They argue that as the f0 perturbation due to intrinsic f0 is filtered out of the signal, intrinsic f0 poses no danger to the tone contrasts. According to Hombert et al.'s three maxims for phonetically biased sound change noted above, the only criterion that does not find convincing support is the occurrence of sound changes motivated by vowel-intrinsic f0 cross-linguistically. It may well be that this type of intrinsic f0 does not lead to sound change, but if this is the case, it is not because the two other maxims do not apply. Instead, we argue that there must be another explanation.

Hombert (1977b) and Hombert et al. (1979) point out that the intrinsic pitch effect they find in perception is smaller than the intrinsic f0 effect found in production, and as discussed above, such a mismatch between intrinsic f0 (production) and intrinsic pitch (perception) has been found elsewhere (Fowler & Brown, 1997; Pape, 2009). Furthermore, Reinholt Petersen (1986) attributes intrinsic pitch to general auditory mechanisms such as virtual pitch (see Stoll, 1984), while according to auditory enhancement theory, vowel-intrinsic f0 is not a phonetic bias but rather a method used to enhance vowel contrasts (Kingston, 1992). If this is the case, we argue that "factoring out" the enhancement effect would be counter-productive. In other languages, intrinsic pitch does not seem to occur at all (see

Section 1.1.1). As such, it is unclear whether intrinsic pitch represents compensation for coarticulation, a general processing mechanism, or something else entirely.

Thus, the relationship between intrinsic f0 and intrinsic pitch is not clear and to our knowledge has not yet been investigated in a language which additionally uses phonological f0 and pitch to distinguish tone contrasts. In particular, in light of the reported mismatch between intrinsic f0 and intrinsic pitch, we predicted consequences for tone contrasts along the lines of Ohala's (1981) misperception account or at least for tone processing in human and machine speech.

1.2.1 Motivations

Our primary motivation for this study was to investigate the relationship between intrinsic f0 (production) and intrinsic pitch (perception) in a complex tone language. In particular, if intrinsic pitch reflects compensation for intrinsic f0 but the effect is only incomplete, as in other languages, any intrinsic f0 not compensated for should be parsed as part of the tone. This would almost certainly affect tone perception and perhaps be relevant for misperception accounts of sound change. Alternatively, it is possible that the degree of compensation for coarticulation depends on the (language-specific) context and specifically the cues involved. For complex tone languages, we might expect complete compensation for intrinsic f0 precisely because the phonological tone contrasts would otherwise be endangered, whereas the incomplete compensation for intrinsic f0 in American English vowels as found by Hombert (1977a) and Fowler and Brown (1997) does not pose a threat to any tone contrast.

Further motivation comes from Hombert's suggestion that consonant-induced intrinsic f0 is "dynamic" and thus perceptually more salient compared with vowel-intrinsic f0, which necessarily remains constant with vowel openness. A search of the literature revealed no studies on the intrinsic f0 of dynamic vowels such as diphthongs, which might conceivably be associated with falling or rising intrinsic f0 during the course of the (changing) vowel openness. However, to our knowledge, two studies so far have looked at intrinsic pitch on diphthongs, as outlined in the following. We wondered if this type of intrinsic f0 would be "dynamic" and perceptually more salient.

In the first study, Niebuhr (2004) investigated the effect of open and closing diphthongs in real German words[11] on listeners' perception of falling and rising contours. Niebuhr (2004) created an f0 continuum from falling to rising and superimposed it on all diphthongs. Listeners participated in an ABX experiment in which A was the stimulus from the continuum with the clearest falling contour and B the stimulus with the clearest rising contour. The listeners' task was to decide whether experimental stimulus X sounded more like A or B, thus avoiding any notions of "falling" or "rising". The results showed that the closing diphthongs were perceived as having falling intrinsic pitch and vice versa, analogous to the effect we find for monophthongs.

[11] The closing diphthongs were /haɪ/ <Hai> and /haʊ/ <hau>, while the so-called opening diphthongs, based on r-vocalisation in German, were /hiːɐ/ <hier> and /ʊːɐ/ <Uhr>.

The second study looked at intrinsic pitch in a tone language. In a study on tone identification and confusion in Cantonese perception, Brunelle, Lim, and Chow (2010) identified what they believed was an effect of vowel-intrinsic f0 on the Cantonese tone space in a set of production data. In particular, four of the six Cantonese tones involved extreme overlap in the tone space in production when they occurred on the syllable /jɐu/ compared with when they occurred on the syllable /si/. For /si/ (Figure 1.1), the tongue position for the close vowel would be high and should remain high throughout the vowel, leading to a consistently high intrinsic f0. On /jɐu/ (semi-vowel onset followed by a closing diphthong, Figure 1.2), however, the tongue position would be expected to be high during the close semi-vowel, then sink for the open vowel before rising as the diphthong closes. The corresponding intrinsic f0 pattern would match that of tongue height: falling-rising. Indeed, this pattern seems to be visible for at least three of the tones in Figure 1.2 (specifically, 22 in red, 23 in green and 33 in turquoise).

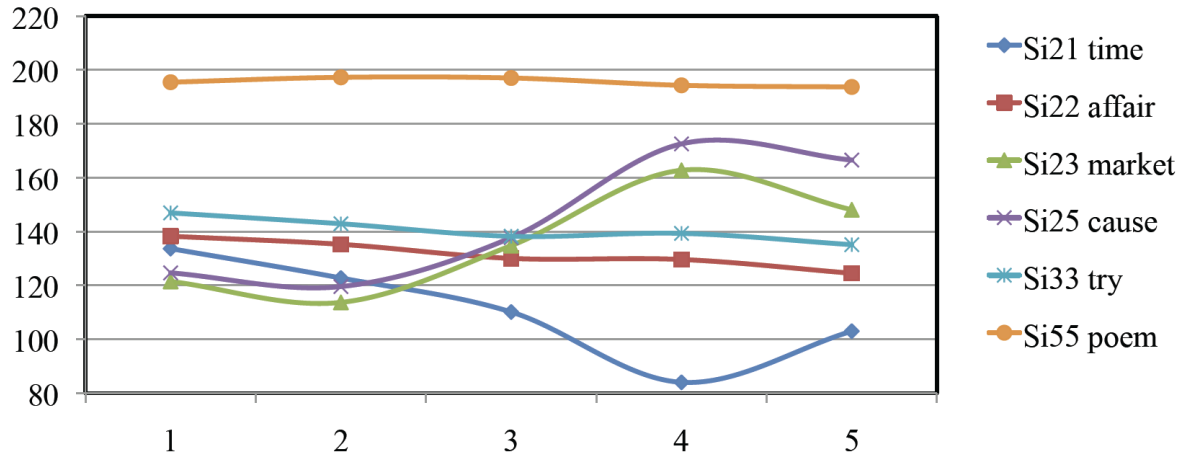


Figure 1.1: Cantonese tone space on syllable /si/, which has constant vowel openness. Characteristically for Cantonese, the tone space is crowded, but all tones are clearly separated. Table from Brunelle et al. (2010, Figure 1).

In addition, Brunelle et al. (2010) found some evidence for intrinsic pitch in perception of these tones. Nevertheless, there was more tone confusion for /jɐu/ tokens than for /si/ tokens, so if intrinsic pitch reflects compensation for coarticulation here, we can only presume that it was incomplete. Unfortunately, this study was relatively small and is described only very generally in the form of a one-page abstract and accompanying poster, so it is not clear where the production data in Figures 1.1 and 1.2 came from. Presumably, they show data from one of the co-authors who produced tokens for resynthesis for the perception experiment described in the abstract. If so, this may not be representative of the wider speech community. In addition, they did not look at a wider range of tokens in order to support the idea that the perception results were explained by intrinsic pitch rather than some other phenomenon (for example, the authors note the probable influence of a lexical effect for certain conditions).

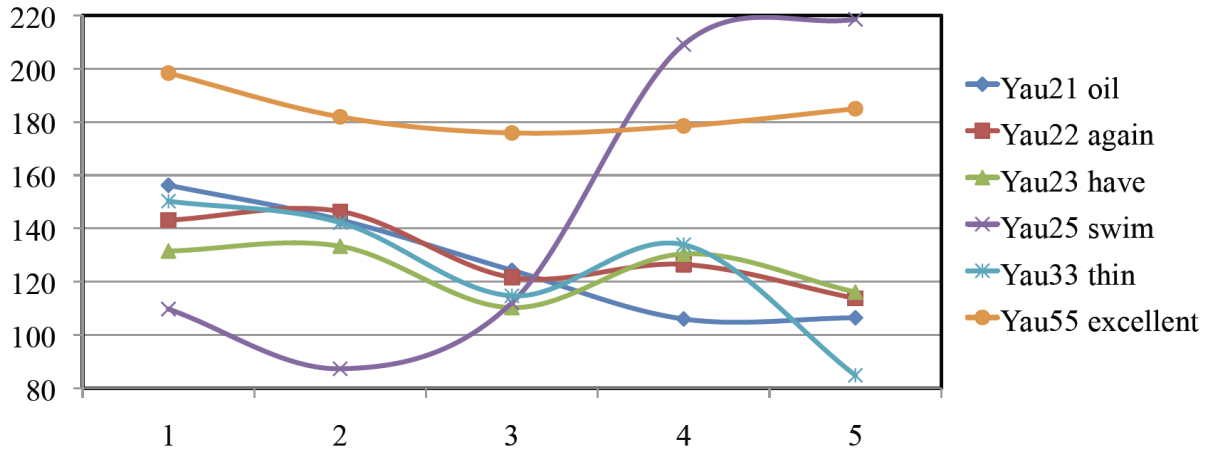


Figure 1.2: Cantonese tone space on syllable /jɐu/, which is characterised by an opening-closing tongue position throughout the syllable and hence might be associated with falling-rising intrinsic f0. Compared with the more neutral syllable in Figure 1.1, there is much more overlap between the tones. Table from Brunelle et al. (2010, Figure 2).

Furthermore, we aimed to fill the gap in the literature on "dynamic" intrinsic f0 on diphthongs in production. Our hope was that we would be able to pick up where Brunelle et al. finished off. If what we know about vowel-intrinsic f0 in monophthongs applies to diphthongs, the effect should transfer. Either way, in order for intrinsic f0 and/or intrinsic pitch not to interfere with tone contrasts, we predicted that the effects should be matched.

Finally, following (conflicting) reports that intrinsic f0 tends to diminish or disappear at the lower end of a speaker's f0 range and on low tones (Whalen & Levitt, 1995; Connell, 2002), we predicted that the f0 distance between contrastive tones would depend on vowel openness. For example, given a simple tone system comprising a high level and a low level tone only, an effect of intrinsic f0 on the high tone would be evident if close vowels had a higher f0 than open vowels on this tone, while the absence of such an effect on the low tone would result in both close and open vowels having the same f0 height. As such, the distance between tone pairs might be predicted to be greater on close vowels than on open vowels if intrinsic f0 decreases at lower f0. Crucially, intrinsic pitch data for the same vowels and tones would be necessary to establish whether such asymmetric intrinsic f0 effects are compensated for accurately.

1.2.2 Aims of this study

The relationship between intrinsic f0 and intrinsic pitch has not yet been examined for a complex tone language. Instead, only single individual aspects of it have been investigated across different languages. Our first aim was to fill this gap in order to establish first the extent to which production and perception are matched and second the relationship between phonetic and phonological determinants of f0.

Furthermore, we aimed to combine and extend the studies by Niebuhr (2004) and Brunelle et al. (2010) by taking a larger set of vowels, as in Niebuhr's study, and replicating his results for a complex tone language, as in Brunelle et al.'s study. Specifically, we were interested in the effects of falling and rising intrinsic pitch patterns on falling and rising tones. Finally, we aimed to investigate the side effect found in Brunelle et al. (2010) that intrinsic f0 is dynamic in speech production and leads to increased tone overlap. By doing so, we would be the first to confirm the presence of intrinsic f0 on diphthongs, and would simultaneously be able to investigate the effects of falling and rising intrinsic f0 on falling and rising contour tones.

As in Brunelle et al.'s study, we chose to use Hong Kong Cantonese for its rich tone inventory where even very slight effects of vowel openness might be predicted to influence f0 and pitch directly and tone indirectly via their acoustic and perceptual correlates. In addition, Cantonese has a large vowel inventory consisting of closing diphthongs as well as a number of opening near-diphthongs (monophthongal nuclei preceded by semi-vowel onsets) with which to investigate dynamic intrinsic f0 and intrinsic pitch.

1.2.3 Hong Kong Cantonese

Cantonese belongs to the Yue family of Chinese languages and is spoken in and around Guangdong and Guangxi provinces and the Special Administrative Regions of Hong Kong and Macau in southern China (see Figure 1.3). While it is related to Standard Chinese, also known as "Putonghua" or "Mandarin", the two languages are mutually unintelligible. Standard Cantonese refers to the variety spoken mainly in Hong Kong and Guangzhou and some other cities in the Pearl River Delta, although even these varieties are somewhat distinct from each other. In this study, we use the term "Cantonese" to refer to the Hong Kong variety of Standard Cantonese (spoken in the region outlined in red in Figure 1.3). Cantonese is often considered a prestige language among Chinese speakers because of the region's (recent) prosperity and special political status (Bauer, 1984; Dong, 2014). Children in Hong Kong learn English and Standard Chinese as foreign languages during their schooling and tertiary education is in English, but the majority of the speech community is monolingual Cantonese in the sense that Cantonese is primary language spoken in the home in Hong Kong (Bauer, 1984). Since the changeover from British to Chinese rule in 1997, Standard Chinese plays an increasingly important role in Hong Kong, and yet Cantonese remains the dominant language. Amongst the participants recruited for our experiments, most speakers reported being more fluent in English than Standard Chinese.

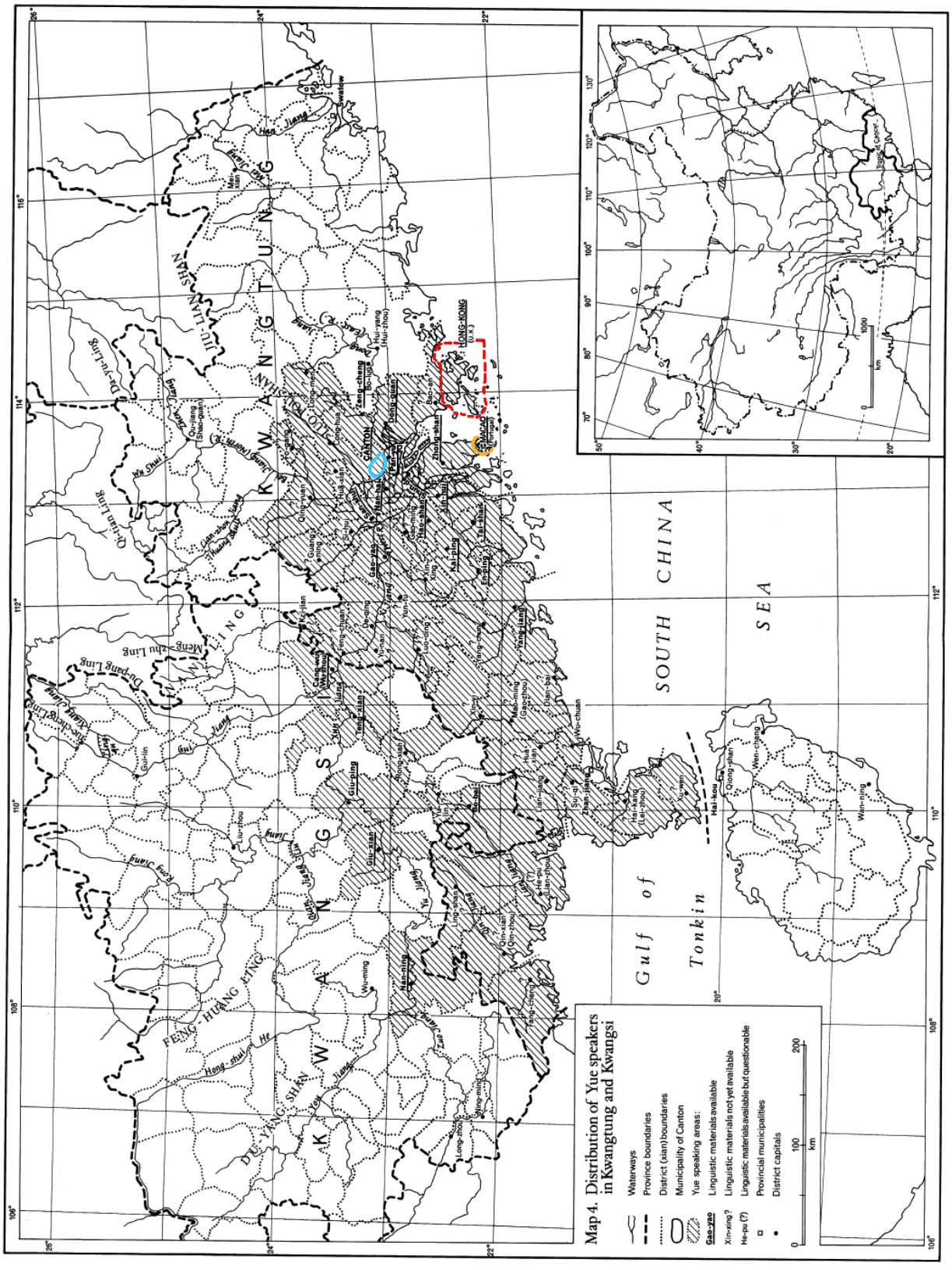


Figure 1.3: Map of Southern China with Yue speaking areas shaded (from Yue Hashimoto, 1972, Map 4). Guangzhou (sometimes referred to as Canton), Hong Kong and Macau are highlighted in blue, red and orange, respectively (our amendments).

Segmental phonology

Cantonese consists of the following phonemic consonants in syllable initial position:

	bilabial	labiodental	alveolar	palatal	velar	labiovelar	glottal
Plosive	p p^h		t t^h		k k^h	kw kw^h	
Fricative		f	s				h
Affricate			ts ts^h				
Nasal	m		n		ŋ		
Approximant			l	j		w	

Alveolar /l/ and /n/ are free variants, with /n/ being the more traditional variant. In addition, syllable initial ŋ is often replaced with a glottal stop.

In syllable final position, the possible phonemic consonants are reduced to the following (note that stops in final position are not released):

	bilabial	labiodental	alveolar	palatal	velar	labiovelar	glottal
Plosive	p̚		t̚		k̚		
Fricative							
Affricate							
Nasal	m		n		ŋ		
Approximant							

A Ø onset or offset is phonotactically legal in both initial and final position.

Cantonese has a rich system of vowel contrasts (Zee, 1999), but it is unclear whether vowel length is contrastive. In most sources, it is presumed to be phonetically conditioned rather than phonologically distinctive, with the exception of /ɐ/ and /aː/, which are presumed to vary both in quality and quantity. The following monophthongs are distinctive in Cantonese:

	front	central	back
high	/iː/, /yː/		/uː/
	/e/ [e] [e^j]	/ø/	/o/ [o] [o^w]
mid	/ɛː/	/ɐ/	/ɔː/
	/œː/		
low		/aː/	

Bauer & Benedict, 1997, pp. 47-48

In this study, we include vowel length (IPA diacritic ː) on the low central vowel only to distinguish it from its mid central counterpart.

In addition, /m/ and /ŋ/ appear alone as syllabic nuclei.

The syllable rime consists of either a nucleus alone (vocalic or syllabic nasal), a vocalic nucleus plus final consonant, or a vocalic nucleus plus /j/, /w/ or fronted /w/ (often

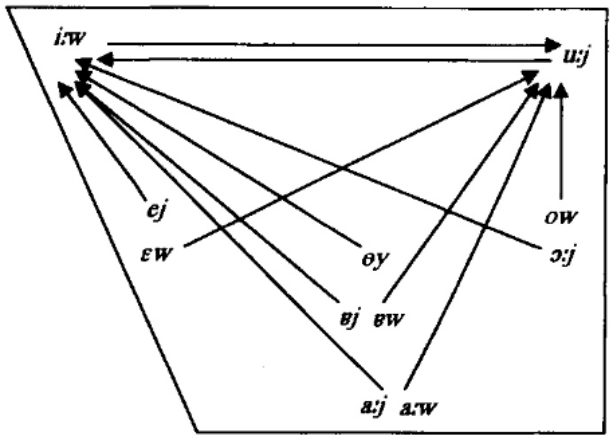


Figure 1.4: Cantonese diphthongs according to Bauer & Benedict, 1997, p. 58.

transcribed as /y/), resulting in one of the eleven Cantonese diphthongs depicted in Figure 1.4. Note that as these are taken to be true diphthongs, our transcriptions will use /u/ in place of /w/ and /i/ in place of /j/ in the final portion of the diphthong.

Thus, the following syllable structures are possible:

- N (nasal syllable nucleus)
- V
- CV
- CVV (where VV is a diphthong)
- VC
- CVC

There are no consonant clusters in Cantonese. In accordance with the majority of the literature, we treat /ts/ and /ts^{h}/ as alveolar affricates and /kw/ and /kw^{h}/ as secondary articulated (labialised) velars.

Suprasegmental/tonal phonology

Cantonese is considered a complex tone system (Maddieson, 2013) with six "full" lexical tones (Guangzhou Cantonese has seven full lexical tones) in unchecked syllables and syllables ending in nasals, as well as three allotones (also known as entering tones or checked tones) in syllables ending in stops. The tones are a feature of the entire syllable rather than, for example, the syllable nucleus, and are superimposed on the syllable onset, nucleus and coda (Kao, 1971, p. 24; see Figure 1.5).

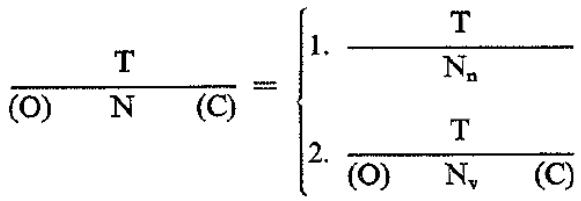


Figure 1.5: The tone (T) is superimposed on the syllable as a whole, including syllable onset (O), nucleus (N) and coda (C), whether the nucleus is vocalic (N_v) or nasal (N_n) (Kao, 1971, p. 24).

	Lexical tones (before m n ŋ & in unchecked syllables)			Allotones (before p˺, t ˺, k˺)			
high register	T1a	52	high falling				
	T1b	55	high level	→	T7	5	high-stopped
	T2	25	high rising				
	T3	33	mid level	→	T8	3	mid-stopped
low register	T4	21	low falling				
	T5	23	mid rising				
	T6	22	low level	→	T9	2	low-stopped

Table 1.1: The Cantonese tones according to their phonological descriptions in the literature (Bauer & Benedict, 1997; Mok et al., 2013). T1a & T1b are still contrastive, at least among older speakers, in Guangzhou but have merged in Hong Kong, where only T1b exists (Bauer & Benedict, 1997, p. 120).

The Cantonese tone system can best be described as displayed in Table 1.1 and Figure 1.6. The tones are described by their Chao tone numerals, already introduced under Tone in Section 1.1.1, the first of which describes the starting level of the tone on a scale from 1 (lowest pitch) to 5 (highest pitch) and the second of which describes the final level of the tone on the same scale. Thus, the Chao tone numerals are able to describe not only level tones but contour tones. Cantonese has three level tones: T1b (tone numerals 55, hereafter referred to merely as T1), T3 (tone numerals 33) and T6 (tone numerals 22). In addition, there are three contour tones: two rising tones, T2 (tone numerals 25) and T5 (tone numerals 23), and one falling tone[12], T4 (tone numerals 21)[13].

Checked tones T7 (tone numeral 5), T8 (tone numeral 3) and T9 (tone numeral 2) are similar to tones T1 (55), T3 (33) and T6 (22) except for their much shorter duration

[12]Many varieties of Cantonese spoken in Guangdong province still have a second falling tone, T1a (52), but this has merged with T1b in Hong Kong Cantonese.

[13]The tone numerals vary slightly from description to description. We chose this system because it best represents both the general consensus and the standard pronunciations.

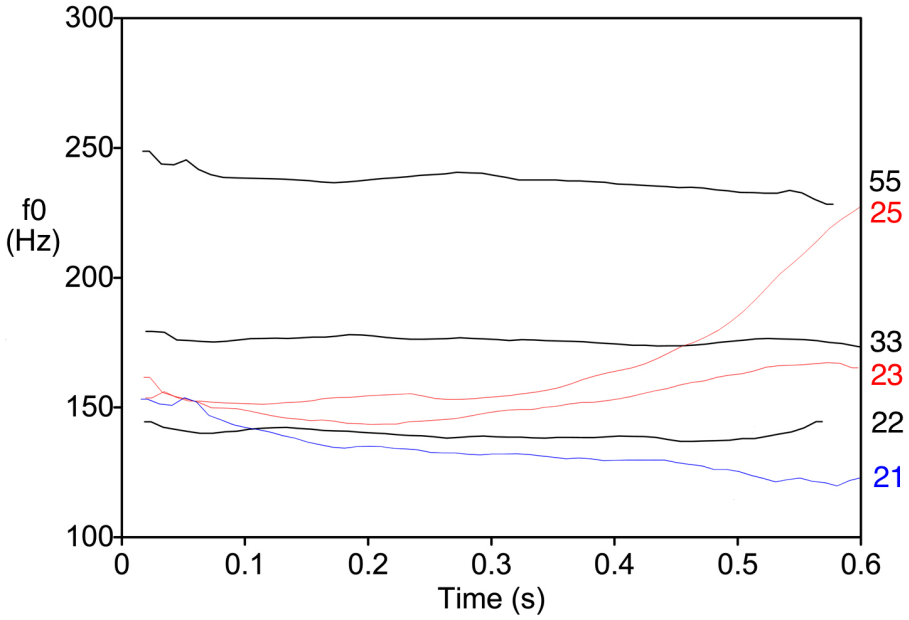


Figure 1.6: The Hong Kong Cantonese tone space according to Mok et al. (2013, p. 343). The tones are colour-coded by tone type: level tones in **black**, rising tones in **red** and the falling tone in **blue**. They are labelled by their Chao letters, i.e. 21 = low-falling, 22 = low level, 23 = mid-rising, 33 = mid level, 25 = high-rising, 55 = high level.

explained by the truncated nature of stop-final checked syllables themselves. Historically, however, tones T7 (5) and T8 (3) are both derived from T1 (55). Many sources consider these three tones to have the same phonemic status as the six full tones described above rather than mere allotones. However, since the checked tones occur *in place of* the level tones in syllables with final stops, they clearly stand in complementary distribution to the full tones. Thus, most modern sources recognise the checked tones as mere variants of the full tones.

Phonetic correlates of tone in Cantonese

Tone is distinguished primarily by f0 modulation in Cantonese, with the exception of the checked tones noted above, which are additionally distinguished by their much shorter duration. Cantonese has both register and contour tones, but the contour tones are distinguished by f0 slope ("direction" in Gandour's terms (1979)) rather than f0 curvature ("contour" in Gandour's terms (1979))[14]. For example, unlike Mandarin tone 3, which is variously described as "dipping" or "falling-rising", the f0 of a Cantonese tone tends to move in one direction only. In a thorough study of both production and perception of Cantonese tones, Khouw and Ciocca (2007) found that the register tones are distinctive

[14]See the description of Tone in Section 1.1.1.

based on average f0 height alone, while the contour tones differ from the register tones in their f0 slope (register tones having flat contours and contour tones having falling or rising contours). Within the contour tones, the rising tones can be distinguished from the falling tone in the direction of f0 slope, while the rising tones can be distinguished from each other by gauging the steepness of the f0 slope (i.e. the magnitude of f0 change during the course of the tone) (Khouw & Ciocca, 2007). The authors argued that these same acoustic correlates are also the most important perceptual correlates of Cantonese tone.

However, in spite of these cues, the Cantonese tone space is very crowded at its lower end (Khouw & Ciocca, 2007). Studies on Cantonese tone perception have identified certain tone pairs that are particularly susceptible to tone confusion: the low level (22) and low-falling (21) tones, the low level (22) and mid level (33) tones, and the mid-rising (23) and high-rising (25) tones (Fok Chan, 1974; Mok et al., 2013). In addition, it has also been suggested that these tone pairs are undergoing sound changes, the most well-documented of which is the merger of the two rising tones (Bauer, Kwan-Hin, & Pak-Man, 2003; Fung & Wong, 2011; Ou, 2012; Mok et al., 2013).

One major caveat applies to both the acoustic and perceptual correlates of Cantonese tones described above. The low-falling (21) tone has long been associated with creaky voice (e.g. Vance, 1977). Recently, Yu and Lam (2014) investigated creaky voice as a cue in tone production and perception in Cantonese. In production, they found creak in over 24% of low-falling (21) tokens in their corpus compared with under 5% overall (for all tones). In perception of natural and synthetic speech, listeners' perception was biased towards the low-falling (21) tone in the presence of creaky voice. Thus, while f0 remains the primary cue for the low-falling (21) tone as for the other Cantonese tones, phonation is an important secondary acoustic and perceptual cue for the low-falling (21) tone.

Due to its irregular pattern, creaky voice often causes problems for f0 algorithms (e.g. de Cheveigne & Kawahara, 2001; Yu & Lam, 2014; Keating, Garellek, & Kreiman, 2015). As creaky voice is common in low f0 regions, and particularly in the Cantonese low-falling (21) tone, when recording speech for this study we chose to isolate the glottal waveform from the rest of the speech signal by carrying out electroglottography (EGG) alongside our audio recordings. With this non-invasive technique, two electrodes are strapped to the subject's throat on either side of the larynx and the changes in impedance between the electrodes measured. Impedance is lowest when the vocal folds are touching (i.e. the glottis is closed). As each glottal cycle consists of an opening and a closing gesture of the vocal folds, the EGG waveform is often considered a more direct and robust measure of f0 and is not subject to artefacts such as pitch-doubling or halving as commonly seen with autocorrelation algorithms based on the speech signal.

A note on Cantonese as a written language

In a study on the phonetics and phonology of intrinsic f0, intrinsic pitch and lexical tone, orthography is hardly important. Nonetheless, a quick note is in order here because orthography was the means by which we elicited tokens from the participants in the production experiments and also represented the response choices in the perception experiments.

The various language groups of Chinese spoken today differ significantly in their historical development from Old and Middle Chinese and the degree to which they are (or are not) mutually intelligible. Today's Standard Chinese, also known as "Putonghua", is a standardised national language based largely on the Beijing dialect of (northern) Mandarin Chinese (e.g. Dong, 2014, pp. 130-151). Standard Chinese orthography is based on this spoken Standard Chinese. Both the Standard Chinese language and orthography came about in the early to mid twentieth century, especially with the rise of the People's Republic of China following World War II, and with it came the simplification of the Traditional Chinese character set predominantly in use up until that time (e.g. Dong, 2014; Snow, 2004). "Simplified Chinese" characters and Standard Chinese became a symbol of the communist revolution and the People's Republic of China. Hong Kong, however, is located in the far south-west of China; far away from the origins of today's Standard Chinese and home to the Yue family of dialects. Hong Kong returned to British control following Japanese occupation during World War II and thus remained culturally and linguistically distinct from the movement on the mainland. Unlike the rest of China, Hong Kong (and Macau) retained the Traditional Chinese character set and remained largely uninfluenced by the Standard Chinese movement.

Today, Hong Kong still uses the Traditional Chinese character set for formal, written Standard Chinese. However, this is not representative of the language spoken in Hong Kong, as the origins of Traditional Chinese go back as far as the Han Dynasty (i.e. around 220AD, also the time of the transition from Old to Middle Chinese) (Dong, 2014, p. 130). Dong (2014, p. 130) compares the use of Traditional Chinese in twentieth-century China with "the use of Latin in Medieval Europe", arguing that the written language was based on an "obselete language, the grammar of which could not be completely grasped and imitated by modern speakers". Thus, over many years, a distinct local variant of the written language emerged based on the Hong Kong Cantonese vernacular (Bauer, 1988; Snow, 2004). This meant not just simple changes such as local pronunciations or phonetic borrowings from existing but lexically unrelated characters, but also adaptation of local grammar and new, unique characters unfamiliar to speakers of other Chinese languages. One lingering effect of the British rule has been the adoption of "English" letters such as <K> and <D> into the written Cantonese character set to represent phonetic similarities shared by Cantonese and English sound systems but foreign to Standard Chinese. While written Cantonese remains unofficial and unstandardised, it is the written correlate of the spoken language and is the language of the local tabloids, magazines and advertisements, is used for emails and text messaging, and is notoriously difficult for outsiders to read (Bauer, 1984, 1988, 2000). While it may be regarded as a "substandard" written form (Bauer, 1984, p. 71), it is a better representation of the colloquial language than standard written forms such as written Standard Chinese or Traditional Chinese. Therefore, in the experiments described in this dissertation, we used characters believed to be most likely to prompt the desired local pronunciations in our participants. For a detailed description and history of written Cantonese, see Snow (2004).

Chapter 2

Experiment 1: Intrinsic f0 in Cantonese monophthong production

While intrinsic f0 is believed to be universal, the size of the effect has been shown to vary by language and, in tone languages, by tone (see Section 1.1.1). To our knowledge, no studies have specifically investigated intrinsic f0 in Cantonese. Thus, in order to observe any interactions between tone and vowel openness, we first needed to confirm the presence of intrinsic f0 in Cantonese and gauge the size of the effect and its robustness across the tone space. We began by looking at the simplest case: open and close monophthongs on level tones. The Cantonese level tones present an interesting opportunity for an interaction with vowel openness, since the three tones are not spread equally across the Cantonese tone space. Hence, some tone contrasts may be more susceptible than others to overlap resulting from intrinsic f0 effects.

2.1 Hypotheses

The first aim of our investigation was to confirm that intrinsic f0 applies to Hong Kong Cantonese and to establish the size of the effect. More importantly, we wanted to compare the magnitude of the intrinsic f0 effect (phonetic) with that of tone (phonological) as a basis for establishing the extent to which intrinsic f0 may or may not interfere with tone.

Specifically, the low (22) and mid (33) level tones are inherently closer together in the Cantonese tone space than the mid (33) and high (55) level tones and are also reported to be more difficult to distinguish (Khouw & Ciocca, 2007; Mok et al., 2013) (see Table 1.1 and Figure 1.6 in Section 1.2.3). This is supported by a possible tone merger in progress involving the low (22) and mid (33) level tones (Mok et al., 2013). Depending on the amount of f0 variation caused by differing vowel openness, and hence the importance of establishing the size of the intrinsic f0 effect, it is possible that the tone spaces of the low (22) and mid (33) level tones actually overlap to some extent while the mid (33) and high (55) level tones do not. This might help to explain the instability and lack of salience in the low (22) and mid (33) level tone contrast as well as in the low tone region in Cantonese

in general[1].

In addition, previous research suggests that the phonetic effect of vowel openness on f0 may be restricted to the upper ranges of a speaker's f0 and thus no longer apply to vowels spoken on low tones (Whalen & Levitt, 1995). If intrinsic f0 is indeed asymmetric in this sense, we would expect the low (22) and mid (33) level tones to encroach on each other's f0 space even further on open vowels due to dampening of the intrinsic f0 effect on the low (22) tone. An f0-lowering effect of open vowels on the mid (33) level tone but not the low (22) level tone could add further confusion to an already difficult contrast by decreasing the distance between the low (22) and mid (33) level tones on open vowels only.

To summarise, we therefore predict:

H1 an effect of vowel openness, in that close vowels have a higher f0 than open vowels,

H2 f0 overlap between the low (22) and mid (33) level tones, but not between the mid (33) and high (55) level tones,

H3 more f0 overlap between the low (22) and mid (33) tones on open vowels than close vowels.

2.2 Method

2.2.1 Participants

The experiment was a repeated-measures design in which all participants took part in all conditions. We recruited 20 native speakers of Cantonese (6 males) born and raised in Hong Kong and aged between 22 and 42 (mean age 29 years). All participants were living in Munich at the time of the recordings with the length of residence ranging between 1 month and 14 years (median length of residence 4.25 months) and were recruited primarily by word-of-mouth and postings to community Facebook groups. In addition, all participants reported being an active part of the local Cantonese community with their main language of everyday communication being Cantonese. While most participants had varying degrees of fluency in English and Mandarin, all grew up in monolingual Cantonese homes (and Cantonese remained their dominant language). Participants were paid in return for taking part in the experiment.

2.2.2 Stimuli

In order to quantify the effect of intrinsic f0, that is the effect of vowel openness on f0, we chose two typical close vowels, /i/ and /y/, and two typical open vowels, /ɐ/ and /aː/[2] from

[1]For example, the low level (22) and low-falling (21) tones are often also reported to be difficult to distinguish perceptually and may be undergoing a merger; see Mok et al. (2013).

[2]While Cantonese generally does not distinguish between short and long vowels, there is an exception for the low vowel pair ɐ/ and /aː/. In addition, /aː/is associated with a slightly more open vowel quality than /ɐ/, perhaps due to the greater amount of time available for proper jaw opening.

the Cantonese vowel inventory. To investigate the effect of vowel openness on tone, and in particular at both high and low f0 ranges, we included the three level tones: high (55), mid (33) and low (22). Based on these four vowels and three tones, we chose a near-minimal set of real Cantonese words all beginning with a voiceless unaspirated alveolar stop and ending in a nasal coda (with alveolar place of articulation as far as possible). By doing so, we aimed to keep coarticulatory effects from neighbouring segments to a minimum. The full set of stimuli is listed in Table 2.1.

	open vowels		**close vowels**	
	/ɐ/	/aː/	/i/	/y/
high level tone (55)	tɐŋ 燈	taːn 丹	tin 顛	tyn 端
mid level tone (33)	tɐŋ 凳	taːn 誕	tim 店	tyn 鍛
low level tone (22)	tɐŋ 鄧	taːn 但	tin 電	tyn 段

Table 2.1: Phonological transcriptions of the stimuli with the Traditional Chinese character used to prompt them. For English translations, see Appendix A.1.

2.2.3 Procedure

Participants were instructed to read out the stimuli at a comfortable speech tempo and volume while sitting comfortably. They read out typical Hong Kong Cantonese sentences during audio and EGG calibration in order to familiarise themselves with the settings and avoid overly hyperarticulated speech in the first few experimental recordings. The stimuli were presented in Cantonese orthography as discussed in Section 1.2.3 to participants via wall-mounted monitor running recording software SpeechRecorder (Draxler & Jänsch, 2004). The stimuli were recorded at a sampling rate of 44 100Hz and bit depth of 16 with one channel audio and one channel electroglottography (EGG). Audio was recorded via a head-mounted microphone (Beyerdynamic Opus 54) and EGG via a Laryngograph Processor. Each stimulus was recorded ten times in isolation and in fully randomised order, resulting in 120 tokens per participant. The first ten participants were recorded in the presence of a native speaker of Cantonese in order to ensure correct pronunciation. As these participants had no difficulty correctly interpreting the Chinese characters and pronouncing the stimuli as intended, the final ten participants were recorded without the native speaker present (and correct tone pronunciation was monitored by a phonetically trained, non-native speaker and subsequently confirmed by plotting f0 for each tone and ensuring clear separation of all three tone categories).

2.2.4 Post-processing

Segmentation and labelling

The audio channel was automatically segmented and labelled using forced alignment system WebMAUS (Kisler, Schiel, & Sloetjes, 2012).

EGG and f0

We extracted the fundamental frequency from the EGG waveform by bandpass filtering the signal and then calculating the short-term zero crossing rate. The original EGG waveform includes noise in the form of micro and macrofluctuations (such as swallowing) outside the frequency range for typical glottal vibration. In order to remove these artifacts, the signal was first filtered with a pass band of 30Hz-400Hz and smoothing of 1Hz (i.e. the width of the transition between pass and stop) using a Praat script (Boersma & Weenink, 2012). The short-term zero crossing rate of the filtered signal was then calculated in Emu (Harrington, 2010) with a window size of 25ms and shift of 5ms. The resulting signal is interpreted as the fundamental frequency of glottis vibration used for our analyses.

2.2.5 Analysis

Statistical analysis was carried out in EmuR (Harrington, 2010; R Development Core Team, 2011).

For f0 analysis, we separated the voiced from voiceless portions by querying all segments which were phonologically voiced. For this stimuli set, our tokens for analysis therefore comprised the rime but not the voiceless onset of the spoken utterance. However, as noted below, these portions may include small devoiced or voiceless portions depending on the phonetic nature of the token or slight errors in WebMAUS' calculation of the boundaries.

f0 was z-score normalised by speaker in order to remove speaker-specific effects such as gender. As segmentation and labelling was carried out automatically and WebMAUS is accurate only to approximantely 10ms, we risked having small voiceless portions included in our data at the beginning and end of each token. In addition, we wanted to avoid including consonant-induced raising of the f0 onset, even if this was consistent across all tokens. Thus, we time-normalised the data by labelling the starting point of each token as 0 and the end point as 1 and selected only the central 80% of each token for analysis. In doing so, we hoped to exclude f0 perturbations at the boundaries of our tokens.

We next checked pronunciation by tone and segment for each speaker individually in order to identify and eliminate any mispronunciations or errorful f0 tracking. This process revealed two sources of noise that would be problematic for our data. Firstly, speakers VP13 and VP15 consistently produced /tyn$_{33}$/ as /tyn$_{22}$/[3]. As a prerequisite for being included in our final dataset, we required each speaker to clearly separate all tones and not systematically vary them on the basis of vowel openness. These mispronunciations would be misleading in our final dataset in that they would indicate lowering of the mid tone on closed vowels. Thus, all repetitions of /tyn$_{33}$/ from speakers VP13 and VP15 were excluded. Secondly, the f0 of speakers VP19 and VP20 was unable to be correctly tracked using the same parameters used for all other speakers; thus, these speakers were also excluded from analysis. The final dataset therefore contained all data from 18 speakers with the exception of /tyn$_{33}$/ for speakers VP13 and VP15.

[3]Our informant ruled out confusion based on the Chinese prompts, so it is possible that these two speakers are merging candidates for this tone contrast; see Mok et al. (2013).

f0 analysis

To establish whether vowel openness had an intrinsic f0 effect on our data, we ran a mixed-effects model with the speaker and time-normalised f0 contours as the dependent variable, Tone (low (22) vs. mid (33) vs. high (55)) and Vowel Openness (open vs. close) as fixed factors, and Speaker and Item as random factors using the `lme4` package in `R` (Bates, Mächler, Bolker, & Walker, 2015; R Development Core Team, 2011).

Machine classification

In addition, in order to estimate how accurately the tones could be separated based on f0, we carried out quadratic discriminant analysis in the `MASS` package for `R` (Venables & Ripley, 2002). Our larger goal was to investigate which tones were misclassified most often, and if these misclassifications could be attributed to an influence of vowel openness. Using the same f0 data as above in Section 2.2.5, we took the mean f0 over all time points and repetitions resulting in one measurement per speaker and per condition. Training and testing were carried out using a "round robin" method (e.g. Watson & Harrington, 1999); that is, training on 17 of the 18 speakers and testing on the eighteenth, iteratively, until all speakers had been tested. The classifier was provided with the real tone identity for training but not the vowel information. The result of the test phase is a three-alternative tone categorisation (low (22), mid (33) or high (55) level tone) for each f0 value. When the tone categorisation based on the f0 value matched the tone intended by the speaker, we labelled this "correct classification"; otherwise, classification was considered "incorrect".

As above, we ran a mixed-effects model with Classification (correct vs. incorrect) as the dependent variable, Tone (low (22) vs. mid (33) vs. high (55)) and Vowel Openness (open vs. close) as fixed factors, and Speaker and Item as random factors using the `lme4` package in `R` (Bates et al., 2015; R Development Core Team, 2011). Note that Tone and Vowel indicate the underlying qualities of the tokens as produced, not as classified.

2.3 Results

2.3.1 f0 analysis

We first checked mean f0 contours for every combination of vowel and tone. Figure 2.1 confirms that, for all tones, f0 increases at least marginally with vowel height (i.e. /i/ and /y/ are higher than /ɐ/, which in turn is higher than /aː/). This justifies our decision to group /i/ and /y/ together as close vowels and /ɐ/ and /aː/ together as open vowels for our analysis as depicted in Figures 2.2 and 2.3. Additionally, for the high (55) level tone, it appears that rounded vowel /y/ is somewhat higher than unrounded vowel /i/[4].

[4]To our knowledge, there are no documented intrinsic effects of lip rounding on f0. Interestingly, however, Pape, Mooshammer, Fuchs, and Hoole (2005) found for German and Catalan that /yː/ required a higher f0 in order for it to sound equal to /iː/ in pitch perception. This is exactly what we would expect for our Cantonese data if listeners normalise for higher f0 on rounded vowels in speech perception.

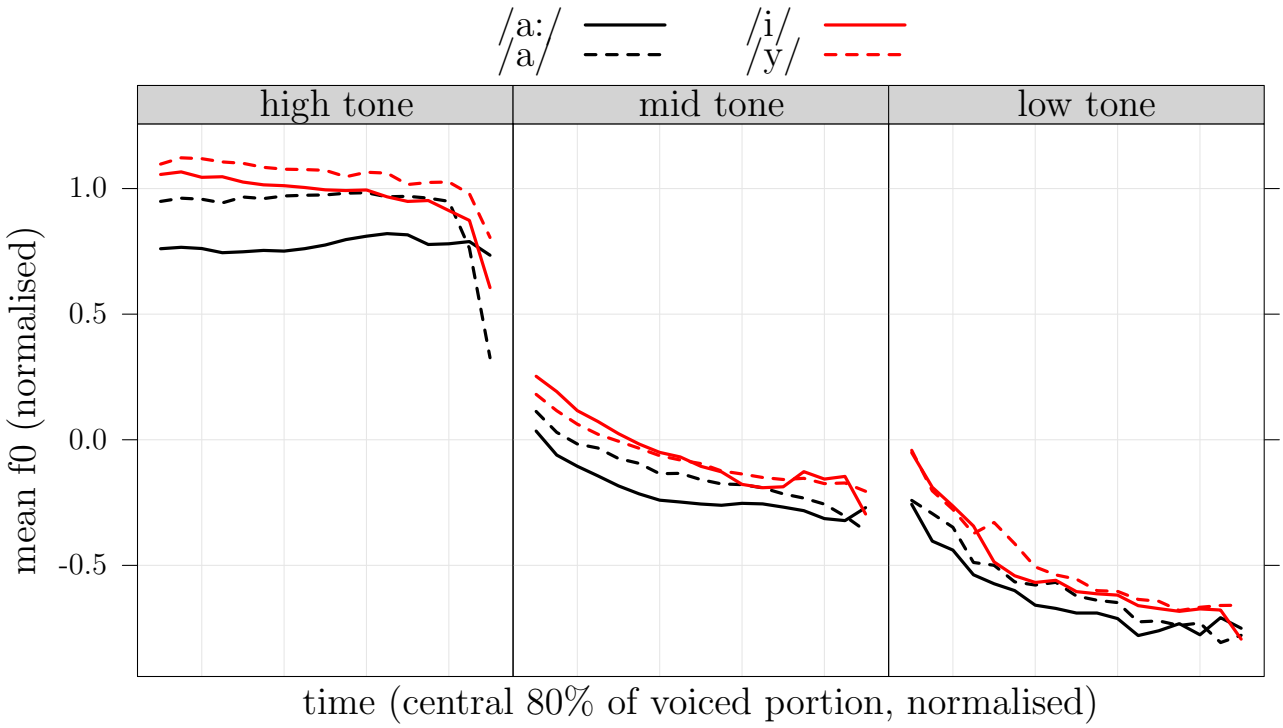


Figure 2.1: Central 80% of normalised mean f0 contours by tone and vowel over all speakers. The colours reflect the tone height, while solid lines represent close vowels and dashed lines open vowels.

Figure 2.2 shows the central 80% of the speaker and time-normalised f0 contours separately for each combination of tone and vowel openness, collapsed over all speakers and repetitions. While tone (represented in colour) is clearly the strongest determinant of f0 height, close vowels (solid contours) indeed appear to have a higher f0 than open vowels (dashed contours). The effect of vowel openness appears to apply to all three tones, although the effect is somewhat greater on the high (55) level tone, as reported in the literature. Figure 2.3 shows the exact data entered into statistical analysis; that is, the mean f0 height from all time points shown in Figure 2.2 separately by tone and vowel openness. The same effect is visible here along with the between-speaker variability (one point per speaker per condition). The effect appears to be weakest on the mid (33) level tone.

The model found main effects of both Tone ($\chi^2[7] = 37872, p < .001$) and Vowel Openness ($\chi^2[7] = 7.5, p < .01$) on f0 with a significant interaction between Tone and Vowel Openness ($\chi^2[9] = 50.6, p < .001$). Post hoc Tukey tests revealed a clear effect of Vowel Openness on tokens spoken with the high (55) level tone ($z = -4.2, p < .001$), a very weak effect of Vowel Openness on the low (22) level tone ($z = -2.6, p = .07$) and no effect of Vowel Openness on the mid (33) level tone ($z = -2.2, p = .2$), despite the numerical

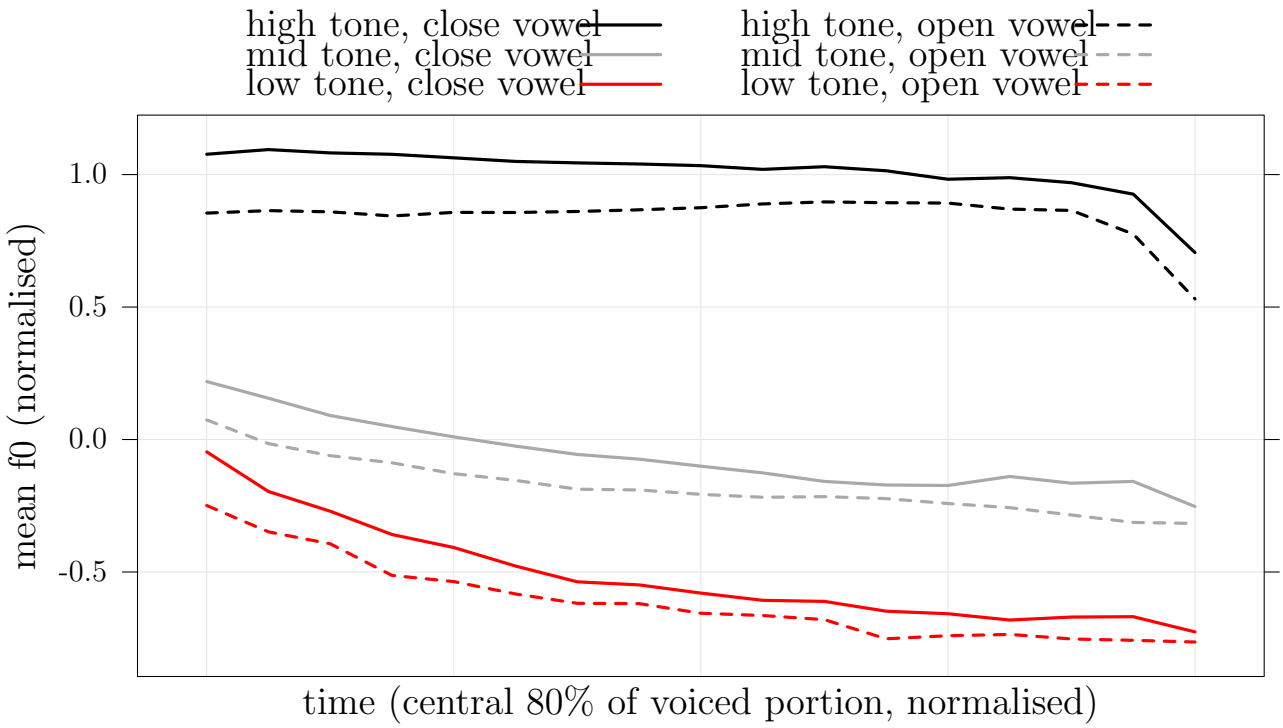


Figure 2.2: Central 80% of normalised mean f0 contours by tone and vowel openness over all speakers. The colours reflect the tone height, while solid lines represent close vowels and dashed lines open vowels.

trend present for this tone in Figure 2.2. All tone comparisons were significant at $p < .001$, but it is worth noting that they differed substantially in their z values[5]. For the low (22) vs. mid (33) level tone contrast, $z = -51.8$ on open vowels and $z = -36.7$ on close vowels, while for the mid (33) vs. high (55) level tone contrast, we found $z = -122.8$ on open vowels and -98.3 on close vowels, indicating a far stronger effect of Vowel Openness on the latter tone contrast.

2.3.2 Machine classification

Figure 2.4 shows the results of the quadratic discriminant analysis. Each box represents classification of its respective tone separately for each degree of vowel openness (left vs. right). The box colouring displays how each tone was categorised (by tone). White indicates that the tone was classed as the low (22) tone, light grey as the mid (33) tone and

[5]i.e. the test statistic in a normal distribution which divides the regression coefficient in a logistic regression model by its standard error in order to determine whether the coefficient is different from zero (Field, Miles, & Field, 2012, pp. 318-319).

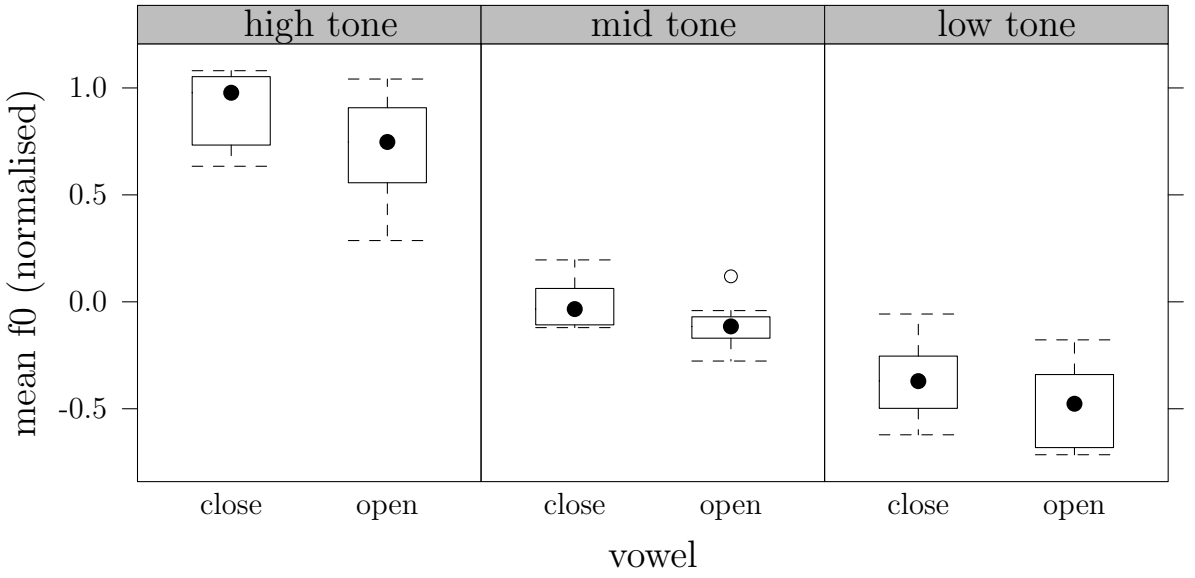


Figure 2.3: Normalised mean f0 height separately for close and open vowels and high (55), mid (33) and low (22) level tones. There is one point per speaker per condition.

dark grey as the high (55) tone. As confirmed in Figures 2.2 and 2.3, the low and mid level tones (22 and 33, respectively) are closer together in the tone space than the mid and high level tones (33 and 55, respectively). This is clearly reflected in the high proportion of correct classifications for the high tone (55, right box) compared with the much lower rate of correction classification for the low (22, left) and mid (33, centre) tones. However, it is not just the close proximity of the low (22) and mid (33) tones that leads to confusion: correct classification depends on vowel openness, at least for the low (22) and mid (33) tones. The low tone (22) appears more likely to be misclassified (as the mid tone (33)) when it is spoken on a close vowel than on an open vowel. To the contrary, the mid tone (33) appears more likely to be misclassified when it is spoken on an open vowel than when it is spoken on a close vowel. Interestingly, when the mid tone (33) is misclassified on close vowels, it is still more likely to be classified as its nearest tone neighbour, the low tone (22), even though close vowels would be expected to raise this tone in the direction of the high tone (55). Thus, the effect of vowel openness on f0 is not so strong as to outweigh the effect of the close proximity between the low (22) and mid (33) level tones.

The model confirmed a main effect of Tone ($\chi^2[7] = 130.5, p < .001$) but not of Vowel Openness ($\chi^2[7] = 0.3, p = .6$) on the tone classification with a significant interaction between Tone and Vowel Openness ($\chi^2[9] = 30.3, p < .001$). Post hoc Tukey tests revealed a clear effect of Vowel Openness on the low (22) ($z = -4.1, p < .001$) and mid (33) ($z = 3.4, p < .01$) tones but not on the high (55) tone ($z = 1.7, p = .6$) (see full results of

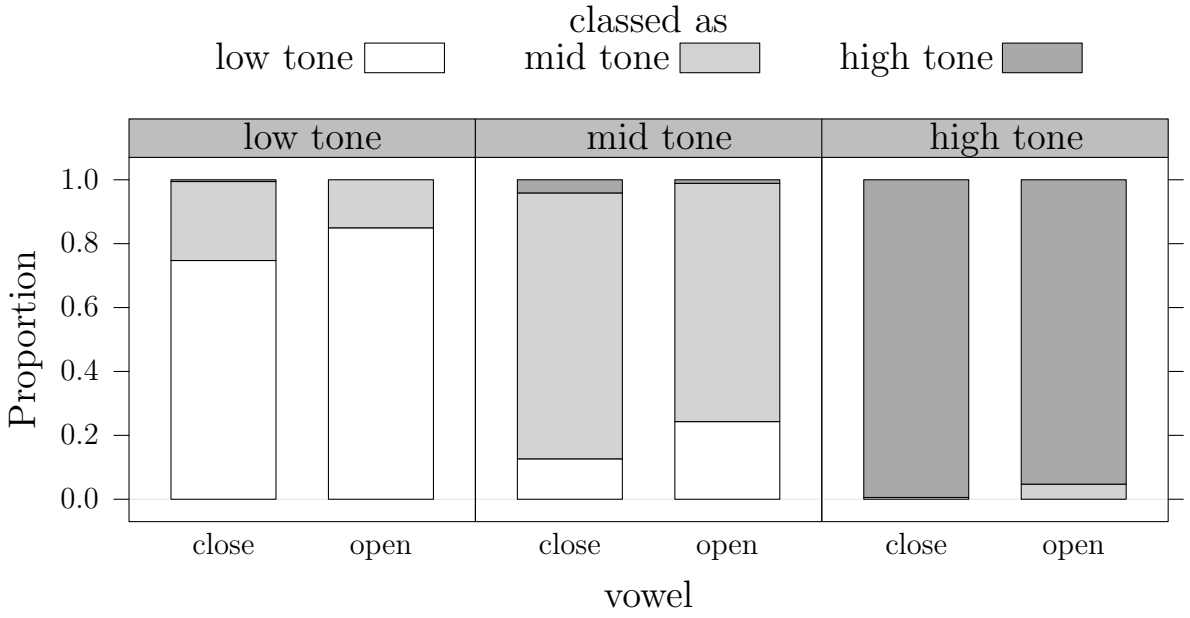


Figure 2.4: Classification of the tones based on f0 after training using tone information only. The individual panels represent the actual tones on which training was based. The boxes and their colouring reflect how the tones were (mis)classified: white indicates that the tone was classed as the low (22) tone, light grey as the mid (33) tone and dark grey as high (55) tone. Separate boxes for close vs. open vowels show the effect of vowel openness on classification.

post-hoc comparisons in Table 2.2).

2.4 Discussion

Naturally, as the phonological determinant of f0, tone was the strongest predictor of f0 height. Numerically, however, vowel openness was also correlated with f0 height, although this was statistically significant for the high tone (55) only. For machine classification of the same data, we found the reverse effects: successful tone categorisation depended on vowel openness for the low (22) and mid (33) tones only, not for the high tone (55).

We discuss the results separately for each of our hypotheses.

H1: Close vowels have a higher f0 than open vowels

Our data supports the presence of intrinsic f0 in Cantonese with a main effect of vowel openness in Section 2.2.5. However, what appears to be a clear trend towards higher f0 on close vowels for all tones in Figure 2.2 is statistically significant on the high (55) tone

	z value	p value
open.high : close.high	1.655	.562
close.mid : close.high	6.311	$< .001$
open.mid : close.high	9.856	$< .001$
close.low : close.high	9.821	$< .001$
open.low : close.high	5.760	$< .001$
close.mid : open.high	4.681	$< .001$
open.mid : open.high	8.203	$< .001$
close.low : open.high	8.166	$< .001$
open.low : open.high	4.106	$< .001$
open.mid : close.mid	3.405	$< .01$
close.low : close.mid	3.361	$< .05$
open.low : close.mid	-0.634	.988
close.low : open.mid	-0.049	1.0
open.low : open.mid	-4.097	$< .001$
open.low : close.low	-4.054	$< .001$

Table 2.2: Post hoc Tukey tests for all combinations of Tone and Vowel Openness

only. As such, our results are compatible with the majority of other studies concluding that intrinsic f0 exists, but is diminished at lower f0 values (Whalen & Levitt, 1995; Connell, 2002). Nonetheless, vowel openness significantly affected the ability of our machine classifier to correctly categorise the low (22) and mid (33) level tones. We interpret this to mean that despite the non-effect in the f0 analysis, intrinsic f0 plays a small but important role even in the lower regions of a speaker's f0. In fact, it is possible that a larger study with more statistical power would result in a small but significant effect of intrinsic f0 on all tones. In our view, the size of a phonetic effect (presuming it exists) is less important than the environment in which it occurs. This idea is discussed in more detail below in the discussion of H3. In Cantonese, four different tones share the low end of the tone space: the high-rising (25), mid-rising (23), low-level (22) and low-falling (21) tones all share an f0 onset at about tone level 2, resulting in a particularly crowded low tone space. It is conceivable that the intrinsic f0 effect is somewhat suppressed in the lower f0 ranges of languages like Cantonese in order to avoid endangering tone contrasts in an already crowded tone space. This is predicted by Connell (2002) and to a certain extent Diehl and Kluender (1989) and Kingston and Diehl (1994), who argue that feature contrasts can be exaggerated (or suppressed) in order to enhance phonological contrasts.

H2: The low (22) and mid (33) tones overlap, while the mid (33) and high (55) tones do not

Our data clearly reflect the assumption implicit in the Chao tone numerals and descriptions in the literature that the low (22) and mid (33) tones are indeed closer in proximity than the mid (33) and high (55) tones. This is depicted in Figure 2.2. When both within and

between speaker variation is included in these contours, it is particularly clear in Figure 2.3 that the boxplots for the low (22) and mid (33) tones overlap substantially (regardless of vowel), while those for the mid (33) and high (55) tones do not (with the exception of a few outliers). This is also apparent from the z values calculated in the post-hoc comparisons detailed in Section 2.2.5: the f0 difference between the two higher tones was much stronger than the difference between the two lower tones.

These results translate directly into the ability of our machine classifier to categorise the tones by their f0 alone (see Figure 2.4). The well-separated high (55) tone was able to be classified correctly almost 100% of the time irrespective of vowel openness, while the overlapping low (22) and mid (33) tones were misclassified approximately 20% of the time. Mostly, the low (22) and mid (33) tones were mistaken for each other rather than for the more unambiguous high (55) tone.

As a result of the asymmetrical distribution of the level tones in Cantonese, we predicted that the less distinct low (22) vs. mid (33) tone contrast should be somewhat more unstable than the mid (33) vs. high (55) tone contrast. Support for this comes from reports from native speakers as well as in the literature that the low (22) vs. mid (33) tone contrast is indeed more difficult to distinguish (Mok et al., 2013; Ou, 2012) and might be merging. Thus, this contrast should be more susceptible to ambiguity as a result of phonetic variation such as intrinsic f0. Indeed, this was one of the motivations behind H3.

H3: The low (22) and mid (33) tones overlap more on open than close vowels

The aim of this hypotheses was to establish whether or not vowel-intrinsic f0 can lead to increased f0 confusion between two overlapping tones in a complex tone system.

Whalen and Levitt's (1995) comprehensive review suggests there is frequent, but not universal, reduction of the intrinsic f0 effect in the lower ranges of a speaker's f0 as the physiological and/or aerodynamic mechanisms underlying the phenomenon weaken. Our data complies with this hypothesis insofar as the effect of vowel openness on f0 is statistically strongest on the high tone (55) and no longer statistically significant on the low (22) and mid (33) tones (Section 2.2.5 and Figure 2.3).

We predicted more overlap of the low (22) and mid (33) tones on open than close vowels because we expected vowel openness to affect the mid (33) tone and not the low (22) tone. Had this been the case, f0 lowering on the open vowel of the mid (33) tone would have encroached on the (non-lowered) space of the same vowel on the low (22) tone.

Interestingly, while there was no effect of vowel openness on f0 of the mid (33) tone (see Figure 2.3), there was a very weak effect ($p = .07$ in the post hoc comparison) on the low (22) tone in which close vowels were somewhat higher than open vowels. In effect, we see the opposite effect of what we predicted. This is clear in Figure 2.3: there is in fact more overlap between the close vowels of the low (22) and mid (33) tones, and this is confirmed by a lower z statistic for the low (22) vs. mid (33) tone contrast on close vowels ($z = -36.7$) than on open vowels ($z = -51.8$). Thus, it is possible that vowel intrinsic f0 can indeed lead to an asymmetry in the degree of overlap between two tones. This might be better tested in a different tone language in which two similar tones are closer to the

intrinsic f0 effect boundary; that is, for Cantonese ideally something like a mid (33) vs. *mid-high (*44) level tone contrast[6].

If we return to our original aim, that is, to investigate whether vowel openness by itself causes greater tonal overlap, it is worth considering once again the results of the machine classification experiment. In this experiment, we found the reverse effects to those found in the pure f0 analysis: vowel openness had no effect on classification of the high (55) tone but did affect classification of the low (22) and mid (33) tones. Low (22) tones on open vowels were classified correctly more often than low (22) tones on close vowels (which were more often misclassified as mid tone (33)) (see Figure 2.4). In our view, it is likely that the stronger effect of vowel openness on the high tone (55) found in the f0 analysis in Section 2.2.5 has no effect on classification as there is no other competitor in this area of the tone space: there is a ceiling effect for classification of the high tone (55), regardless of vowel. The same cannot be said for the low (22) and mid (33) tones: the overlap between these two tones might be interpreted as a region of uncertainty or instability in which even very weak effects such as those of vowel openness are taken into account in an attempt to classify ambiguous f0 values. This supports the idea of tone as the primary, phonological determinant of f0 and vowel openness as an important secondary, phonetic component of f0.

These results may be interpreted in terms of their consequences for speech perception. Specifically, we might hypothesise that the size of the vowel openness effect in absolute terms is less important than the region in which it occurs. Specifically, the secondary or phonetic f0 components may be less important for tone categorisation, whether by humans during speech perception or by machines in automatic speech recognition, in regions of relative phonological stability/certainty such as the f0 region surrounding the high (55) tone in Cantonese. Instead, it is likely that this phonetic variation is more problematic in regions of phonological ambiguity/instability, such as the overlapping low (22) and mid (33) tones tested here.

[6]where * marks an illegal construction.

Chapter 3

Experiment 2: Intrinsic pitch in perception of Cantonese monophthongs

In Experiment 1 (Chapter 2), we were able to confirm the presence of intrinsic f0 in Cantonese as well as an interaction with tone, in that the effect of vowel openness was minimal in the low tone region. However, because of the close proximity of the low (22) and mid (33) level tones to each other (compared with the high (55) level tone), even minimal effects of vowel openness led to f0 overlap between the tones. These results pose interesting questions for tone processing. To what extent do listeners compensate perceptually for the effects of vowel openness on f0? Do they adjust appropriately for differing effect sizes, and can they separate intrinsic f0 from tonal f0?

The only research to date on intrinsic pitch in Cantonese was conducted by Brunelle et al. (2010) (see Section 1.2.1), but it was designed primarily to investigate the amount of confusion between the six Cantonese tones and the acoustic cues used for tone identification. Brunelle et al.'s research question and findings with regard to intrinsic pitch appear to be a mere side effect of the stimuli they chose for their main research questions. Thus, any implications of the study for our understanding intrinsic pitch should be treated with caution. For one, they compared a close monophthong with a semi-vowel + closing diphthong sequence rather than an open vowel. Additionally, their experiment required that listeners choose between all six Cantonese tones; that is, including the three contour tones. The authors also concluded that there was a confound for some combinations of segments + tone, which differed in lexical frequency. On the basis of this study alone, then, we have some reason to expect an effect of vowel openness in Cantonese, but it is difficult to judge whether the effect is similar in size and distribution (across the tone space) to the intrinsic f0 effects we observed in Experiment 1.

In light of this and the conflicting results in the literature on the universality of intrinsic pitch (see Section 1.1.1), our aim was to confirm whether vowel openness affects pitch perception in Cantonese, and if so, whether it patterns with the intrinsic f0 patterns found in Experiment 1. If it does, this might provide evidence for intrinsic pitch as perceptual

normalisation for intrinsic f0 and against the idea that intrinsic f0 is actively used by the speaker to enhance perceptual contrasts (Diehl & Kluender, 1989; Kingston, 1992, 2007).

3.1 Hypotheses

The larger goal of this experiment was to investigate the effect of vowel openness on listeners' pitch perception. Previously, this question has been asked using metalinguistic tasks in which listeners participated in AB-type experiments, as in Hombert (1977b) and Stoll (1984). However, the relevance of such tasks for real speech processing is not entirely clear. In addition, most previous studies used entirely synthetic and thus unnatural speech, so it is possible that listeners were treating it more like a complex tone than natural speech. Thus, we used natural spoken (laboratory) speech as the basis for our manipulations and subsequent resynthesis of the signal. We are therefore confident of our stimuli being perceived as close to natural speech as currently possible. In Cantonese we were able to make use of the complex lexical tone system to create a more natural speech perception task. We aimed to create a synthetic continuum varying only in the height of its (level) f0 contours, which would be based on the low (22), mid (33) and high (55) level tones in Cantonese. This same continuum would then be overlaid onto two differing vowels, one open and one closed, so that all natural intrinsic f0 effects are eliminated and each vowel at the same step has identical f0. All possible vowel and tone combinations should be real words, so that the listener's task is merely to click on the word they hear (corresponding to the tone they hear). As a result, vowel openness should influence listeners' tone categorisation if they normalise for an (in this case absent) effect of intrinsic f0 in the spoken speech they hear. An effect of vowel openness would thus provide us with evidence for intrinsic pitch in Cantonese.

In light of previous research, we expect the following:

H1 On an f0 continuum between a low (22) and mid (33) level tone or between a mid (33) and high (55) level tone with identical f0 values at each step, the position of the category boundary will depend on vowel openness (= intrinsic pitch). That is, open vowels should sound higher than close vowels.

H2 Because intrinsic f0 is believed to disappear at lower f0 values, and if intrinsic pitch reflects normalisation for intrinsic f0 effects, there should be an interaction between vowel openness and tone region. That is, we expect vowel openness to influence tone categorisation for the mid-high, but not the low-mid category boundary. Alternatively, and in line with the results from Chapter 2 showing a trend towards an intrinsic f0 effect on the mid and low tones, we might also see intrinsic pitch at both category boundaries, but the effect should be greater at the mid-high boundary in line with our production results.

H3 As the low (22) and mid (33) level tones are closer together in the tone space than the mid (33) and high (55) level tones, we would expect more difficulty with the

distinction between low and mid tones than between mid and high tones. This would be visible in the form of flatter sigmoid slopes due to larger periods of ambiguity at the category boundary.

3.2 Method

3.2.1 Participants

Model speaker

In order to create the speech stimuli for our experiment, we required a model speaker of Hong Kong Cantonese. We recruited a 27 year old male speaker born and raised in Hong Kong to native Cantonese speaking parents. This speaker clearly produced and perceived all six canonical tones correctly and was judged by our informant to speak typical Hong Kong Cantonese. He agreed to produce our targets for use in this experiment and was paid for his time. Five repetitions of each of the target stimuli (see Table 3.1) embedded in the carrier sentence "看見X快講出來。" (/hɔn^{33} gin^{33} X faːi^{33} gɔŋ^{25} tsœt^{5} lei^{21}/) [1] were recorded in a sound-treated cabin. The carrier sentence was chosen for its neutral tone context, in that the mid (33) level tone preceded and followed the target. Previous research indicates that context for speaker normalisation is crucial for accurate tone perception in Cantonese (Wong & Diehl, 2003; Francis, Ciocca, & Ng, 2003).

Figure 3.1 shows f0 values for the voiced portion of each combination of vowel openness and tone. Each point represents the mean f0 value across all time points of a single repetition[2].

All tokens from the model speaker were segmented and labelled by hand with particular attention paid to the onset and offset of voicing.

Listeners

We recruited 15 native Cantonese speakers (five male) born and raised in monolingual Cantonese households in Hong Kong but living in Munich at the time of the experiment (length of residence ranging between 1 month and 14 years, median length 6 months). All were active members of the local Cantonese speaking community and reported speaking Cantonese on a daily basis. The participants were aged between 23 and 43 (mean age 30) and reported no history of speech or language disorders. All participants were paid for their participation.

[1] Carrier sentence taken from Gu, Hirose, & Fujisaki, 2004.

[2] Notably, for this speaker, there is a visible effect of vowel openness on f0 height for all three level tones. This effect was removed when creating the perception stimuli as described in Section 3.2.2 below in order to test listeners' sensitivity to the effect of vowel openness on f0.

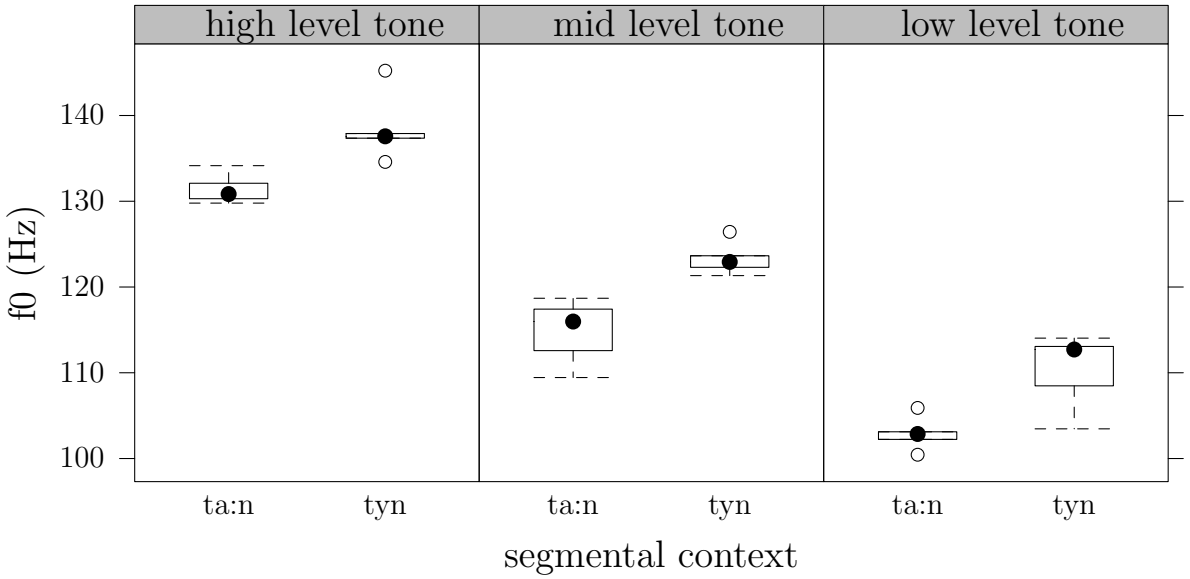


Figure 3.1: Model speaker's mean f0 height (in Herz) by tone and vowel, separately for each of the five repetitions. f0 was measured in the voiced portion of the signal only. Note the clear effect of vowel openness on f0 for all three tones for this speaker.

3.2.2 Stimuli

Targets

For the perception experiment, we chose open vowel /a:/ and close vowel /y/ from the stimuli used in Experiment 2, as we were able to build exact minimal sets with identical onsets and codas and with similar lexical and grapheme frequency across all words. The stimuli and the Traditional Chinese characters used in the perception experiment are listed in Table 3.1.

	open vowel	**close vowel**
	/a:/	/y/
high level tone (55)	/ta:n/ 單 *single, sole*	/tyn/ 端 *beginning*
mid level tone (33)	/ta:n/ 誕 *birth*	/tyn/ 鍛 *exercise*
low level tone (22)	/ta:n/ 但 *but*	/tyn/ 段 *section, piece*

Table 3.1: Phonological transcriptions of the stimuli with the Traditional Chinese character used to prompt them.

Resynthesis and preparation of experiment

The aim was to create an f0 continuum from an unambiguous low (22) level tone via an unambiguous mid (33) level tone to an unambiguous high (55) level tone. To do this, we first needed to select natural stimuli for resynthesis and identify suitable f0 endpoints.

We first analysed the characteristic values of the target stimuli produced by our model speaker. Specifically, we took not only f0 values but also the first and second formants as well as stimulus duration into account in order to create stimuli as typical and as natural as possible. We calculated the median durations of all /taːn/ and all /tyn/ repetitions and shortlisted those individual tokens that were closest to the median in duration. We then plotted the mean F1 and F2 contours of all /taːn/ and all /tyn/ repetitions and, from the shortlisted tokens, took a natural token of each vowel that was closest to the average formant values for its vowel openness. Both tokens happened to be natural utterances of the mid (33) level tone. Our informant confirmed that these tokens were typical with no unusual qualities.

We subsequently created one level f0 continuum with 22 equidistant steps ranging from the lowest to the highest f0 in our sample database and overlaid these contours onto our two natural /taːn/ and /tyn/ tokens using the overlap-add function in `Praat` (Boersma & Weenink, 2012). This resulted in 22 /taːn/ and 22 /tyn/ stimuli varying in f0 height only. Finally, we embedded all stimuli into the same carrier for presentation in the experiment. Embedding the targets in the carrier provided an f0 context within which listeners could frame their tone judgements.

In order to confirm the suitability of the stimuli for our purposes, we carried out a small pilot study prior to beginning the perception experiment. We designed an informal three-alternative forced choice experiment based on the stimuli described above and presented it to our informant, the model speaker and another native-speaking volunteer in order to ensure that all three tones were able to be identified clearly. After some adjustments to our original continuum (by lowering both the lower and upper ends of the continuum), this pilot study was successful in that none of the participants had difficulty identifying any of the three tones in either vowel context, and all participants considered the stimuli to sound natural. The final continuum ranged in f0 from a level 87Hz contour at step 1 to a level 139Hz contour at step 22.

3.2.3 Procedure

Listeners were seated at a computer in a quiet room with high-quality studio headphones (model Beyerdynamic DT770 Pro). They were told that on each trial they would be presented auditorily with the utterance “看見X快講出來。” (/hɔn^{33} gin^{33} X faːi^{33} gɔŋ^{25} tsœt^{5} lei^{21}/) and that word X would be one of two words presented in Traditional Chinese characters on the computer screen. They were instructed to click on the word that was the closest match to the stimulus they heard as fast as possible and then on OK to continue. The stimulus was presented once only.

For steps 1 to 12 of the continuum (hereafter the low-mid continuum), listeners chose

between the low (22) and the mid (33) level tone. All 12 steps on both vowels were repeated 10 times, resulting in 240 tokens per person. The same applied to steps 11 to 22 (hereafter the mid-high continuum), for which listeners chose between the mid (33) and high (55) level tones. The total of 480 tokens was presented in randomised order in six blocks of 80, between which listeners were offered a short break if they wished. Button order was counterbalanced. The experiment took approximately 30 to 40 minutes per listener.

3.2.4 Analysis

As all listeners participated in all conditions, we are dealing with a repeated-measures or within-subjects design. We fit four sigmoids (because the responses were binary) to the responses over the 22 steps: one for each combination of Tone Condition and Vowel Openness. This was done separately for each listener, so that listener could later be included as a random factor in the statistical model. The coefficients of these sigmoids were then extracted for statistical analysis. Coefficient k corresponds to the intercept and m to the slope of each sigmoid, and with the formula $\frac{-k}{m}$ we arrive at turning point u, which corresponds to the category boundary between the two tones in the respective Tone Condition.

Before entering the coefficients "blindly" into a statistical model, we closely examined the data for each listener in order to identify errors that would lead to a poor model fit and/or misleading results. We checked for three types of errors:

A errors These are errors in which listeners were unable to consistently identify one or both endpoints for the Tone Condition in question, resulting in a turning point that is literally "off the plot" or so decentralised that a proper sigmoid fit can no longer be achieved. Some of these listeners may have profited from a wider continuum, while others simply showed no evidence of categorical perception of the two tone categories. Listeners with this error type were removed entirely from analysis for each condition in which the error occurred. This applied to three male subjects and one female subject for the /tyn$_{22}$-tyn$_{33}$/ model, two of these male subjects for the /taːn$_{22}$-taːn$_{33}$/ model, and one of the same male subjects for the /tyn$_{33}$-tyn$_{55}$/.

B errors The steps on the continuum were not narrow enough for these listeners, so that their perception was extremely categorical and less than two steps near the category boundary were considered ambiguous (ambiguous being response proportions between 0 and 1). For these cases, the slopes that are calculated are indefinitely steep and thus meaningless, but accurate category boundaries are still able to be calculated and for this reason listeners with this error type were not excluded from analysis. However, in order to carry out statistical analysis on the slopes themselves (which may be interpreted as a measure of uncertainty), these listeners would need to be removed from the conditions affected (and the resulting slopes would be somewhat flatter than the whole picture if we were able to calculate slopes for listeners with B errors). Nevertheless, indefinitely steep slopes are a result in themselves, as they

show that listeners had little trouble identifying the tone. Table 3.2 in Section 3.3 shows the number of B errors (i.e. infinitely steep slopes) in each condition.

C errors This error occurs when the proportion at one or both endpoints in the unambiguous regions of the continuum either side of the category boundary rises above 0 or sinks below 1. We interpret small deviations from 0 or 1 at these endpoints as "mistakes", in that the listener simply clicked on the wrong response button or did not hear the stimulus correctly and thus was forced to choose by chance. By including these errors in our model, our sigmoids would be flatter than the raw data would otherwise suggest. Thus, this type of error was filtered out by correcting these data points so that the endpoints were unambiguous (proportions of 0 or 1). To be precise, we interpreted a deviation of 10% (or just one repetition) as such a "mistake"; any deviation from an endpoint of 0 or 1 by more than 10% was regarded as reflecting ambiguity in the signal or uncertainty on behalf of the listener rather than an error.

For transparency, plots demonstrating this careful by-listener pre-examination of the data set can be found in Appendix B.

Once we had removed A errors, corrected C errors and taken note of B errors, we ran a linear mixed-effects model with Category Boundary u as the dependent variable and Tone Condition, Vowel Openness and an interaction between the two as fixed factors with by-Subject random intercepts.

3.3 Results

Figure 3.2 shows the results for description purposes only. The plot on the left shows the low-mid condition (the first 12 steps of the f0 continuum), while the plot on the right shows the mid-high condition (the last twelve steps of the f0 continuum). The proportion of low (22, left plot) and high (55, right plot) responses are shown as a function of the step of the f0 continuum on the x-axis. Solid black curves represent open vowels, while dashed red curves represent close vowels. The curves themselves are taken directly from the coefficients of our statistical model above, except that subjects with indefinite slopes (B errors) have been removed from the plot for a better fit (otherwise all conditions in which at least one B error occurred would be infinitely steep). The points show population means of the response for each vowel and condition with those subjects removed whose category boundary was so far removed from the centre that calculation of sigmoids for these speakers was not possible (A errors) but including subjects with indefinite slopes (B errors) and conditions in which mistakes at an endpoint (C errors) were corrected.

The category boundaries for open vowel /aː/ (black solid lines) are left-shifted for both tone conditions, meaning that there were more mid (33) responses in the low-mid condition and more high (55) responses in the mid-high condition than for close vowel /y/ (red dashed lines). This indicates that open vowel /aː/ sounded higher than close vowel /y/. Note, however, that the curves for open vs. close vowels in the mid-high condition

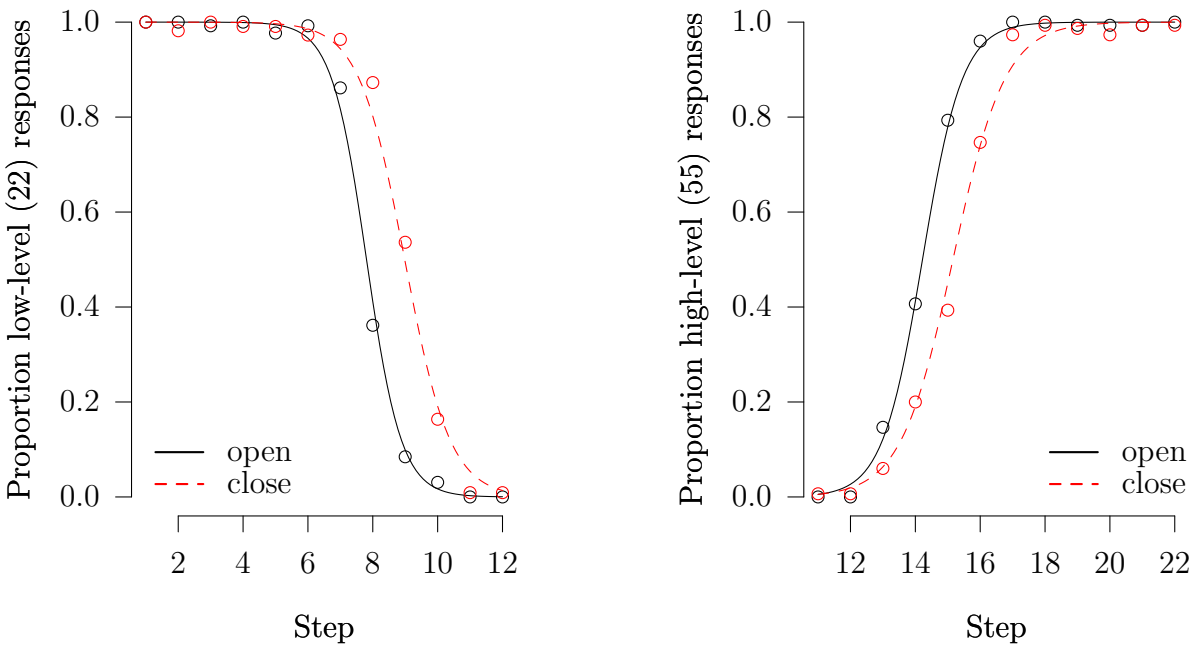


Figure 3.2: Proportion of low-level responses (left) and high-level responses (right) for each step of the continuum, separately for open (solid) vs. close (dashed) monophthongs.

are slightly closer together than in the low-mid condition: hence, the effect of vowel might be somewhat larger in the low-mid condition.

Figure 3.3 shows the speaker means for the location of Category Boundary u (the dependent variable described in Section 3.3) as a function of vowel and tone contrast, and even with the between-speaker variation included in this plot, the pattern remains the same as that shown in Figure 3.2.

As explained above, we can fully rely on the category boundaries but not the slopes of the curves plotted in Figure 3.2. In particular, the slopes for open vowels (solid black curves) in both tone conditions are most likely somewhat steeper than those plotted, as we detected infinitely steep slopes in those conditions listed in Table 3.2.

	open vowel	**close vowel**
low (22) vs. mid (33	13.3%	0%
mid (33) vs. high (55)	20%	0%

Table 3.2: Percentage of participants for whom slopes were infinitely steep in each condition.

Thus, based on the slopes in Figure 3.2 and the distribution of listeners with infinite slopes in Table 3.2, we can conclude that the task was simpler and the categories more clear-cut on open vowels than close vowels. A direct comparison of the tone conditions

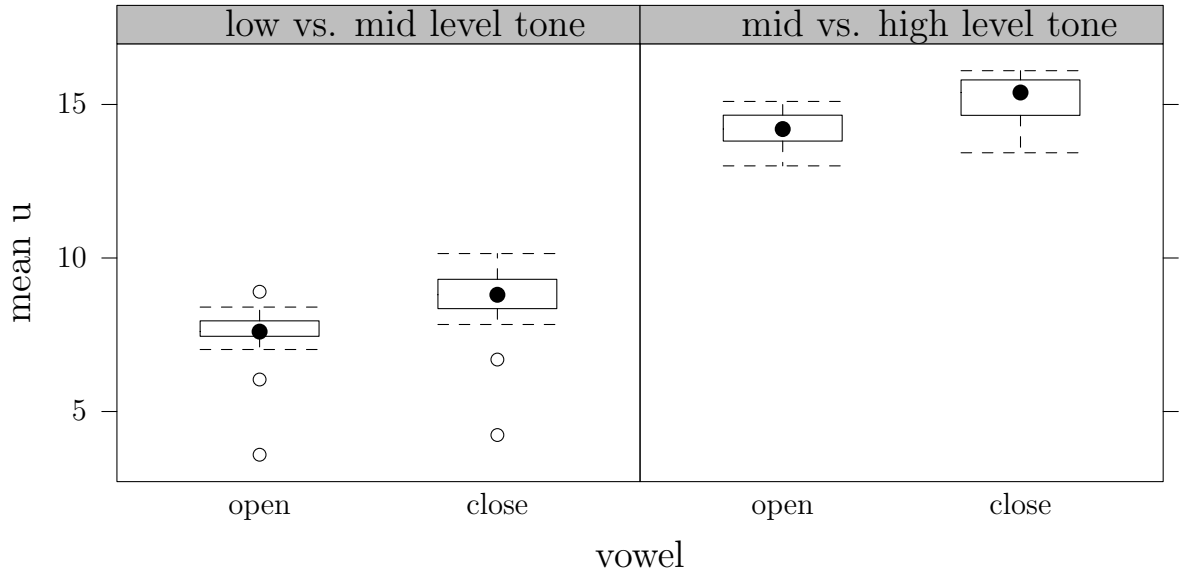


Figure 3.3: Listener means of category boundary u by vowel and tone contrast. u corresponds to the turning point in each of the four sigmoids, i.e. the step of the continuum at which the category boundaries occurred.

is difficult, but it is not obvious that one condition was easier than the other merely by judging the slopes.

However, four subjects revealed A errors for the low-mid condition (compared with just one speaker for the mid-high continuum), as they showed very flat curves (more linear in shape than sigmoid) and an inability to consistently identify either one or both endpoints of the condition. These subjects showed little evidence of categorical perception in the low-mid condition, especially on the close vowel. As such, if these subjects were able to be included, we would expect to see somewhat flatter curves for the low-mid condition (left), especially for the close vowel (red). This indicates that the low-mid continuum, and probably the close vowel, proved a more difficult task for the subjects.

Removing the effect of Vowel Openness from the statistical model was significant ($\chi^2[6] = 19, p < .001$), as was removing the effect of Tone Condition ($\chi^2[6] = 155.3, p < .001$), while removing the interaction between Vowel Openness and Tone Condition was not ($\chi^2[6] = 0.11, p = .74$). However, as we were specifically interested in any interactions between Vowel Openness and Tone Condition, we carried out post hoc Tukey tests. These revealed a weak effect of Vowel Openness on the low-mid condition ($t = 2.8, p < .05$) but not on the mid-high condition ($t = 2.4, p = .09$) (see full results of post-hoc comparisons in Table 3.3).

	t value	p value
close.mid-high : open.mid-high	2.396	.09
open.low-mid : open.mid-high	-16.979	$< .001$
close.low-mid : open.mid-high	-14.226	$< .001$
open.low-mid : close.mid-high	-19.375	$< .001$
close.low-mid : close.mid-high	-16.622	$< .001$
close.low-mid : open.low-mid	2.753	$< .05$

Table 3.3: Results of post hoc Tukey tests for all combinations of Tone Condition and Vowel Openness

3.4 Discussion

Again, Tone Condition was the strongest predictor of response, while Vowel Openness also affected response. While open vowels biased responses towards the higher of the two tones in each tone condition, the effect was only significant for the low-mid condition. As such, we find intrinsic pitch only at the lower end of our tone continuum in Cantonese.

In addition, while we could not analyse the slopes of our models for technical reasons, in Table 3.2 we provided descriptive statistics that may enlighten us as to the difficulty of the task. Open vowels were unambiguous for more listeners than close vowels (B error analysis), while fewer listeners showed proper categorical perception at the low-mid end of the f0 continuum than at the mid-high end (A error analysis).

The results are discussed below in light of our hypotheses.

H1: At exact same f0 values, open vowels should sound higher than close vowels (= intrinsic pitch)

The intrinsic pitch effect, in which open vowels are judged as sounding higher than close vowels at equal f0, is much less robust in the perception literature than the intrinsic f0 found in production and believed to be universal. It has even been proposed by some (e.g. Pape, 2008) that this effect is language-specific rather than universal. In the case of complex tone languages such as Cantonese, we hypothesised that it must be crucial to efficient tone processing that any intrinsic f0 inherent in the signal be filtered out or at least taken note of in perception in order to maintain clear tone contrasts. In agreement with Zheng (2014) for Mandarin, we found an intrinsic pitch effect in a real speech context using a linguistic task. Thus, tone language users must be able to separate the phonological f0 (tone) from the phonetic f0 (vowel-intrinsic f0) in the signal, at least to some extent.

H2: Intrinsic pitch effects should lessen or disappear completely in the low-mid tone condition

In line with the observation that intrinsic f0 lessens or even disappears at lower f0 values (Whalen & Levitt, 1995; Connell, 2002), we predicted that where there is less coarticulation

(vowel-intrinsic f0) there should also be less perceptual compensation for the effect in the form of intrinsic pitch. Thus, we predicted intrinsic pitch to affect the mid-high tone condition to a greater extent than the low-mid tone condition. Analogously, we predicted that if intrinsic pitch affected both category boundaries to the same extent, we would have a case in which perception is misaligned with production and possibly an interesting test case in terms of current sound change theory (cf. Section 1.1.3).

Our results were not in line with these predictions. While there was a trend toward intrinsic pitch in both tone conditions, vowel openness significantly affected the low-mid tone boundary only. At first glance, this may seem counter-intuitive and not in line with compensation for coarticulation. However, this result makes sense in light of the results of the machine classification experiment conducted in Chapter 2, in which it was the low (22) and mid (33) level tones that were most susceptible to confusion despite the smaller (and non-significant) effect of intrinsic f0. As we suggested in the discussion in Section 2.4 of Chapter 2, the close proximity of the low (22) and mid (33) level tones in Cantonese necessitates much more careful perceptual normalisation for vowel-intrinsic f0 than the mid (33) vs. high (55) tone contrast. In this respect, the lack of a statistically significant effect of intrinsic pitch on the mid-high tone boundary is not so surprising, as this well-defined tone contrast is not endangered by the vowel-intrinsic f0 effect.

Thus, we propose that perceptual compensation for phonetic effects is only necessary where it would otherwise endanger a phonological contrast. In a study on the influence of f0 on vowel openness perception, Reinholt Petersen suggests that speech perception is "task-dependent" and

> [...] focuses selectively upon the [segmental or prosodic] level where the acoustic input is ambiguous in relation to the identity of the linguistic categories and/or where the greater importance is attached to the correct categorization (Reinholt Petersen, 1986, p. 40).

Nevertheless, he does not exclude the possibility in real speech situations of ambiguity at both levels that might result in occasional misperception.

Perhaps the most important implication of this proposal would be that each specific case of compensation for coarticulation would require phonological knowledge of the language in question and therefore be language-specific rather than universal. Previous authors, such as Beddor and colleagues, have drawn similar conclusions based on their studies. In a study looking at cross-linguistic production and perception of vowel-on-vowel coarticulation in English and Shona, Beddor et al. (2002) found that listeners showed perceptual normalisation only for types of coarticulation familiar from their native language.

As Beddor et al. (2002) point out, this finding, as well as our finding that compensation for intrinsic f0 only occurs in certain conditions, contradicts the gestural hypothesis posed by motor theorists and direct realists (A. M. Liberman & Mattingly, 1985; Fowler, 1986). According to a gestural account, the listener perceives not the auditory object but the articulatory gestures underlying them and accordingly is automatically aware of and filters out coarticulatory effects. There should be no reason to compensate for some types of

coarticulation and not for others, although Lindblom (1990) hints that compensation for coarticulation may be selective in that it applies only where necessary in relation to word stress patterns.

A further model of pitch perception should be mentioned here. According to the virtual pitch model (Terhardt, 1972b, 1972a; Stoll, 1984), pitch shifts such as vowel-intrinsic pitch are explained by general auditory processing rather than any type of perceptual normalisation or indeed compensation for coarticulation. Crucially, this model predicts that vowel-intrinsic pitch should decrease with decreasing f0. Instead, our data shows that intrinsic pitch affects the lower but not the higher end of the Cantonese level tone space. While the virtual pitch model would not predict this finding, it does not automatically follow that our data discount this theory. For example, it is possible that speech perception is governed by general auditory principles as well as more speech-specific compensation mechanisms where necessary. Indeed, we interpret our findings as evidence that successful speech communication depends on perceptual normalisation for confounds that may be language-specific. Lindblom argues that normalisation patterns may be selective depending on context and optimisation processes (Lindblom, 1990; Lindblom et al., 1987).

H3: The low-mid tone condition should be more difficult than the mid-high tone condition

This hypothesis was based on the greater proximity between the low (22) and mid (33) level tones in comparison to the mid (33) and high (55) level tones. We predicted that the more overlap between any two tones, the more ambiguous the contrast and the more difficult the distinction[3]. According to prevalent sound change theories, it is in these conditions that sound change can occur (Ohala, 1981; Beddor, 2009; Harrington et al., 2008; cf. Section 1.1.3). These theories predict that accurate compensation for coarticulation is necessary in order to avoid misinterpretation of the signal and re-phonologisation according to a new set of cues (sound change).

Unfortunately, as the task we created was too simple for some listeners in some conditions, the slopes calculated in our model are infinitely steep and thus statistically impossible to determine. However, our data point towards more ambiguity in combination with close vowels and as well as less categorical perception at the low-mid end of the f0 continuum.

Four of our 15 listeners showed very gradual slopes and an inability to consistently identify one or both endpoints on the low-mid continuum compared to just one listener for the mid-high continuum. This might be interpreted as indicating more difficulty distinguishing between the low (22) and mid (33) level tones than between the mid (33) and high (55) level tones. On the other hand, for a further two listeners the slopes we calculated for this condition were infinitely steep, indicating that for these two subjects, the task was too simple and the continuum would have benefited from finer steps.

Another method of measuring the level of uncertainty or difficulty by condition is to analyse response times. We timed the duration between the end of the stimulus presen-

[3]Increased ambiguity or uncertainty in such a task is generally visible as greater (within listener) variation in the location of the category boundary on the continuum; that is, flatter sigmoid slopes.

tation to the click on the OK button to continue to the next trial. However, response times based on motor skills such as these are useful only as a method of sorting out extremely long response times (several seconds long) which often reveal that the subject was distracted. For more accurate and informative reaction times, it is better to use the more fine-grained timing associated with, for example, eye-tracking or electroencephalography.

Thus, without better procedures and/or more data, we are unable to draw stronger conclusions regarding this hypothesis.

Chapter 4

Linking production and perception (Experiments 1 and 2)

In the production and perception experiments described in Chapters 2 and 3 we had three primary goals. Firstly, we wanted to confirm that intrinsic f0 applies to complex tone languages such as Cantonese. Secondly, we wanted to establish whether there is perceptual normalisation for the effect in the form of intrinsic pitch. Finally, we aimed to establish to what extent asymmetries in the Cantonese tone space and in the effect of vowel openness may lead to increased overlap between the tones in production and confusion in perception. We hypothesised that any vowel-induced asymmetries in the relationship between speech production and perception could have consequences not just for (human) speech processing but also for speech technology and sound change theory. The aim of this chapter is to hone in on the third and final objective and try to link production (intrinsic f0) and perception (intrinsic pitch).

We briefly review the results of the two experiments before evaluating the production-perception relationship first at the level of the speech community (i.e. across all participants) and then at an individual level.

4.1 Overview of results

For production (Experiment 1, Chapter 2), our f0 analysis showed an effect of intrinsic f0 on all three tones, although statistical significance was only reached on the high (55) level tone. However, due to the inherent overlap between the low (22) and mid (33) level tones in Cantonese, a machine classification analysis revealed that the success of tone categorisation on the basis of f0 alone was significantly affected by vowel openness for the low (22) and mid (33) level tones only. We concluded that more statistical power would probably reveal a significant effect of vowel openness on the f0 of all tones that would increase in size with tone height. Due to the overlap between the low (22) and mid (33) level tones but not the mid (33) and high (55) level tones, we took the machine classification effect as evidence that even a very small intrinsic f0 effect for the lower tones is far more detrimental to this

tone contrast than a larger effect on the non-overlapping contrast.

For perception (Experiment 2, Chapter 3), there was an effect of intrinsic pitch at both tone category boundaries, but it was only statistically significant at the boundary between the low (22) and the mid (33) level tones. At first glance, this result was surprising because a compensation for coarticulation account would have predicted more normalisation at the mid (33) vs. high (55) tone boundary where the intrinsic f0 effect was greater in production. However, and in line with the machine classification experiment in Chapter 2, we reasoned that it was not so much the size of the intrinsic f0 effect that was detrimental to tone perception, but the location of the effect. As the low (22) and mid (33) level tones clearly overlap in production while the mid (33) and high (55) level tones do not, perceptual normalisation in the form of intrinsic pitch should be more important at the low (22) vs. mid (33) boundary than at the mid (33) vs. high (55) boundary.

4.2 Are production and perception aligned?

The central question is how production and perception are aligned: how does intrinsic f0 compare with intrinsic pitch? On the one hand, we found a significant effect of intrinsic f0 on the high (55) level tone while intrinsic pitch was only significant at the low (22) vs. mid (33) level tone boundary. As such, intrinsic f0 and intrinsic pitch do not seem to be well-matched, at least not if intrinsic pitch is to be interpreted as some kind of perceptual normalisation for intrinsic f0. We attributed this paradox to compensation for coarticulation only where a phonological contrast depends on it. On the other hand, if we consider the trends rather than the statistical analysis, we find that there is intrinsic f0 on all tones and intrinsic pitch at both tone boundaries. From this perspective, and in light of the relatively small data set, it is possible that the relationship between intrinsic f0 and intrinsic pitch is more balanced than the results of the statistical analysis would indicate. In order to consider this question in detail, we need a measure of effect size for both production and perception.

As the majority of subjects participated in both experiments, it is possible to compare the results of each speaker-listener more directly. But even so it is not clear what kind of relationship we should expect between intrinsic f0 and intrinsic pitch. The intrinsic f0 values come from the subjects themselves, while the intrinsic pitch values are based on their perception of the speech of just one model speaker. Thus, it is difficult to predict whether the intrinsic pitch values should generalise to speech from other speakers as well. Nevertheless, we will attempt to compare the two values.

In the production experiment in Chapter 2, we carried out our analysis based on speaker-normalised f0 values in Hz. In the perception experiment in Chapter 3, the response variable was turning point u which was equivalent to the step on the f0 continuum (in Hz) at which the listener showed maximal uncertainty, that is, a tone category boundary. As such, it is impossible to carry out a direct comparison without rescaling these values.

We restricted the production and perception data to subjects and conditions for which

we had data from both analyses[1]. For this comparison, we converted the raw production data into semitones with a reference frequency of 1Hz. Then, for perception, we interpolated linearly between the lowest and highest steps on our 22 step f0 continuum and converted the u values first into their equivalent f0 value in Hz and from there into semitones with a reference frequency of 1Hz as for production. For production, we then calculated an average value for each combination of vowel, tone and speaker, and for perception an average value for each combination of vowel, tone condition and listener. In order to compare intrinsic f0 in production with intrinsic pitch in perception, we subtracted the f0 of open vowels from the f0 of close vowels (again, for each speaker and tone or tone condition). We interpreted the f0 difference between close and open vowels in production to be the intrinsic f0, and the difference in perception to be the intrinsic pitch. However, there were three tones in production but, logically, just two category boundaries in perception, so the resulting values still could not be directly compared. Thus, we split the production data into two sets: one set including the low (22) and mid (33) level tones and another including the mid (33) and high (55) level tones. We then averaged the intrinsic f0 values for each speaker in the low-mid set to achieve an intrinsic f0 value that could be compared with the intrinsic pitch value at the low (22) vs. mid (33) category boundary and did the same for the high-mid set. The resulting measures of intrinsic f0 and intrinsic pitch could be matched for each tone combination and participant and are displayed in Table 4.1.

Figure 4.1 displays the intrinsic f0 (production) and intrinsic pitch (perception) values in Table 4.1 side by side, separately for each tone combination. It is immediately clear that there is much more between-subject variation in production than in perception, but this is perhaps not surprising considering that the production data comes from a range of speakers from both sexes, while the perception data reflects perception of just the (male) model speaker. Comparing production and perception, we see that the intrinsic pitch effect in perception is roughly equal to the intrinsic f0 effect in production for the low (22) vs. mid (33) level tones but not for the mid (33) vs. high (55) leveltones. That is, the subjects produced an intrinsic f0 effect that is roughly equal to the intrinsic pitch effect they perceived for the low (22) vs. mid (33) level tone contrast. For the mid (33) vs. high (55) level tone contrast, the intrinsic f0 effect speakers produced was larger than the intrinsic pitch effect they perceived for the model speaker.

If the perception results based on just one model speaker can be generalised to the general population, these results indicate that production and perception are indeed well-matched for the lower contrast but not for the higher contrast. This would explain how it is possible that the overlapping low (22) vs. mid (33) level tone contrast is able to be maintained in spite of intrinsic f0 and intrinsic pitch effects. It also supports our (newly formed) hypothesis that where a phonological contrast is in danger, full compensation for coarticulation is essential whereas otherwise it is not.

Numerous previous studies have shown that compensation for coarticulation is rarely

[1] for example, speakers whose f0 could not be reliably tracked in production were removed from perception as well.

subject	tone combination	intrinsic pitch	intrinsic f0
F1	**low-mid**	0.475	0.696
F10	**low-mid**	0.316	0.649
F2	**low-mid**	0.238	0.75
F3	**low-mid**	0.157	0.452
F4	**low-mid**	0.346	0.833
F5	**low-mid**	0.418	0.293
F6	**low-mid**	0.476	-0.051
F7	**low-mid**	0.506	0.24
F8	**low-mid**	0.231	0.447
F9	**low-mid**	0.255	0.542
M1	**low-mid**	0.307	1.226
M2	**low-mid**	0.472	0.368
M3	**low-mid**	0.128	0.439
F1	**mid-high**	0.443	0.12
F10	**mid-high**	0.479	0.43
F2	**mid-high**	0.584	0.483
F3	**mid-high**	0.257	0.169
F4	**mid-high**	0.626	0.618
F5	**mid-high**	0.221	-0.027
F6	**mid-high**	0.405	0.522
F7	**mid-high**	0.548	0.143
F8	**mid-high**	0.369	0.565
F9	**mid-high**	0.585	0.338
M1	**mid-high**	0.476	1.23
M2	**mid-high**	0.394	0.407
M3	**mid-high**	0.161	0.85

Table 4.1: For each subject, we calculated the average f0 difference (in semitones with a reference frequency of 1Hz) between open and close vowels at each tone boundary for perception (=intrinsic pitch) and for each tone combination in production (=intrinsic f0).

complete, in that the coarticulatory effect is usually larger than the compensatory effect (Lindblom, 1967; Hombert et al., 1979; Lindblom et al., 1987; Fowler & Brown, 1997; Beddor et al., 2002; Fowler, 2005). This certainly applies to the mid (33) vs. high (55) tone contrast in our data, but in this contrast it is not essential that intrinsic f0 be filtered out of the signal because these two tones are inherently better separated. We predicted that incomplete compensation for intrinsic f0 on the low (22) vs. mid (33) tone contrast would have severe implications for tone processing as well as possible consequences for sound change because this tone contrast is less well-defined. Instead, it seems that complete perceptual normalisation is indeed possible where a phonological contrast depends on it as is the case here for Cantonese but not, for example, for American English as reported in Hombert (1977b) and Fowler and Brown (1997). In a shadowing experiment of voiced

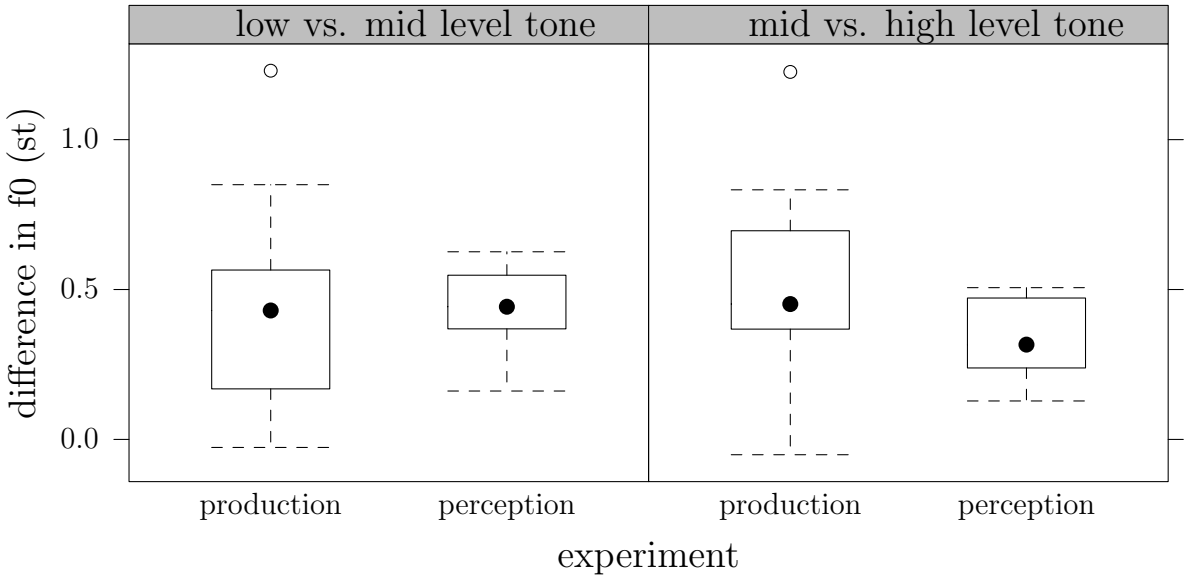


Figure 4.1: The effect of vowel openness on each tone contrast, measured in semitones with a reference frequency of 1Hz for perception and production and calculated from the data in Table 4.1. The y-axis shows the difference in f0 between close and open vowels.

stops in Dutch, Mitterer and Ernestus (2008) found that "[the] phonologically relevant difference between no versus some pre-voicing was imitated, while the exact amount of pre-voicing, which is phonologically irrelevant, was not" (Mitterer & Ernestus, 2008, p. 173). Additionally, they found that response latencies for imitation of two free allophones of the Dutch /r/ phoneme were the same even if the gesture of the token produced differed from that of the stimulus being shadowed. That is, shadowing patterned with phonological principles rather than phonetic detail. The authors argued that this was evidence that production and perception are linked only loosely based on abstract categories and phonological relevance rather than gestural principles along the lines of motor theory or direct realism. Applied to our data, the idea of abstract categories and phonological relevance might account for differing degrees of compensation for coarticulation.

On a side note, we find that mean intrinsic f0 in production over all 13 speakers analysed in this section and over all tones is approximately 0.5 st (range -0.3 st to 1.6 st; standard deviation 0.4 st), which is lower than the valued cited in Whalen & Levitt's (1995, p. 356) meta-analysis suggesting the average across all languages is around 1.65 st (re: 1 Hz). However, this same meta-analysis emphasised the substantial variability between the languages compared (p. 358) and also observed different behaviour between tone and non-tone languages, in that intrinsic f0 generally disappeared on low tones in tone languages (pp. 357-358). The mean intrinsic f0 across all speakers and tones for each of the four

African tone languages in Connell's study lay between 0.3 st and 1.0 st (Connell, 2002, p. 118), which is more in line with our data for Cantonese. Connell tentatively suggests that intrinsic f0 in tone languages is generally smaller than Whalen and Levitt's 1.6 st value.

4.3 Individual variation

A further question often asked in connection with compensation for coarticulation and sound change is the extent to which production and perception are aligned within each speaker-listener (Beddor, 2015). In other words, do subjects who coarticulate more also compensate more? With the same caveats as above, we can also attempt to answer this question based on the data in Table 4.1. Plotting intrinsic f0 (production) against intrinsic pitch (perception) results in the scatterplots in Figure 4.2 in which each point represents one speaker-listener.

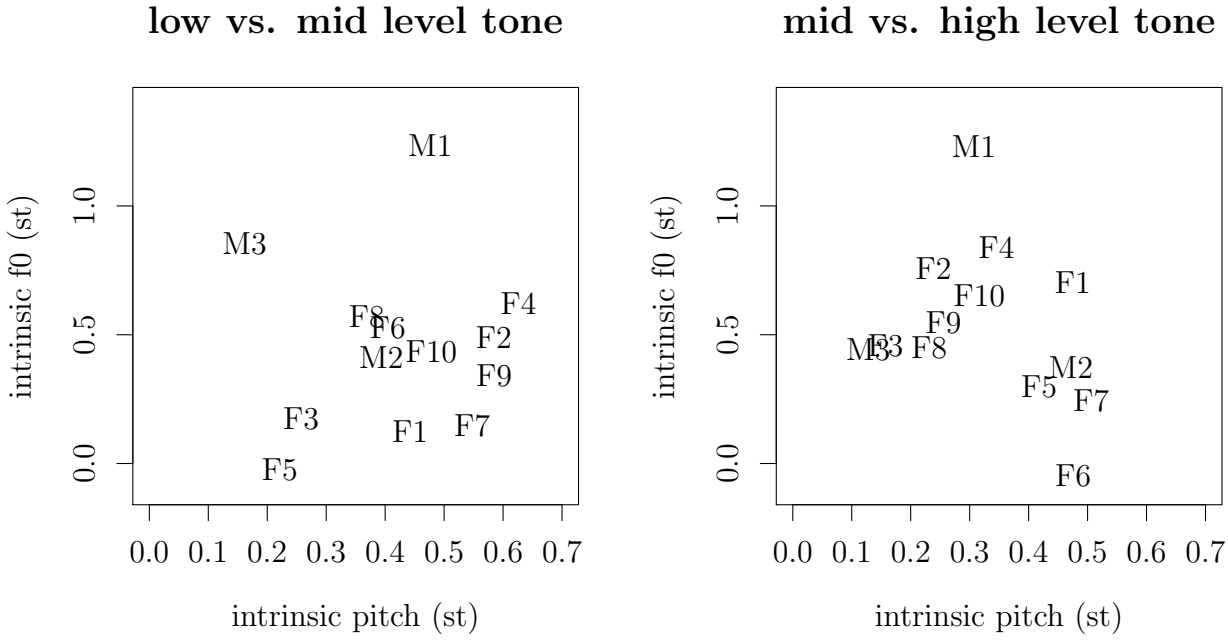


Figure 4.2: Intrinsic pitch on the x-axis as a function of intrinsic f0 on the y-axis, both in semitones, for the low (22) vs. mid (33) tone contrast on the left and the mid (33) vs. high (55) tone contrast on the right. Each point is a participant average and is labelled with the participant's code number (F=female, M=male).

There is no clear pattern in these plots. The variation between participants is very large, and most participants are not even consistent between the tone combinations. For example, for the mid-high tone combination, participants M3 and F3 both produce low to

moderate intrinsic f0 but show only minimal evidence of hearing intrinsic pitch. For the low-mid tone combination, however, while their intrinsic pitch perception remains roughly the same, participant M3 produces a very large intrinsic f0 effect while participant F3 produces almost none. Thus, at an individual level the results look very different, and production and perception are not so well-aligned. It remains to be seen what patterns might emerge in a similar study in which perception data is based on a number of different model speakers. Based on this data, we cannot assume that compensation and coarticulation are correlated at the individual level. This is in line with previous results from Kataoka (2011), who found no correlation between production and percetion of /u/-fronting in American English at an individual level, but contrasts with preliminary evidence from Beddor (2015) showing an individual link between production and perception of coarticulatory vowel nasalisation in American English as well as of voicing and f0 in emerging tonogenesis in Afrikaans[2].

[2]Beddor's work on Afrikaans was carried out together with Andries Coetzee and Daan Wissing; see Coetzee, Beddor, and Wissing (2014).

Chapter 5

Experiment 3: Intrinsic pitch in perception of Cantonese diphthongs

Just as monophthongs can be described in terms of vowel openness, so can diphthongs. Closing diphthongs refer to a vowel in which there is dynamic movement from an open vowel such as /a/ to a close vowel such as /i/ within one segment. Similarly, opening diphthongs are characterised by dynamic movement from a close vowel to an open vowel tongue position.

In a perception experiment using real German words, Niebuhr (2004) found falling intrinsic pitch on closing diphthongs and rising intrinsic pitch on opening diphthongs. This finding extended our knowledge of intrinsic pitch as a static phenomenon on monophthongs (high vs. low pitch) to the more dynamic realm of diphthongs (rising vs. falling pitch). In Experiment 2 (Chapter 3), we found some evidence of intrinsic pitch on Cantonese monophthongs. We asked whether dynamic intrinsic pitch on diphthongs would affect perception of contour tones in a language such as Cantonese.

We expect a dynamic intrinsic pitch effect on diphthongs to be problematic in Cantonese because of its rich tone inventory including a four-way minimal set of tones beginning at tone level 2 and ending at tone levels 1 (low-falling), 2 (low level), 3 (mid-rising) and 5 (high-rising) (see Figure 1.6 in Section 1.2.3). Among this set of tones, even very slight changes in pitch may be contrastive and interpreted as such by the listener.

Thus, the aim of this investigation was to replicate Niebuhr's study for Cantonese to see if dynamic intrinsic pitch based on diphthongs applies to languages with contour tones. Niebuhr's study used a psychoacoustic but rather non-linguistic task in which listeners participated in an AXB-style experiment. He created a continuum between rising and falling f0 and superimposed this on closing and opening diphthongs. Listeners had to judge whether token X (a random token from the continuum) sounded more like token A or token B in order to establish whether X sounded more "rising" or "falling". However, this type of task may not transfer to real speech perception. With Cantonese, we had the advantage of being able to create a more speech-like task by using the linguistic tone categories as responses, whereby listeners should perceive a tone contrast rather than judging tokens' auditory similarity. We chose to look at the lower end of the Cantonese

tone space described above because of the high density of tones differing only slightly in their contours.

5.1 Hypotheses

When presented with closing and opening diphthongs on an f0 continuum from low-falling (21) to mid-rising (23) via low level (22) tones, earlier decision boundaries (i.e. more level and mid-rising responses) for opening diphthongs compared to closing diphthongs would be evidence for the dynamic intrinsic pitch effect described by Niebuhr. In addition, if intrinsic pitch is a form of perceptual normalisation for intrinsic f0, and intrinsic f0 also diminishes with decreasing f0 (Whalen & Levitt, 1995), then intrinsic pitch should also diminish with decreasing f0.

In short, if intrinsic pitch of diphthongs is dynamic (i.e. changes with vowel openness), we would predict the following patterns:

H1: On an f0 continuum between the low-falling (21) and low level (22) tones, there should be more low-falling responses for closing diphthongs than for opening diphthongs;

H2: On an f0 continuum between the low level (22) and mid-rising (23) tones, there should be more mid-rising responses for opening than closing diphthongs; and

H3: A smaller effect of intrinsic pitch at the low-falling (21) vs. low level (22) tone boundary might suggest that intrinsic pitch is perceptual compensation for intrinsic f0 rather than a general auditory effect.

5.2 Method

This experiment was much like the one described in Chapter 3 for monophthongs. Because of the transient nature of the local Cantonese population in Munich, it was no longer possible to recruit the same participants as in the previous experiments. In addition, it is important to note that this particular study was also designed to accompany a larger event-related potentials (ERP) experiment[1]. As such, listeners first participated in the ERP study, in which they heard selected stimuli from the set described below repeated over and over while they watched a silent film in a passive oddball paradigm. After a short break, the same participants then carried out the experiment described in this chapter. As such, they were already familiar with the model speaker's voice and the lexical items involved in the experiment. In addition, as they had already participated in the ERP

[1]ERPs are electrical potentials (or responses) in the brain resulting from an event such as presentation of a stimulus. They are measured using electroencephalography, which is a non-invasive technique in which multiple active electrodes are attached to the scalp with a conductive gel in order to measure electrical activity in the brain. Reference and ground electrodes are used to filter out artifacts such as static electricity on the participant or mechanical potentials caused by muscle movements. For a detailed introduction, see Luck (2014).

study, repetitions in this behavioural experiment were kept to a minimum (details below) in order to prevent further fatigue. In this chapter, we describe the behavioural experiment only, which was primarily the work of the author, and not the ERP experiment, which was the work of a larger group of collaborators.

5.2.1 Participants

Model speaker

The model speaker used for the first perception experiment described in Chapter 3 was no longer available for recording when this experiment was carried out. Thus, we contacted a 37 year old female native speaker known to us from previous experiments. This speaker was born and raised in Hong Kong to native Cantonese speaking parents[2], was judged by our informant to be a typical speaker of Hong Kong Cantonese and clearly distinguished all six tones in production. She agreed to produce the target stimuli for use in this experiment. Ten repetitions of each target and some unrelated fillers were produced in isolation.

Listeners

Ten listeners aged 21 to 30 (3 male) were selected from the local Hong Kong Cantonese community as for the previous experiments. Again, subjects had no known history of speech or language disorders.

5.2.2 Stimuli

As for the other experiments, all stimuli were real and frequent words in Cantonese.

As depicted in Section 1.2.3, Cantonese has a large inventory of closing diphthongs, including both short and long diphthong pairs such as /ɐu/ and /aːu/. The longer diphthongs show greater dynamic vowel movement and thus were particularly well-suited to our research questions.

Unfortunately, however, Cantonese has no opening diphthongs in the traditional sense of a vowel whose articulatory configurations and acoustic patterns change gradually between vowel onset and offset. What Cantonese does have, however, are semi-vowel + open monophthong sequences which are articulatorily and acoustically similar to opening diphthongs (such as those used in Niebuhr (2004)). As approximants, semi-vowels have a very close articulatory configuration, similar to high vowels, and exert considerable coarticulatory influence on surrounding monophthongs in the form of long formant transitions (similar to the dynamic movement in diphthongs). To our knowledge, no studies have investigated intrinsic f0 or intrinsic pitch on semi-vowels, but due to their acoustic and articulatory similarity with high vowels and in line with the tongue-pull theories of intrinsic f0 described in Section 1.1.1, we hypothesise that their intrinsic f0 must be similiar (i.e.

[2]The speaker spent five years of her childhood in Sweden and spoke some Swedish, but we judged her Cantonese to be unaffected by this.

high). It follows, then, that intrinsic pitch would be low, just as it usually is for close vowels.

We therefore chose semi-vowel + open monophthong sequences to mimic opening diphthongs for this experiment. Thus, while our opening diphthongs might not be traditional diphthongs consisting of a monophthongal nucleus with semi-vowel onset, the intrinsic pitch effects of this segmental pattern should be the same as if we were able to use true opening diphthongs.

Targets

We used opening and closing diphthongs in f0-neutral consonantal contexts as listed in Table 5.2.2.

	closing diphthong	**opening diphthong**
	/aːu/	/waː/
mid-rising tone (23)	/maːu/ 牡 *peony*	/waːn/ 輓 *to mourn*
low level tone (22)	/maːu/ 貌 *appearance*	/waːn/ 幻 *illusion*
low-falling tone (21)	/maːu/ 矛 *spear*	/waːn/ 還 *to return*

Resynthesis and preparation of experiment

The basis for our prosodic manipulations was a token of our model speaker producing /maːu$_{22}$/ in which the word was pronounced clearly and the tone contour matched the average low level (22) tone contour of all this speaker's low level tone productions. Figure 5.1 shows a spectrogram of this natural token including overlaid formant and f0 contours.

Using the overlap-add function in Praat Version 5.3.42 (Boersma & Weenink, 2012), we resynthesised the pitch contour as follows. First, time markers were set at three time points in the original token: the beginning, centre (50%) and end of the token. Then, based on our analysis of the model speaker's speech, we set the f0 onset at the first marker to 160Hz, the f0 turning point at the second marker to 140Hz, and then created ten different f0 offsets spaced equally between 105 and 175Hz for the final marker. Ten different contours were then created by linearly interpolating between the f0 onset and the f0 at the turning point as well as between the turning point and the f0 offset[3]. Figure 5.2 represents schematically the f0 contours used to resynthesise the 10-step continuum, while Appendix C depicts spectrograms of all manipulated stimuli including overlaid f0 contours.

[3]Some confusion may arise as to how the resulting level and rising contours can be described as such due to the initially falling f0 in the first half of the stimuli. f0 contours in other studies on Cantonese as well as data from our own model speakers show that this initial "dip" in f0 is completely normal. Level tones are known to have a somewhat negative slope due to declination and the same pattern can be observed both in the natural base tokens used resynthesis displayed in Figures 5.1 and 5.3 as well as in the production data in Figures 6.1 and 6.4 in the next chapter. This pattern was perceived as most natural to our two native-speaking informants mentioned below, so we feel that our decision to design our synthetic f0 contours this way is well justified.

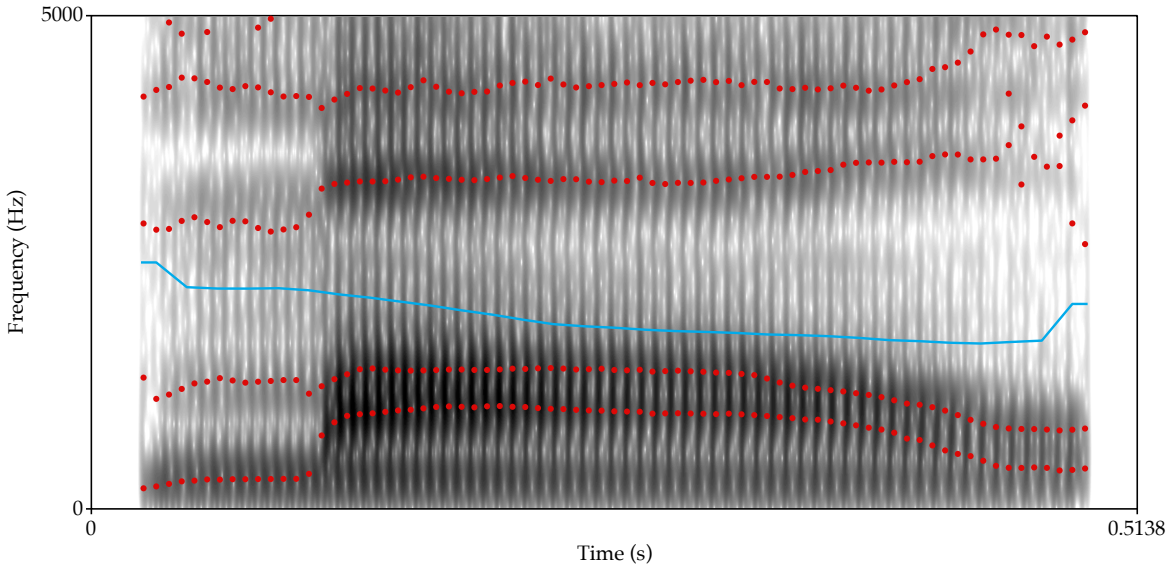


Figure 5.1: Model speaker's natural production of /maːu$_{22}$/ used as the basis for resynthesis. The red dotted lines reflects Praat's estimation of formants, while the solid blue line demonstrates the f0 contour.

The exact same process was carried out for /waːn/ using the same f0 values and same (relative) time markers, so that we had two identical f0 continua with the endpoint ranging from 105 to 175Hz (i.e. continua from low-falling (21) to mid-rising (23) via low level (22)) superimposed on two differing vowel patterns. A spectrogram of the natural token of /waːn$_{22}$/ used for resynthesis is displayed in Figure 5.3[4].

Each of these steps (on both continua) were played to two native speakers of Cantonese, who were asked for their judgements on the best instances each of the low-falling (21), low level (22) and mid-rising (23) tones. Independently of each other, our informants judged the first step of each continuum to be the most natural production of the low-falling tone and the fifth step the most natural production of the low level (22) tone. For the mid-rising (23) tone, both speakers chose the eighth and tenth steps as the most natural productions of

[4]Figure 5.3 is also interesting with regard to the discussion above in Section 5.2.2 about the our use of a semi-vowel + open monophthong sequence in place of an opening diphthong. We predicted that the articulation and acoustics of our semi-vowel + open monophthong should closely resemble those true diphthongs, thus making an adequate substitute. In Figure 5.3, we see a gradual increase in both F1 and F2 throughout the /w/ until about half-way into the /aː/. This closely resembles the reverse pattern visible in the closing diphthong /aːu/ displayed in Figure 5.1). We would therefore argue that our semi-vowel + open monophthong is indeed diphthong-like and a suitable substitute for a true opening diphthong. In addition, regarding the lack of information in the literature about the intrinsic f0 of approximants, we do see some decrease in f0 with increasing F1 between the semi-vowel and the monophthong in Figure 5.3, as predicted, but it must be noted that this is likely exacerbated due to f0 declination (which is also visible in the closing diphthong in Figure 5.1).

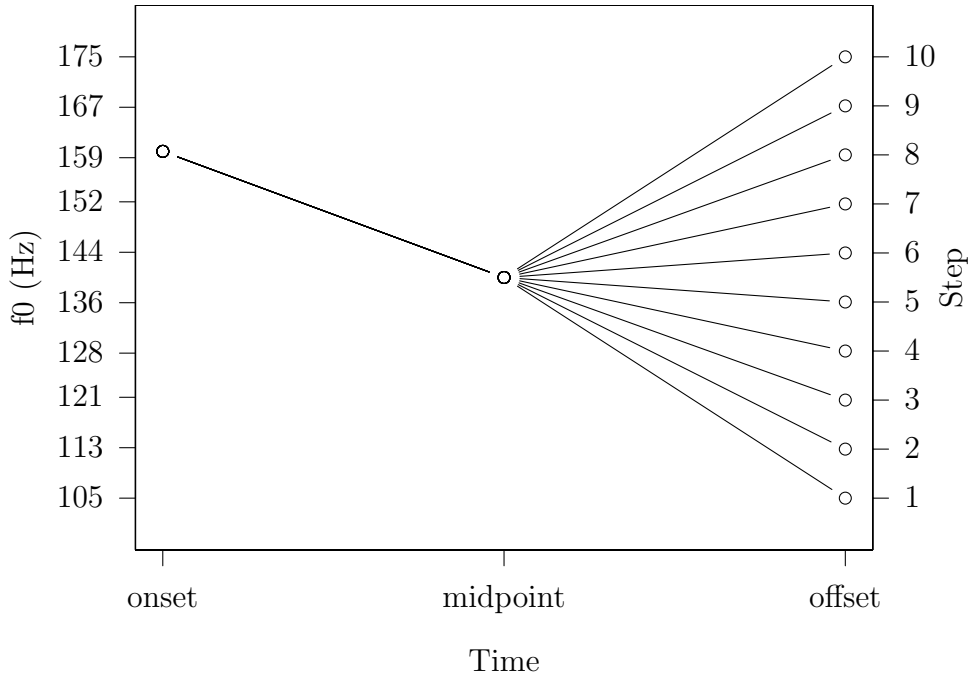


Figure 5.2: Ten step f0 continuum from the low-falling (21) tone via the low level (22) tone to the mid-rising (23) tone.

/maːu/ and /waːn/, respectively. Thus, we were confident that our resynthesised continua contained all three target stimuli.

In the first perception experiment in Chapter 3, the targets were embedded in a carrier sentence to allow listeners to normalise for speaker-specific characteristics, which is very important for correct tone perception in Cantonese. However, as this experiment was restricted to three tones with similar f0 height but quite different f0 slope (one rising, one level and one falling), and as the listeners had previously participated in the ERP experiment and were familiar with the model speaker's voice, we deemed a carrier sentence unnecessary. In addition, the nature of the primary ERP experiment demanded the tokens be presented in isolation.

5.2.3 Procedure

Participants carried out this experiment following a short break after the longer ERP experiment in which they were repeatedly presented with steps 1, 5 and 9 of the same two continua in a passive oddball paradigm. For the current study, they were seated at a computer in a quiet room with high-quality studio headphones (model Beyerdynamic DT770 Pro). The experiment was programmed in PsychoPy (Peirce, 2007). On each trial, participants were played one of five repetitions of each stimulus in isolation and blocked by diphthong but otherwise in random order. On the computer screen, they were presented with a choice of three Traditional Chinese characters for the real words representing our

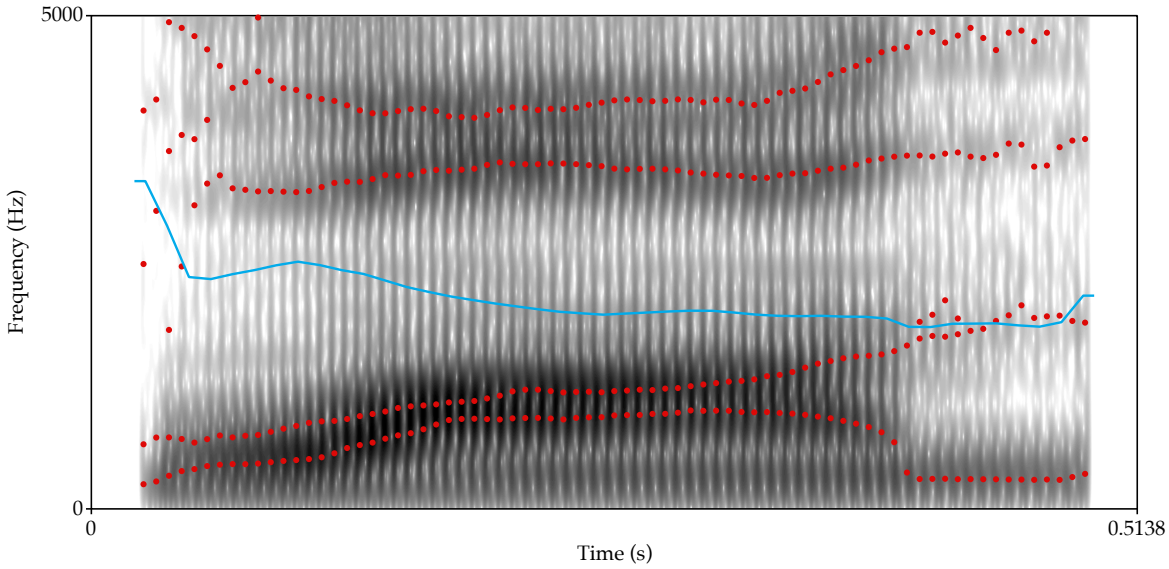


Figure 5.3: Model speaker's natural production of /waːn$_{22}$/ used as the basis for resynthesis. The red dotted lines reflects Praat's estimation of formants, while the solid blue line demonstrates the f0 contour.

three target tones (low-falling (21), low level (22) and mid-rising (23)) and an arrow key on the keyboard was assigned to each tone for the duration of the experiment. Each stimulus was presented once only on each trial. Participants were instructed to press the arrow key for the word that was the closest match to the stimulus they heard as fast as possible, after which a new trial began automatically. The response keys were inactive during presentation of the stimulus so that listeners were forced to listen to each stimulus in its entirety before making a decision. In total, participants heard 100 trials each (2 diphthongs * 10 steps * 5 repetitions) and the experiment took less than ten minutes. Our final table of results comprises data from 1000 trials (100 trials * 10 participants).

5.2.4 Analysis

Unlike the two-alternative forced choice experiment described in Chapter 3, this time listeners were presented with a three-way choice between the low-falling (21), low level (22) and mid-rising (23) tones. Thus, we were no longer able to carry out simple binomial logistic regression in this case. However, the low-falling (21) and mid-rising (23) tones were each expected to be confused with the low level (22) tone and not with each other, and a pre-examination of our results confirm that this was the case. Thus, analysis for this data set was simplified by re-coding listener responses as binary dependent variables so that a similar analysis to that described for Experiment 2 in Chapter 3 could be performed. We created a factor called Tone Condition which specified whether the crucial choice was

between the low-falling (21) and low level (22) tones or between the low level (22) and mid-rising (23) tones. Then, for the low-falling (21) vs. low level (22) Tone Condition, we categorised the listeners' responses as "low-falling (21)" vs. "other", and for the low level (22) vs. mid-rising (23) Tone Condition the responses were categorised as "mid-rising (23)" vs. "other". The vowel pattern was encoded in the variable "Diphthong" with the levels "opening" (for /waːn/) and "closing" (for /maːu/). Using this procedure, we were able to carry out binomial logistic regression as for Experiment 2 (Chapter 3).

We then proceeded as described for Experiment 2 (Chapter 3). First, for each listener we fit four sigmoids to the responses over the ten steps of the continuum: one for each combination of Tone Condition and Diphthong. The coefficients of the sigmoids, intercept k, slope m and turning point u were collected separately for each listener for data analysis. We then examined the data for each listener in order to identify the same errors described in Chapter 3 (Section 3.2.4) that would lead to a poor model fit[5]. C errors were corrected, data from listeners with A errors were checked but not removed, and B errors were counted for consideration during the interpretation of the results.

We then collected the revised regression coefficients for each speaker and ran our models for statistical analysis. For the full model, we calculated a linear mixed-effects model with category boundary u (i.e. the turning point) as the dependent variable, Diphthong and Tone Condition and an interaction between the two as fixed factors, and by-Participant random intercepts with random slope adjustments for Diphthong and Tone Condition. To test for interactions and main effects, we re-calculated subsequent models by removing the effect to be tested from the full model and comparing the full and subsequent models using likelihood ratio tests (Winter, 2013).

5.3 Results

Figure 5.4 shows the population means of the four sigmoids we calculated. The y-axes represent the proportion of low-falling (21) tone responses (left plot) and mid-rising (23) tone responses (right plot) at each step of the continuum on the x-axes.

Figure 5.5 plots the speaker means of category boundary u (i.e. the response variable in the statistical models) separately for each diphthong and each tone condition. The lower the value of u, the lower the step in the continuum at which the category boundary was heard. There appears to be a difference between closing and opening diphthongs at the category boundary between the low-falling (21) and low level (22) tones (left half of plot), but not at the category boundary between the mid-rising (23) and low level (22) tones (right half of plot). Thus, it appears that diphthong does not affect both tone conditions equally.

We hypothesised that dynamic intrinsic pitch would be shown if category boundary u

[5]In brief, A errors described situations in which listeners were unable to consistently identify one or both endpoints; B errors those in which perception was extremely categorical, leading to indefinitely steep slopes at the category boundaries; and C errors those in which the proportion of responses at one or both endpoints in an otherwise unambiguous region deviated slightly from 0 or 1.

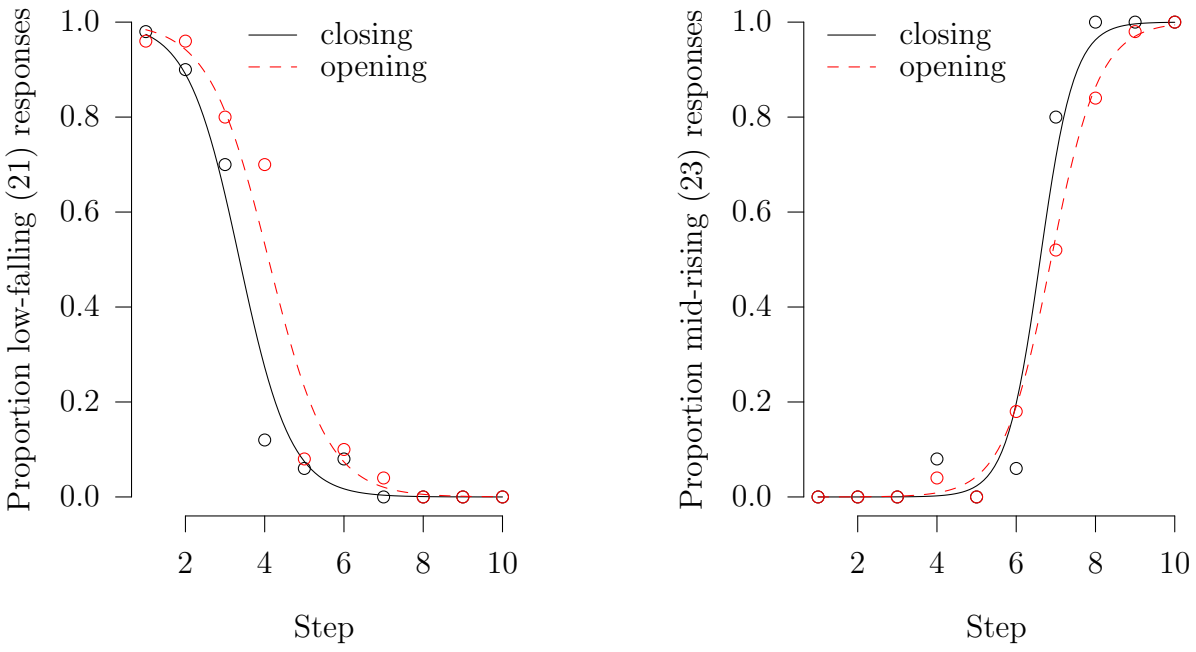


Figure 5.4: Proportion of low-falling responses (left) and mid-rising responses (right) for each of the ten steps on the continuum, separately for closing (solid) vs. opening (dashed) diphthongs.

was later (i.e. higher) for closing than for opening diphthongs, indicating that at equal f0 values (i.e. steps of the continuum in Figure 5.4) a low-falling (21) tone was more likely to be heard on closing diphthongs than on opening diphthongs. Instead, Figures 5.4 and 5.5 show the exact opposite pattern: opening dipthongs were more often (i.e. for a larger portion of the continuum) perceived as falling than closing diphthongs.

The likelihood ratio tests between the full and subsequent models revealed the following. Firstly, there was a main effect of Tone Condition ($\chi^2[9, 11] = 42.5, p < .001$). Secondly, there was a main effect of Diphthong ($\chi^2[9, 11] = 12.04, p < .01$). While there was no significant interaction between Diphthong and Tone Condition ($\chi^2[10, 11] = 3.28, p = .07$), planned post-hoc Tukey tests showed that the Diphthong affected u (the category boundary) in the low-falling (21) vs. other Tone Condition ($z = 3.7, p = .001$), but not the mid-rising (23) vs. other Tone Condition ($z = 1.2, p = .6$).

The slopes m reflect the region of (within-listener) uncertainty between two category boundaries. A very steep slope, for example in which a participant hears Step 3 as the low-falling (21) tone on 100% of trials but by Step 4 hears the low level (22) tone 100% of trials, indicates that the boundary between the two categories is very clear and the decision easy. On the contrary, a very gradual slope, for example in which numerous steps of the continuum are perceived as low-falling (21) only on some trials and low level (22) on others, shows indecision as to which step best reflects a low-falling (21) or low level (22)

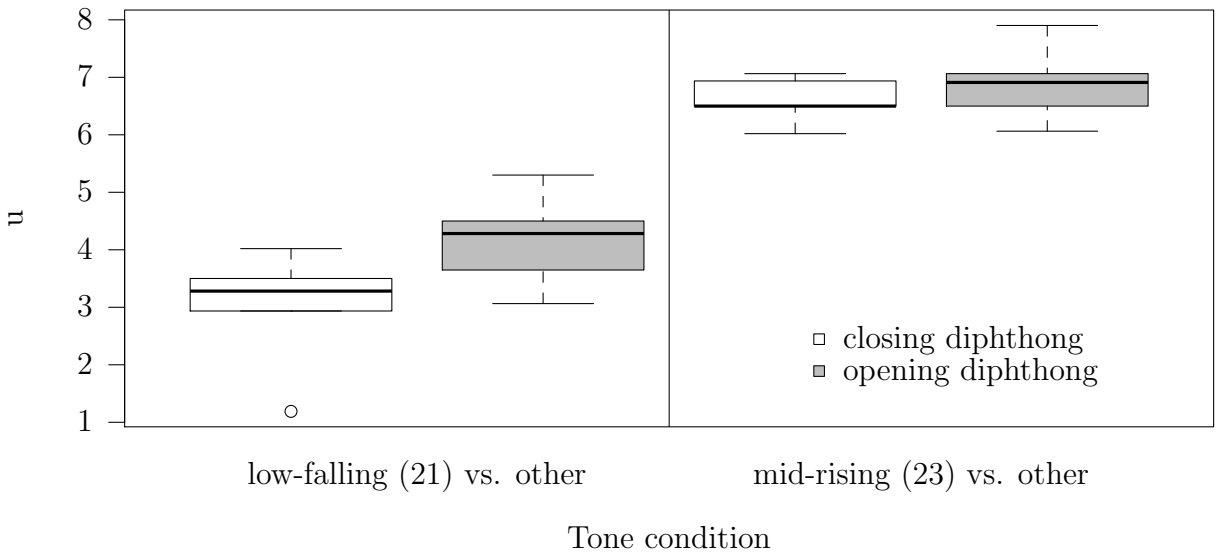


Figure 5.5: Category boundary u plotted as a function of Diphthong and Tone Condition (averaged for each listener). u is analogous to the step of the continuum at which the category boundary was heard. Lower numbers represented falling f0 contours and higher numbers rising contours.

tone. During the pre-analysis (see Section 5.2.4), we discovered a large number of B errors (infinite slopes)[6]. This prevented us from carrying out a test with the slopes m as the response variable, which may have provided us with some information about the difficulty the listeners had in each condition. However, as in Chapter 3, even without a meaningful estimation of the slopes m, we can attempt to draw conclusions about the difficulty of the task and the amount of indecision from the number of B errors and their distribution alone.

Table 5.1 displays the percentage of participants for whom slopes were infinitely steep in each condition. Regardless of Tone Condition (the rows in Table 5.1), tone identification was much easier for closing than for opening diphthongs. Tone Condition seems to have had little influence on closing diphthongs, but for opening diphthongs it appears that tone discrimination was somewhat easier when the choice was between the mid-rising tone (23) vs. other. These descriptive statistics fit with the between-listener variation in u (depicted in the sizes of the box and whisker plots) visible in Figure 5.5.

[6]Error types are described in Section 3.2.4 and reviewed again briefly in Section 5.2.4.

	opening diphthong	closing diphthong
low-falling (21) vs. other	30%	80%
mid-rising (23) vs. other	50%	70%

Table 5.1: Percentage of participants for whom regression coefficient m (slopes) were infinitely steep in each condition.

5.4 Discussion

Following Niebuhr (2004), we predicted that if intrinsic pitch is dynamic, closing diphthongs should have naturally falling intrinsic pitch and opening diphthongs rising intrinsic pitch (i.e. the inverse of the intrinsic f0 effect). For Cantonese, this translates into an earlier category boundary for opening diphthongs than closing diphthongs at both ends of our continuum from falling to rising f0. In addition, we predicted that if intrinsic pitch is not just a case of pitch shift or a general auditory effect but rather a compensatory effect for intrinsic f0, then the size of the intrinsic pitch effect should decrease with decreasing f0.

We found an effect of Diphthong (closing vs. opening) on the boundary between the low-falling (21) and low level (22) tones, but not on the boundary between the low level (22) and mid-rising (23) tones. However, the effect of Diphthong at the low-falling (21) vs. low level (22) category boundary was that, on identical f0 contours, opening diphthongs sounded more falling, which is not predicted by intrinsic pitch. As such, we find no evidence of dynamic intrinsic pitch on diphthongs in Cantonese. In the following, we will separate the discussion for each of these results. We will then compare our study with Niebuhr's before discussing general implications and the motivation for our production experiment in Chapter 6.

Before we discuss the results in more detail, it is important to point out some limitations of our design that may have influenced our results. Firstly, the low-falling (21) and low level (22) tones are sometimes difficult to distinguish and might even be in the midst of a tone merger (Mok et al., 2013). Secondly, in natural speech the low-falling (21) tone is characterised not only by its f0 contour, but also by creaky voice as an important secondary cue (Yu & Lam, 2014); see also Section 1.2.3. As we wanted to vary f0 only and keep all other factors constant in our experiment, we did not include creak in any of our stimuli. Thirdly, the time course of our stimuli differed in that the shift from close to open tongue position happens very early in opening diphthong /waːn/, while the shift from open to close tongue position is considerably later in closing diphthong /maːu/. As a result, any change in intrinsic pitch that might be expected to occur would perhaps take place earlier, relatively speaking, for the opening diphthings than for the closing diphthong. Furthermore, it is possible that /waːn/ was perceived as an open monophthong rather than the opening diphthong we intended. We will come back to these points during our discussion of the results.

H1: More low-falling responses for closing than opening diphthongs at low-falling (21) vs. low level (22) end of continuum

The category boundary between the low-falling (21) and low level (22) tones was significantly later for opening diphthongs: listeners heard a larger portion of the continuum as falling when the diphthong was opening than when it was closing. Contrary to our expectations and the intrinsic pitch hypothesis, this is evidence that listeners perceive opening diphthongs as sounding more falling than closing diphthongs. In the following, possible explanations for this result are considered.

The lack of creak on the low-falling (21) tone must be considered very seriously. A recent study showed that not only is creak prevalent in the production of the Cantonese low-falling (21) tone, it also serves as a perceptual cue to tone identity (Yu & Lam, 2014). Nevertheless, there is no reason to believe that excluding this cue directly caused the unexpected result. The creak in the Cantonese low-falling (21) tone increases over the time course of the tone as f0 sinks lower and lower until regular glottal vibration is difficult to maintain. Yet creaky voice is known to be more prominent on open vowels than close vowels (Panfili, 2016). As such, creak should occur more frequently on opening diphthongs in which the very low f0 at the end of the falling tone coincides with the open vowel. Thus, if missing creak in the low-falling (21) tone were to bias our responses, it should be the case that listeners would be more likely to interpret opening diphthongs without creak as the low level (22) rather than the low-falling (21) tone. According to this idea, the sigmoid for the opening diphthong on the right-hand side of Figure 5.4 should be left-shifted compared to the closing diphthong - but it is not[7]. Although creaky voice is no doubt important for the low-falling (21) tone, we do not believe it could be responsible for the reverse effect we found.

Another factor that may have contributed to this result is the listeners' acceptance of /waː/ in /waːn/ as an opening diphthong as we intended. Let's assume for a moment that this was not successful, and listeners perceived the /aː/ portion as an open monophthong following a semi-vowel onset. We know from previous studies and Experiment 2 in Chapter 3 that, all else being equal, open vowels should sound higher than close vowels (intrinsic pitch). As /aːu/ in /maːu/ is indisputably a (closing) diphthong, according to Niebuhr's study it should sound more falling and thus still be biased toward the low-falling (21) tone (and the /waːn/ toward the higher of the two tones: the low level (22) tone). Again, this would result in the curve for the opening diphthong on the right side of Figure 5.4 being left-shifted compared with the closing diphthong, and this is not what we find.

One explanation that would fit with the result is to do with the time pressure of the task, as listeners were asked to answer as quickly as possible. It may be that listeners' decision-making was based largely on the very beginning of the tokens. From this perspective, the closing diphthong (in which the intrinsic pitch should begin high) should have sounded higher than the opening diphthong (in which the intrinsic pitch should begin low). This may have prompted listeners to enter more low-falling (21) responses (as the lower of

[7]Incidentally, had we found the pattern we expected, this may have been a confounding factor limiting support for our hypothesis.

the two tones) for the opening diphthong than the closing diphthong. Although it might explain our result, this explanation is not particularly satisfying for several reasons. For one, the response keys were only activated after the token had been presented in full so that the listeners could not make a premature decision. Of course, this does not rule out the possibility that their decisions were already biased by the early time course of the stimuli. Furthermore, if listeners only attend to the very early time course of the signal, then Niebuhr should not have found dynamic intrinsic pitch in his study. Most of all, we have no explanation as to why listeners should associate only the beginning of the vowels with the f0 trajectory.

H2: More mid-rising responses for opening than closing diphthongs at low level (22) vs. mid-rising (23) end of continuum

The category boundary between the low level (22) and mid-rising (23) tones occurred at approximately the same step of the f0 continuum for both diphthong patterns (for opening diphthongs: $\overline{u} = 6.86$, $\sigma_u = 0.56$; for closing diphthongs $\overline{u} = 6.53$, $\sigma_u = 0.32$; see Figure 5.5). Again, this result is at odds with the dynamic intrinsic pitch effect found by Niebuhr (2004), if not quite as surprising as the reverse result found for H1. There are several explanations for this result that would each have very different consequences for pitch processing.

The simplest explanation for the non-effect of intrinsic pitch is just that: contrary to Niebuhr's hypothesis, what we know about intrinsic pitch on monophthongs does not generalise to diphthongs. Niebuhr himself had some problems with responses to his stimuli and his results were not equally straightforward in all conditions. In short, intrinsic pitch simply may not be dynamic.

On the other hand, the lack of an intrinsic pitch effect does not exclude that the effect is automatic, as a result of pitch shift or some other general auditory process, and should thus be universal, but can be suppressed by the listener - for example, in order not to interfere with tone contrasts. This would explain the presence of the effect in Niebuhr's data for German listeners but not in our own.

The main problem with either of the above explanations arises when we consider the link between production and perception. Vowel openness is robustly linked with intrinsic f0 in production of monophthongs, and there is no reason to believe that this should not apply to diphthongs as well. Closing diphthongs, in which tongue position rises from open to close, should be associated with rising intrinsic f0. Just as for monophthongs (Chapters 2 and 3), we predict that in order to avoid tone confusion, rising intrinsic f0 on diphthongs would have to be balanced out in perception with falling intrinsic pitch. Otherwise, tone perception would surely depend (to a small degree) on diphthong pattern. Thus, it was essential that we establish whether or not the intrinsic f0 effect we found for Cantonese monophthongs in Experiment 1 (Chapter 2) applied to diphthongs. This was an important motivation for following up on this study with the production experiment in Chapter 6. It may well be that there is no rising or falling intrinsic pitch on diphthongs and that what we know from monophthongs does not transfer. For Cantonese, however, this would require

that there is also no dynamic effect of diphthongs on f0 in production or else the intrinsic f0 would confound with phonological f0 (tone). Otherwise, we might be confronted with a mismatch in the tradition of Ohala (1981).

Finally, it is possible that our experimental design is responsible for the lack of a dynamic intrinsic pitch effect. We return to Table 5.1, which listed the percentage of listeners for whom regression coefficient m (slope) was infinitely steep. These percentages are in stark contrast to those listed in Table 3.2 for Experiment 2 (Chapter 3). The large number of infinitely steep slopes is a clue that the steps on our f0 continuum (Figure 5.2) were not spaced finely enough for this experiment. The f0 offset of each step differed in increments of approximately 7Hz. These increments were not too fine for the test run with our informants, whose tone perception was categorical but not so steep as that of our test participants. Yet due to the extremely categorical nature of the sigmoids and the large number of infinite slopes it is more than likely that the steps were not fine enough to reveal the very small effects of vowel openness (if any).

One previous study in this regard unfortunately escaped our notice until after our data had been collected. Klatt (1973, p. 13) describes an otherwise unpublished study on Mandarin tone identification conducted by Victor Zue at the Massachusetts Institute of Technology. Zue created four synthetic f0 contours based on real speech to match the four Mandarin tones as reproduced in Figure 5.6 (from Klatt, 1973, p. 13) and superimposed these on a synthetic /ba/ syllable. Three native speakers had no difficulty identifying the tones and made no errors. When Zue reduced the frequency range of his synthetic tone space only very slightly, tone identification remained good with few errors. Astonishingly, when he reduced the frequency range to just 4Hz, tone identification was 90% correct. By compressing the tone space even more, however, his listeners were unable to distinguish between Mandarin tones 2 and 3. Note that Mandarin tones 2 and 3 (as displayed in Figure 5.6) are similar in shape, so they are most representative of the type of tone contrasts we modelled in our f0 continuum. We can conclude from Zue's study that even two similar tones can still be correctly identified when there is (considerably) less than 7Hz difference between them. As such, in order to tease apart the effects of phonological and phonetic f0 (due to tone and vowel-intrinsic pitch, respectively), we would need to re-run the experiment with much finer steps.

H3: Smaller effect at low-falling (21) vs. low level (22) boundary than at low level (22) vs. mid-rising (23) boundary

The reasoning behind this hypothesis was that if intrinsic pitch reflects perceptual normalisation for intrinsic f0, and the intrinsic f0 effect is reduced at the lower end of the tone space (cf. Whalen & Levitt, 1995, and Chapter 2), then intrinsic pitch should also be reduced at lower f0. As we did not find intrinsic pitch (falling pitch on closing diphthongs and rising pitch on opening diphthongs) at either category boundary, this hypothesis cannot be addressed.

In the following, general implications and problems are considered that are related to the

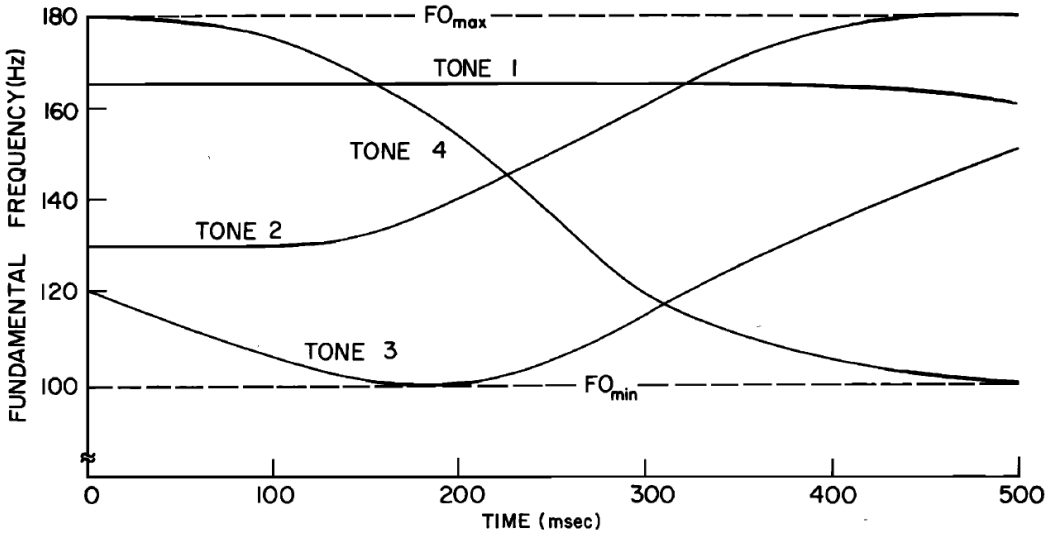


Figure 5.6: The synthetic f0 contours modelled on real speech and representing the four Mandarin tones in Victor Zue's study (from Klatt, 1973, Figure 5, p. 13).

study as a whole and not just to specific hypotheses.

As always, lexical effects may have played a role in listener's tone perception. According to our informant, all items were equally common, well-known words in Hong Kong Cantonese. It was difficult to confirm this from a more scientific point of view, but according to the Hong Kong Chinese Lexical Lists for Primary Learning (Chinese Language Education Section, 2008), all characters are listed in the "list of written form of commonly-used Chinese characters". In addition, this same source provides two lexical lists for primary learning: Key Stage I contains 4 914 essential words for primary school children in grades 1 to 3, while Key Stage II includes 4 792 essential words for primary school children in grades 4 to 6. While these lists are compiled with general language use and not just lexical frequency in mind, they are a good base for checking for lexical bias in our study. Indeed, our stimuli are categorised as follows:

Key Stage I: /waːn$_{21}$/ 還 *to return*, /maːu$_{22}$/ 貌 *appearance*

Key Stage II: /maːu$_{23}$/ 牡 *peony*, /waːn$_{22}$/ 幻 *illusion*, /maːu$_{21}$/ 矛 *spear*

"List of written form of commonly-used Chinese characters": /waːn$_{23}$/ 輓 *to mourn* (i.e. not in the Lexical Lists for Primary Learning)

If we assume that words listed for Key Stage I (primary school grades 1 to 3) have the highest lexical frequency and are most important in general language usage, then we might expect bias toward these items (/maːu$_{22}$/ and /waːn$_{21}$/) in our continua. For the continuum between the low-falling (21) and low level (22) tones, this would be manifested in more low-falling (21) responses for opening diphthongs and more low level (22) responses

for closing diphthongs, which is exactly the reverse of the intrinsic pitch hypothesis but in fact matches what we found (and discussed under H1 above). For the low level (22) vs. mid-rising (23) continuum, closing diphthongs should have been biased toward the low level (22) tone; that is, there should have been fewer mid-rising (23) responses and the curve for closing diphthongs in Figure 5.4 should have been right-shifted. In this case, this would have been a confounding factor with our hypothesis and it is not what we found, so it is unlikely it played a role.

It is difficult to determine to what extent lexical effects impacted our results. All stimuli were common words according to both our informants and the lexical lists for Hong Kong primary schools. If there were lexical effects, they might help to explain the reverse effect for the continuum between the low-falling (21) and low level (22) tones but not the non-effect for the continuum between the mid-rising (23) and low level (22) tones.

Perhaps more important than lexical effects are token and type frequency. Token frequency reflects how often a given token appears, while type frequency records the number of different token types in existence. For example, the token frequency of items with the high level tone in a database of spoken Cantonese sums the number of occurrences of the high level tone. Any one word with a high level tone that comes up several times in this database will be counted each time it occurs. Instead, type frequency records the number of unique words occurring in conjunction with the high level tone. A word with a high level tone that reoccurs in identical form throughout the database will only be counted once.

One previous study on Cantonese tone hyperarticulation found that for the mid level (33) and mid-rising (23) tones, f0 was higher on low frequency words (Zhao & Jurafsky, 2009). If our stimuli vary not only in vowel pattern but also in frequency, this may have affected our results.

Leung, Law, & Fung (2004) compiled type and token frequencies on the basis of the Hong Kong Cantonese adult language corpus (Leung & Law, 2001), which is comprised of orthographic and phonetic transcriptions of eight hours of spontaneous radio speech in Hong Kong. Their most relevant findings for the interpretation of our results are as follows. In terms of onset frequency, /m/ was more frequent (token frequency = 6 711, type frequency = 119) than /w/ (token frequency = 4 462, type frequency = 75). However, this was balanced out by the higher frequency of rime /aːn/ (token frequency = 2 189, type frequency = 57) compared with /aːu/ (token frequency = 604, type frequency = 32). In terms of tone, the low level (22) tone is the most frequent of the three tones we used in this study (in terms of both type and token frequency). This might cause a very slight bias toward the low level (22) tone in either of our tone conditions, but if so this should apply equally to both vowel patterns and thus not confound with diphthong. In addition, while the low-falling (21) and low level (22) tones are about equally frequent on syllables with final /n/ coda, the mid-rising (23) tone is considerably rarer (token frequency = 310, type frequency = 18). While this does not necessarily affect the actual tokens we chose, which were judged to be equally common by our informants, if it did it would have the effect of a response bias against mid-rising (23) responses on our opening diphthong (i.e. bias against /waːn$_{23}$/ or toward /waːn$_{22}$/). Relative to the closing diphthong, the opening diphthong in Figure 5.4 is not biased toward the mid-rising (23) tone, although it is certainly right-

shifted compared to our hypothesis. While we cannot rule out that this frequency effect might have overridden any intrinsic pitch effect here, this does not seem plausible, as there was no intrinsic pitch in the low-falling (21) vs. low level (22) condition either.

On top of the methodological considerations outlined above, we only had ten participants and five repetitions per condition for this small-scale experiment. We also cannot rule out participant fatigue, habituation or other side effects that might have been caused by the experiment being conducted subsequent to the ERP study referred to in Section 5.2. In light of the contrasting results in our study and Niebuhr's, further studies are necessary to uncover the mechanisms behind intrinsic pitch on diphthongs in general and especially in contour tone languages. Such studies should consider the following improvements.

Firstly, the f0 distance between the offsets of the steps in our continuum needs to be decreased. Judging by the results of Victor Zue's study as well as Klatt's (1973) own data, distances of 2Hz at most would be reasonable, although possibly as low as 0.5Hz. In addition, more repetitions per condition would decrease the chances of complete category changes from one step to the next. Of course, there is a trade-off between step size and number of repetitions if participant fatigue is to be avoided. Our short experiment was an offshoot from a larger ERP study and was limited in size, but a full-scale study devoted to this topic could at least afford smaller step sizes or more repetitions, if not both.

Furthermore, it would be advantageous to study the phenomenon of dynamic intrinsic pitch in the upper regions of the tone space where we would expect any effect to be largest and where there is no confound with differing phonation quality. However, there are no contour tones in Hong Kong Cantonese that are restricted to the upper regions of the tone space. For older speakers of Guangzhou Cantonese and speakers of other dialects spoken in Guangdong province, a distinction remains between the high level (55) and high-falling (51) tones[8], much like the distinction between Mandarin tones 1 and 4 in Figure 5.6. A continuum between these two tones differing in f0 offset (with very fine f0 increments as discussed above) could be superimposed on words with undisputed closing diphthongs such as /au/ or /ai/ and words with open monophthongal controls such as /a/. The closing diphthongs should have falling intrinsic pitch and the close monophthongs high (level) intrinsic pitch. Consequently, if diphthongs have dynamic intrinsic pitch, listeners should perceive a larger portion of the continuum as the high level (55) tone on the close monophthong and more of the continuum as high-falling (51) on the closing diphthong.

Finally, if intrinsic pitch reflects perceptual compensation for intrinsic f0 in speech production, it is possible that a lack of dynamic intrinsic pitch on Cantonese diphthongs is evidence that intrinsic f0 is simply not dynamic (i.e. rising or falling) in production of diphthongs. To test this, we ran a production experiment (Experiment 4) analagous to Experiment 1 (Section 2) for Cantonese diphthongs and contour tones. Should intrinsic f0 indeed rise and fall with closing and opening diphthongs, respectively, it might be evidence

[8] Although, as discussed in the description of Cantonese tonal phonology in Section 1.2.3, this distinction is merging in Guangzhou Cantonese. In addition, while the Cantonese high-falling tone clearly falls from about tone level 5, there is no consensus as to how far it falls. Bauer & Benedict (1997) describe it as a 51 contour, but it has traditionally been described as 53.

that, unlike for monophthongs, listeners are simply unable to compensate for the effects of vowel openness on diphthongs, with implications for tone processing.

Chapter 6

Experiment 4: Intrinsic f0 in Cantonese diphthong production

In Experiment 3 (Chapter 5), we failed to find any evidence of dynamic intrinsic pitch in the perception of Cantonese diphthongs. This led us to question whether intrinsic f0 is dynamic in speech production and rises and falls with the changing tongue configurations and formants of diphthongs. If intrinsic f0 is indeed an automatic and thus universal phenomenon, it should be dynamic in that f0 should rise with increasing vowel openness in closing diphthongs such as /au/ and vice versa for opening diphthongs. Cantonese provides an excellent test case for such an analysis because of its large inventory of diphthongs and tones differing only slightly in their f0 contour (rising, falling, level). For example, while open monophthong /a/ should generally have a low intrinsic f0 on any given tone, we might expect the f0 of closing diphthong /au/ to rise slightly relative to /a/. In turn, this may have consequences for tone perception in complex tone languages such as Cantonese, in which very slight changes in f0 slope lead to categorical changes in tone perception - especially if effects of vowel on f0 are not compensated for in perception (see Chapter 5).

6.1 Hypotheses

To our knowledge, intrinsic f0 of diphthongs has not been tested empirically. By repeating Experiment 1 as outlined in Chapter 2 with diphthongs, we aimed to test whether intrinsic f0 is indeed dynamic. We compared the f0 of diphthongs and triphthongs with that of a monophthongal control. Relative to monophthongal controls, a dynamic effect of vowel openness on intrinsic f0 would predict:

H1 more positive f0 slopes on closing diphthongs (rising intrinsic f0),

H2 more negative f0 slopes on opening diphthongs (falling intrinsic f0), and

H3 more f0 curvature on triphthongs.

As such, the slope of a level tone on a closing diphthong should be slightly steeper in the positive direction than a monophthongal control, while the same level tone on an opening diphthong should be slightly steeper in the negative direction than the same monophthongal control. On the other hand, a triphthong should show a very slight concave or convex f0 pattern (depending on the pattern of vowel openness) compared to the monophthongal control.

6.2 Method

6.2.1 Participants

Again, the experiment was a repeated-measures design in which all participants took part in all conditions. We recruited 10 native speakers (3 males) of Cantonese aged between 26 and 30 (mean age 28.2 years) who also participated in the production and perception experiments described previously in Chapters 2 and 3. They had been living in Munich for 1 to 4 months (mean 2.2 months) but were active members of the local Cantonese community. For further information, please refer back to Chapters 2 and 3.

6.2.2 Stimuli

In order to elicit the strongest possible effects of varying vowel openness within a diphthong, we chose stimuli with closing diphthongs /aːu/, /aːi/, /ɐu/ and /ɐi/ (described phonologically by Bauer & Benedict, 1997, in Table 1.4 as /aːw/, /aːj/, /ɐw/, /ɐj/ sequences, respectively) and opening counterparts /waː/, /jaː/, /wɐ/ and /jɐ/.

In addition, we included stimuli with triphthong /jɐu/ (opening-closing vowel pattern) to establish whether any dynamic intrinsic f0 patterns that apply to diphthongs are equally transferable to more complex vowel patterns.

Table 6.1 displays the chosen stimuli alongside the Traditional Chinese characters used to represent them to participants.

tone	closing diphthongs				opening diphthongs			control	opening-closing triphthong
	/maːi/	/maːu/	/ŋɐi/	/ŋɐu/	/jɐn/	/wɐn/	/waːn/	/ji/	/jɐu/
high level (55)								醫	幽
high-rising (25)					讞	穩		綺	柚
mid level (33)								意	幼
mid-rising (23)	買	牡	蟻	偶	蚓	尹	挽	耳	友
low level (22)	賣	貌	偽	吽	刃	運	幻	貳	又
low-falling (21)	埋	矛	危	牛	人	雲	鬟	移	油

Table 6.1: Phonological transcriptions of the stimuli with the Traditional Chinese characters used to prompt them. Only cells containing a character represent stimuli used for our experiment. For English translations, see Appendix A.2.

6.2.3 Procedure

The procedure for this experiment was the same as that described for Experiment 1 in Section 2.2.3 with the exception that all ten speakers were recorded in the presence of a native Cantonese speaker who monitored correct pronunciation.

6.2.4 Post-processing

In light of the difficulties selecting appropriate stimuli for Experiments 3 and 4 (see discussion in Chapter 5, Section 5.2.2), we have included example spectrograms of the audio signal for a number of the target stimuli in Appendix D. For each speaker, one realisation each is depicted of a short and a long closing and opening diphthong as well as a triphthong. While these tokens were selected randomly and therefore may not be perfectly representative of each speaker, they demonstrate the durational characteristics and formant structures typically associated with each token.

Segmentation and labelling

The audio channel was automatically segmented and labelled using the MAuS forced alignment system (Schiel, 1999). MAuS uses the same algorithm as WebMAUS used in Experiment 1 (Section 2.2.4) but is run from the command line rather than a web app and allows more flexible parameter settings not available through the web app. We required this increased flexibility in order to accommodate pronunciation variants and chunk segmentation for more accurate location of the word boundaries.

For this stimuli set, we expected and observed variation in pronunciation of stimuli with phonological /ŋ/ onset. In Hong Kong Cantonese, initial /ŋ/ is undergoing a sound change in which it is increasingly being substituted with a glottal stop (Bauer & Benedict, 1997; Matthews & Yip, 2013). Initially, we assigned a rule to our MAuS parameters so that MAuS would automatically identify the onset as either /ŋ/ or /ʔ/ and label the segment accordingly. However, MAuS was unable to reliably distinguish between these two sounds and many tokens were thus segmented and labelled incorrectly. Thus, we manually identified the onset sound (auditorily and visually) and then re-ran MAuS with the correct labels already pre-assigned for each token. The new positions of the segment boundaries with the correct labels were very reliable and were able to be used as for all other segments. In line with our conventions for analysing f0, utterances with initial [ŋ] were taken in their entirety, but utterances with initial [ʔ] were restricted to the voiced portion (in this case the vowel only).

EGG and f0

f0 was extracted from the EGG waveform in the same manner as for Experiment 1 (Section 2.2.4). However, f0 from one speaker was unable to be tracked reliably and thus needed to be excluded from analysis.

6.2.5 Analysis

Statistical analysis was carried out in EmuR (Harrington, 2010; R Development Core Team, 2011).

For f0 analysis, we separated the voiced from voiceless portions by calling all segments which were phonologically voiced. Thus, for this stimuli set, our tokens for analysis were taken as a whole unless the syllable-initial /ŋ/ was produced as allophone /ʔ/, in which case only the rime was used. As noted in Chapter 2, these portions may include small devoiced or voiceless portions due to slight errors in MAUS' calculation of the boundaries.

As in Experiment 1 (Chapter 2), f0 was z-score normalised by speaker, and again we time-normalised the data by labelling the starting point of each token as 0 and the end point as 1 and selected only the central 80% of each token for analysis

To test our hypotheses, we needed dynamic f0 information revealing changes in the contours of diphthongs compared with monophthongs. Thus, we reduced the f0 information across all time points to just two values per token representing the slope and the curvature of each f0 contour. We achieved this by carrying out a discrete cosine transformation (DCT) of the f0 signal as described in Watson & Harrington (1999) for formant trajectories. Specifically, a DCT breaks the signal down into half-cycle frequency cosine waves, the amplitudes of which are the DCT coefficients. Given N time points in a signal trajectory $x(n)$ extending from $n = 0$ to $N - 1$, we can calculate the m^{th} DCT coefficient, $C(m)$, as follows:

$$C(m) = \frac{2}{N} k_m \sum_{n=0}^{N-1} x(n) cos\left(\frac{(2n+1)(m-1)\pi}{2N}\right) \tag{6.1}$$

$$m = 1, ...N \tag{6.2}$$

$$k_m = \frac{1}{\sqrt{2}}, m = 0; k_m = 1, m \neq 0 \tag{6.3}$$

In the context of this study, it is the second coefficient, $C(1)$, that is particularly useful, as it provides a measure of the direction and degree of linear slope in the f0 contour. As $C(1)$ is almost perfectly *negatively* correlated with the slope (Harrington, 2010, p. 312), positive values indicate a falling (negative) slope and negative values indicate a rising (positive) slope. This is particularly important to note in light of our hypotheses. In addition, we will make use of the third coefficient, $C(2)$, as this coefficient represents the trajectory's curvature. A positive $C(2)$ value indicates a ∪-shaped trajectory, while a negative value indicates a ∩-shaped trajectory (Harrington, 2010, pp. 201 & 312).

The DCT coefficients can be added together to reconstruct the original signal, much like a discrete Fourier synthesis. By limiting the number of coefficients summed together, it is possible to create a smoothed version of the original trajectory using only the most informative coefficients[1]. For our purposes, the above three coefficients are capable of reproducing the original signal very nicely while eliminating smaller perturbations due

[1]For details, see Harrington (2010).

to outliers. For comparison, for the plots below we have included equivalent plots with speaker-normalised f0 in Hz (i.e. before the DCT smoothing) in Appendix E.

f0 analysis

We ran two full mixed-effects models as base models using the `lme4` package in `R` (Bates et al., 2015; R Development Core Team, 2011) and tested for interactions and main effects by comparing, using likelihood ratio tests, the full model with subsequently calculated models in which the effect to be tested was removed from the full model (Winter, 2013).

In Model 1, aimed at investigating the influence of diphthong pattern on f0 slope, we used coefficient $C(1)$ (equivalent to f0 slope) as the dependent variable, Tone (low-falling (21) vs. low-level (22) vs. mid-rising (23)) and Vowel (opening diphthong vs. closing diphthong vs. close monophthong) as fixed effects. In addition, we set random intercepts by Speaker and Item with by-Speaker random slope adjustments for Vowel and Tone.

In Model 2, aimed at investigating the difference between triphthongs and monophthongs, coefficient $C(2)$ (representing f0 curvature) was the dependent variable, Tone (all six tones) and Vowel (closing-opening triphthong vs. close monophthong) were fixed effects. Again, we set random intercepts by Speaker and Item with by-Speaker random slope adjustments for Vowel and Tone.

6.3 Results

6.3.1 Slopes

Figure 6.1 shows the DCT-smoothed f0 contours of the falling (21), mid-rising (23) and low-level (22) tones separately for closing and opening diphthongs and close vowel monophthongal control[2].

On the low-level (22) tone, it seems that while all three vowels begin at the same f0 height, the closing diphthong ends somewhat lower than the other two vowels and hence has a steeper, more negative slope. This is not what we would expect; rather, we predicted the closing diphthong to have more positive slopes and the opening diphthong more negative slopes than the monophthongal control.

On the low-falling (21) tone, the close monophthong appears to have the flattest slope, beginning lower and ending higher than the other two tones. The opening diphthong begins higher and ends lower than the monophthongal control, revealing a more negative slope in line with our predictions. Interestingly, the closing diphthong begins at the same f0 as the opening diphthong and yet ends even lower than the opening diphthong. Again, we predicted the closing diphthong to have the most positive slope.

For the mid-rising (23) tone, there appears to be no difference in slope depending on vowel, although the monophthongal control may have a slightly more positive slope than the diphthongs. It appears that the diphthongs for this tone additionally differ from the

[2]The original signal (before DCT analysis) is reproduced in Appendix E for comparison.

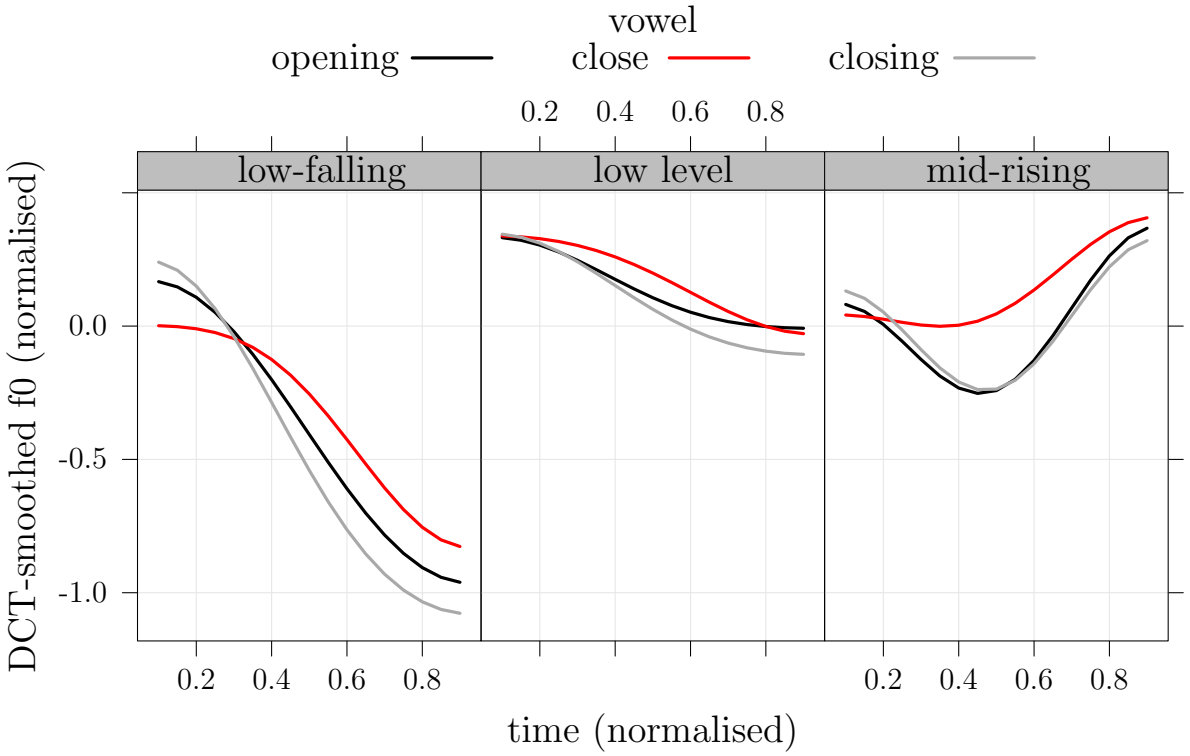


Figure 6.1: DCT-smoothed normalised f0 contours for each tone separately for the opening and closing diphthongs and close monophthong control.

monphthongal control in their curvature: that is, the diphthongs have a more concave shape than the monophthongal control.

Figure 6.2 shows the second coefficient, $C(1)$, of the discrete cosine transformation. That is, instead of analysing f0 over time as depicted in Figure 6.1, we have just one value for each f0 contour. These are the values used for statistical analysis. $C(1)$, on the vertical axis, is correlated with the slope of the f0 over time. As outlined above, when $C(1)$ equals zero, the slope is flat; $C(1)$ above zero corresponds to negative slopes and below zero to positive slopes. As in Figure 6.1, the mid-rising (23) tone has the most positive slopes, the low-level (22) tone has slightly negative slopes, and the falling (21) tone has the most negative slopes and also the most variation. The $C(1)$ values depending on vowel seem to closely match those in Figure 6.1, but the lack of real differences between vowels indicate no effect of vowel on tone. If anything, vowel affects the slopes ($C(1)$) of the low-falling (21) tone only, in that both closing and opening diphthongs have slightly steeper (negative) slopes than the monophthongal controls, with the effect size being greater on the closing diphthongs.

Statistically, there was an effect of Tone ($\chi^2[20, 22] = 22.2, p < .001$) but not of Vowel

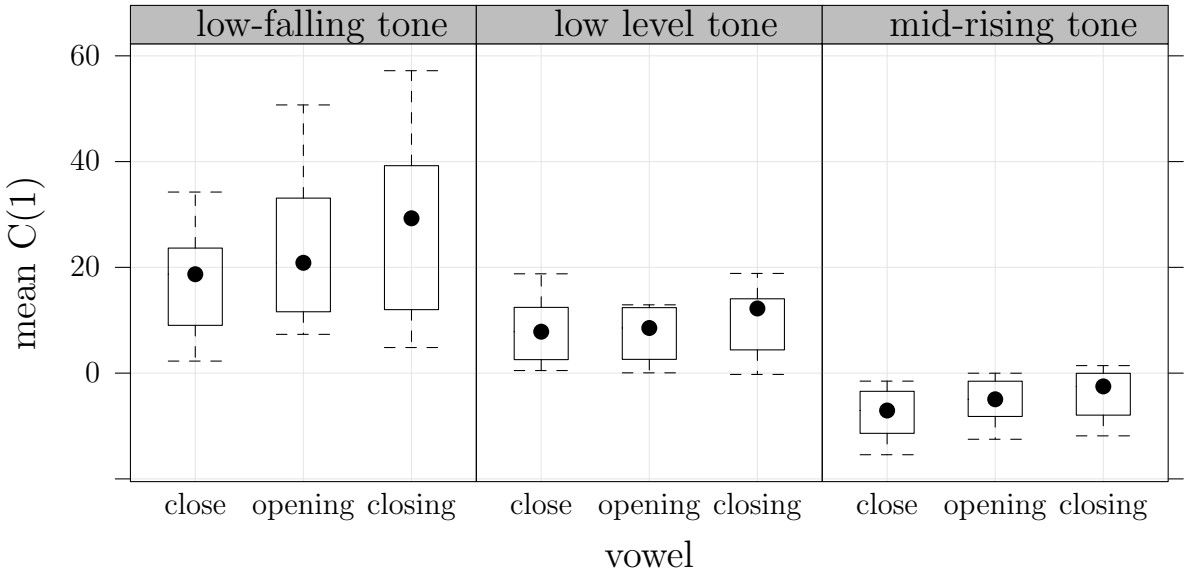


Figure 6.2: Speaker means of DCT coefficient $C(1)$ as a function of vowel and tone. A positive $C(1)$ value reflects a negative slope and vice versa.

($\chi^2[20, 22] = 5.3, p = .07$) on $C(1)$. However, there was an interaction between Tone and Vowel ($\chi^2[22, 26] = 36.1, p < .001$). Post-hoc Tukey tests revealed that the effect of Tone and the non-effect of Vowel was consistent for all pairwise comparisons but for a significant difference between $C(1)$ of the closing diphthong and the monophthongal control on the falling tone (21) ($z = 3.8, p < .01$), and no difference in $C(1)$ of the low-level (22) and low-falling (21) tones on the monophthongal control ($z = 2.7, p = .9$). Thus, closing diphthongs increased $C(1)$ (and thus decreased the slopes) of the falling tone compared to the monophthongal control, but there were no other effects of Vowel on slope.

Figure 6.3 is exactly the same as Figure 6.1, except that Figure 6.3 separates the diphthongs according to their length. Short vowels (closing diphthongs /ɪɐi/ and /ɪɐu/; opening diphthongs /jɐn/ and /wɐn/; close monophthongal control /ji/) are plotted in solid lines, while long diphthongs (closing /maːi/ and /maːu/; opening /waːn/) are plotted in dashed lines.

The same general pattern is visible in Figure 6.3 as in Figure 6.1: on the low-falling (22) and mid-rising (23) tones, the monophthongal control has the flattest slope, while on the low level (22) tone, the close monophthong has the highest f0 for most of the contour. With regard to diphthong length, however, the long diphthongs (dashed black and grey lines) are better differentiated from the control monophthong (red) than the short diphthongs (solid black and grey lines). This is hardly surprising, given that the shorter the segment, the less time is available to reach articulatory and acoustic targets (e.g. Lindblom, 1963).

However, there appears to be little difference overall between the opening (black) and closing (grey) contours, with the exception that short opening diphthongs seemed to have patterned with the control monophthong on the low level (22) tone and, to a lesser extent, on the low-falling (21) tone. If intrinsic f0 were to pattern with vowel openness, however, the grey contours (closing diphthongs) should all have more positive slopes than both the black contours (opening diphthongs) and the red (control). This is not the pattern we see in Figure 6.3, neither for the short nor the long closing diphthongs, so our unexpected results are not an artefact of a vowel length confound.

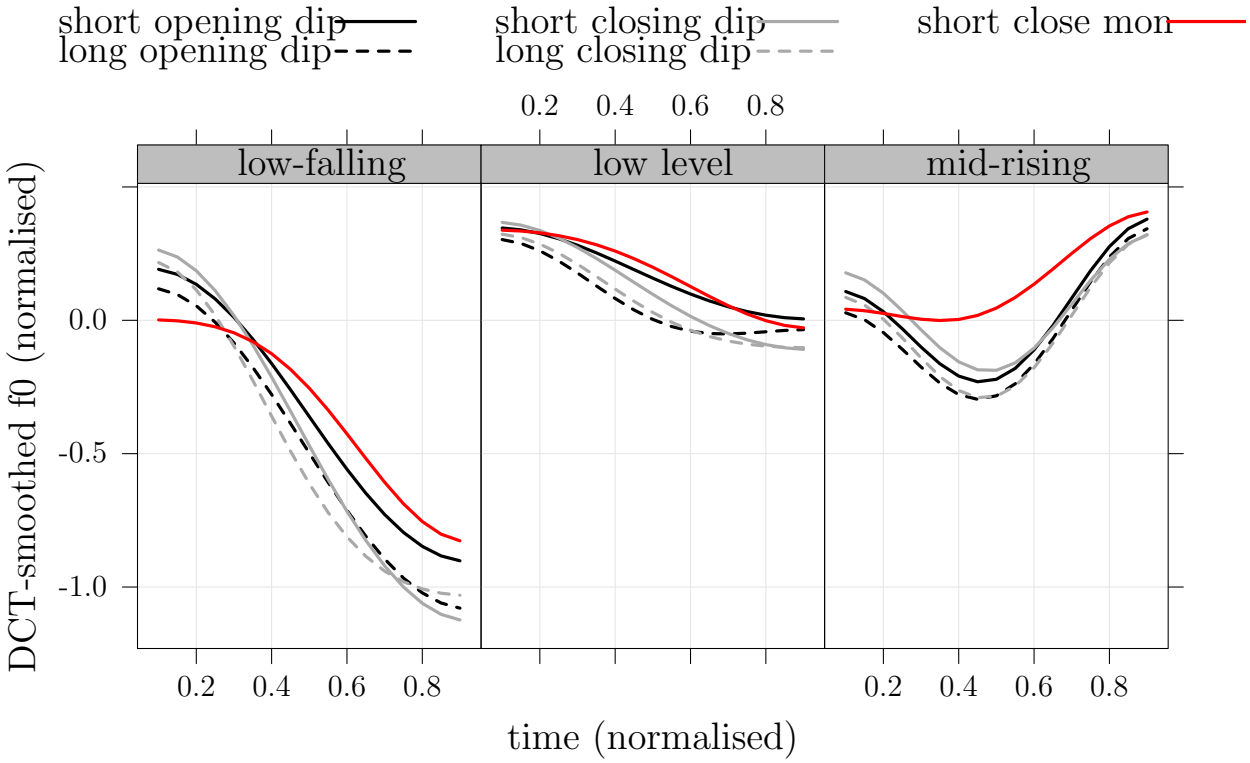


Figure 6.3: DCT-smoothed normalised f0 contours are plotted for each tone (low-falling on the left, low level in the centre, and mid-rising on the right) and vowel length (short in solid and long in dashed lines) separately for the opening (black) and closing (grey) diphthongs and close monophthong control (red).

6.3.2 Curvature

Figure 6.4 shows the DCT-smoothed f0 contours for all six tones separately for the opening-closing triphthong and the close vowel monophthongal control[3]. We expect intrinsic f0 on

[3]The original signal (before DCT analysis) is reproduced in Appendix E for comparison.

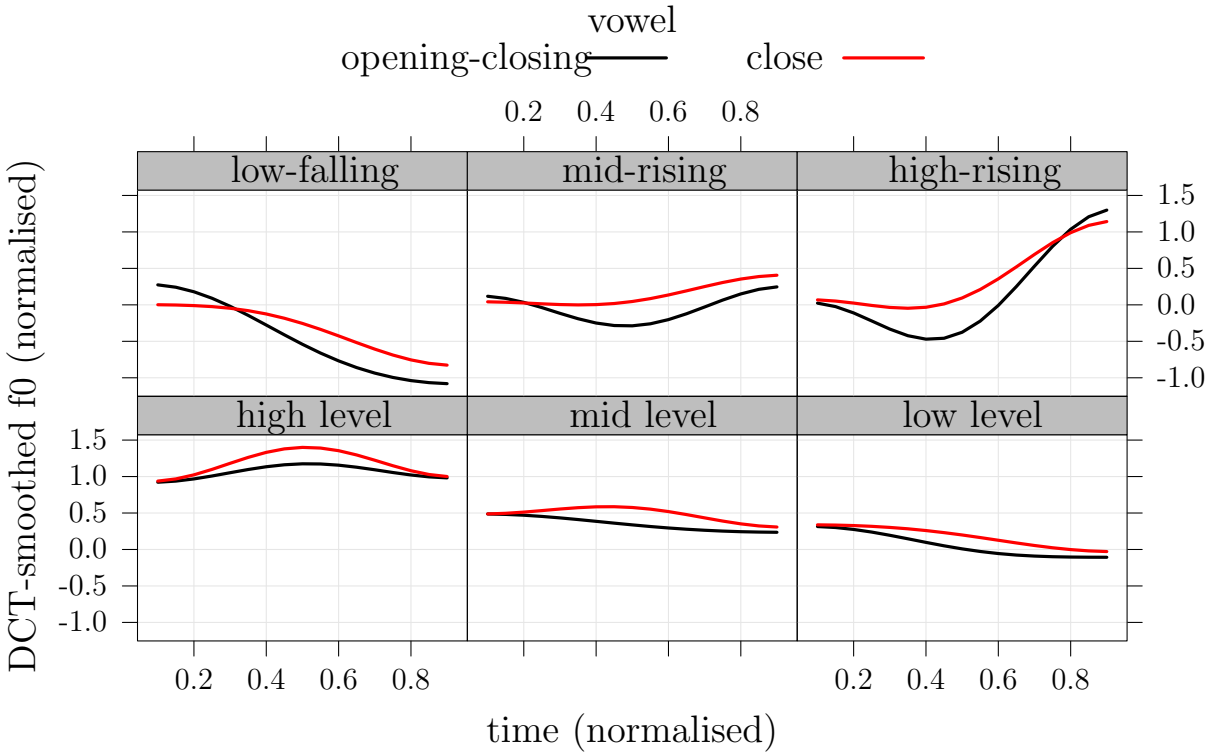


Figure 6.4: DCT-smoothed normalised f0 contours for each tone separately for the opening-closing triphthong and close monophthong control.

the triphthong to fall and rise again compared to the monophthongal control. For the two rising tones (mid-rising 23 and high-rising 25), this does indeed appear to be the case, while on the low-falling tone the triphthong merely seems to have a steeper slope and on the level tones a minimally lower f0 height compared to the close vowel control.

Figure 6.5 shows the third coefficient of the discrete cosine transformation. Again, we have just one value for each f0 contour and this is the value used for statistical analysis. $C(2)$, on the vertical axis, is correlated with the curvature of the f0 over time. When $C(2)$ equals zero, the curvature is flat; $C(2)$ above zero corresponds to a concave shape and below zero to a convex shape. We see very little effect of vowel on the level tones or the low-falling tone (although the triphthong tends to have slightly greater curvature than the control in each of these tones), but there appears to be an effect on the rising tones (mid-rising 23 and high-rising 25). For these tones, the opening-closing triphthong appears to increase curvature relative to the monophthongal control, and this effect is largest on the high-rising tone (25).

Statistically, we found an effect of Tone ($\chi^2[5] = 27.2, p < .001$) and of Vowel ($\chi^2[1] = 8.4, p < .01$) on $C(2)$. In addition, there was an interaction between Tone and Vowel

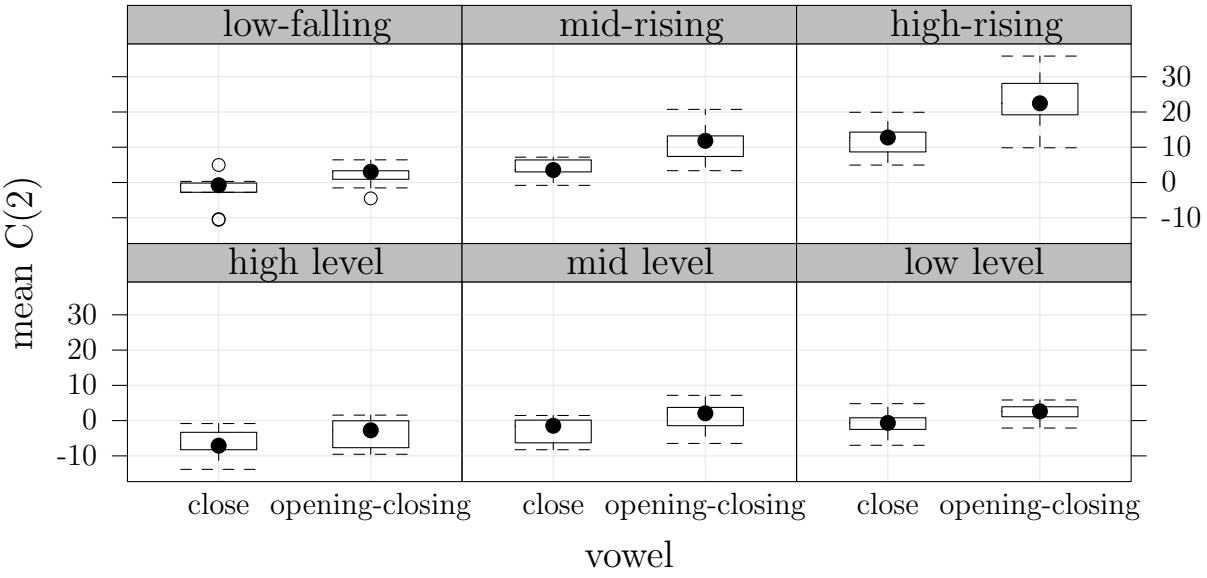


Figure 6.5: Speaker means of DCT coefficient $C(2)$ as a function of vowel (close monophthong vs. opening-closing triphthong) and tone. A positive $C(2)$ value indicates a ∪-shaped trajectory, while a negative value indicates a ∩-shaped trajectory.

($\chi^2[5] = 31.2, p < .001$). However, post-hoc Tukey tests showed that the interaction was because Tone affected $C(2)$ in most, but not all, pairwise comparisons. Vowel had no effect on $C(2)$ in any of the pairwise comparisons. Thus, we cannot conclude that Vowel itself has any significant influence on $C(2)$.

6.4 Discussion

Just as for f0 height in Experiment 1 (Chapter 2), in this study f0 shape was predominantly determined by tone. We expected closing diphthongs to create a slightly more positive f0 slope and opening diphthongs a slightly more negative slope relative to the control monophthong. In addition, we expected the f0 contour to be more ∪-shaped on opening-closing triphthongs than on the close monophthongal control. DCT coefficients $C(1)$ and $C(2)$ were interpreted as reflecting the slope and curvature of the signal, respectively. Vowel pattern had no significant effect on $C(1)$ with the exception of one condition: closing diphthongs increased $C(1)$ on the low-falling (21) tone (i.e. they *decreased* the f0 slopes on this tone). We noted that this was the opposite direction to the expected effect. Coefficient $C(2)$ depended on tone but not on vowel (triphthong vs. monophthong), neither as a main effect nor in any of the pairwise comparisons, despite numerical trends (Figures 6.4 and

6.5) hinting at increased curvature on the rising tones. We discuss the results separately for diphthongs (slopes analysis) and triphthongs (curvature analysis).

6.4.1 Slopes

We were unable to confirm either of our hypotheses for f0 slopes. Closing diphthongs showed no evidence of rising intrinsic f0 and opening diphthongs no evidence of falling intrinsic f0 in our Cantonese data. Curiously, however, closing diphthongs significantly increased $C(1)$ (i.e. decreased slope) on the low-falling (21) tone. Our discussion of slopes begins with the first result and ends with the latter, more surprising effect.

It seems that, at least for Cantonese, what we know about intrinsic f0 on monophthongs does not carry over to diphthongs, or at least not without some qualifications and further investigation. In our view, there are three possible explanations for this result.

Firstly, this could be taken as evidence against the hypothesis that intrinsic f0 is automatic with a physiological or aerodynamic basis, in accordance with, for example, the contrast model (Diehl & Kluender, 1989; Kingston & Diehl, 1994). However, the overwhelming majority of literature so far[4], admittedly based on studies of monophthongs only, shows clear evidence of intrinsic f0 even in tone languages. Vowel-intrinsic f0 has even been documented in infants' babbling (Whalen, Levitt, Hsiao, & Smorodinsky, 1994). As a result, we think this explanation is highly unlikely.

Alternatively, it is possible that intrinsic f0 is indeed automatic, but in this case is suppressed. This explanation would also be somewhat compatible with the contrast model if the suppression of vowel-intrinsic f0 is seen as auditory enhancement of the tone contrast. Proponents of this model generally do not believe that intrinsic f0 is automatic but rather a further (controlled) cue used to help further distinguish vowel quality contrasts. However, as Hoole and Honda (2011) point out, there is no reason to exclude what they refer to as a "hybrid" model in which intrinsic f0 is both automatic and may also be enhanced (or in this case suppressed) where necessary. According to Connell (2002):

> Languages that rely on f0 movement (contour tone languages), those that require the achievement and maintenance of specific targets at a range of heights (register tone languages) and those that involve the achievement of targets at specific but relatively infrequent locations in the utterance (nontone and pitch accent languages) likely all involve different degrees of f0 control and it is not unreasonable to assume that the mechanisms used to achieve this control should vary accordingly (Connell, 2002, p. 121).

Connell goes on to argue that vowel-intrinsic f0 is probably universal and is either "suppressed where phonological concerns are overriding" (Connell, 2002, p. 122) or constrained in languages in which "the degree of control, and therefore the mechanism of control, required is antagonistic to the occurrence of [intrinsic f0]... [intrinsic f0] can perhaps be viewed as no less universal for this, as its absence is in a sense accidental" (Connell, 2002,

[4]see Connell (2002) for an exception in the case of Mambila.

p. 122). One way of testing this hypothesis would be by looking for dynamic intrinsic f0 effects on diphthongs in a non-tone language while simultaneously measuring muscle activity using electromyography as in Hoole & Honda's 2011 study mentioned above. In their view, it is most likely the genioglossus posterior that is responsible for automatic vowel-intrinsic f0, while it is primarily the cricothyroid and sternohyoid muscles that are responsible for actively raising and lowering f0 (Sagart, Hallé, de Boysson-Bardies, & Arabia-Guidet, 1986; Hoole & Honda, 2011). It would be important that the target vowels not be pitch accented. By repeating this study using electromyography, it would be possible to monitor activity of these muscles in order to establish whether the cricothyroid or sternohyoid muscles are active during production of diphthongs. If they are, it would be evidence that intrinsic f0 on diphthongs is indeed suppressed (rather than non-existent).

Finally, before exposing test participants to the discomfort of electromyography, there is at least one more possible explanation for the lack of dynamic intrinsic f0 on diphthongs.

Some previous studies have shown that the vocalic onset of a diphthong tends to be more prominent and produced with less target undershoot than the offset (Gay, 1968)[5]. An investigation into the formant values of our data was beyond the scope of this dissertation, but if these findings apply to our data as well, we would expect the intrinsic f0 of the initial open vowel in the onset of the diphthong to outweigh the close vowel offset. Note also in Table 6.1 that our closing diphthongs comprised two phonologically long diphthongs (/aːi/ and /aːu/) as well as two phonologically short diphthongs (/ɐi/ and /ɐu/). In Cantonese, the onset of the longer diphthong is particularly prominent, as can be seen in the example spectrograms in Appendix D. Thus, the intrinsic f0 of these tokens might be expected to be more similar to that of an open vowel rather than an opening diphthong. In addition, if the offset vowels were produced with considerable target undershoot (which is doubtful, considering the pronunciation of the stimuli in isolation), this very undershoot would surely also result in intrinsic f0 undershoot. In this case, the intrinsic f0 should be very similar to that of an open monophthong, which would differ from our close monophthongal control in f0 height, but not in slope. Figure 6.1 shows that this may be the case: the height of f0 does indeed seem to be considerably lower than the close vowel monophthongal control on the mid-rising (23) tone and may be slightly lower on the low-level (22) and low-falling (21) tones. Yet another contributing factor may have been our decision to take the central 80% of our target stimuli in order to factor out the effects of inaccurate boundaries and/or neighbouring consonants. In many cases, the close vowel offsets of our diphthongs may have been too close to the end of the vowel for their spectral effects to show up in our data - again adding to the evidence that our data for closing diphthongs may not be representative of real-world f0 patterns for these diphthongs. Manual segmentation of the boundaries and perhaps correction of any f0 miscalculations might show different results.

Of course, this does not explain the lack of an effect on the opening diphthongs, but this condition was not without its own problems. For one, these may not be considered "real" diphthongs, and indeed they cannot be considered as diphthongs in traditional Cantonese phonology. We argued that they are "phonetic" diphthongs in the sense that they are

[5]But see (Hu, 2013).

comprised of a semi-vowel + open monophthong sequence linked by dynamic formant transitions and that their effect on f0 should be the same as a "real" opening diphthong (falling intrinsic f0). However, by extracting only the central 80% of the target stimuli for analysis, we may also have eliminated the small portion of the signal in which the close vowel onset/glide could have been expected to affect f0. In addition, as the syllable nucleus there can be no doubt that the open vowel portion of this condition is more prominent than the glide. In this light, it is perhaps no surprise that the intrinsic f0 values might again resemble those of an open vowel monophthong, and indeed Figure 6.1 indicates that at least for the mid-rising (23) tone, intrinsic f0 is lower than on the close vowel monophthongal control.

We also cannot exclude the possibility that, with just nine speakers (one speaker's f0 could not be tracked and was removed from analysis), we simply do not have enough data. However, compared with the studies in Whalen & Levitt's comprehensive review, our sample is by no means unusually small (many of the studies listed had less than five speakers). Thus, in the absence of any numerical trends in our data, it is difficult to say whether our results reflect the big picture when it comes to intrinsic f0 on diphthongs or whether our experiment design was fundamentally flawed.

Finally, we briefly turn to the one significant finding: the apparently more negative slopes on closing diphthongs in connection with the low-falling (21) tone. Figure 6.2 also shows the most variation between speakers in this condition, which may reflect a lack of stability. The low-falling (21) generally appears to be plagued by greater between-speaker variation than the other tones.

6.4.2 Curvature

If our analysis of diphthongs is indeed representative of intrinsic f0 over all dynamic vowel sequences, we should not expect to find any effect of triphthongs on curvature. Thus, it may come as no surprise that in our analysis of coefficient $C(2)$, we were also unable to demonstrate a significant effect of opening-closing triphthongs on the shape of the f0 trajectory. In our view, this is also not surprising when the temporal domain is considered. Natural speech is inherently coarticulated and involves both gestural and spectral overlap. However, the amount of overlap surely increases with the amount of targets within a complex vowel such as diphthong or triphthong. We therefore consider it likely that gestural and/or spectral undershoot either reduces any intrinsic f0 patterns to insignificance or masks them. In future work, a formant analysis of our acoustic data would hopefully shed light on this matter. We could compare the spectral characteristics of our open and close monophthongs from Experiment 1 with the different vowel openness phases of the triphthongs studied here in order to establish the extent of any target undershoot.

Having said that, and in line with Brunelle et al.'s (2010) (limited) data, there does appear to be a slight numerical trend in the right direction for all tones, but particularly for the two rising tones (see Figures 6.4 and 6.5). It is not clear to us why intrinsic f0 of triphthongs should affect the rising tones more than the other tones. Nonetheless, it is possible that an effect of intrinsic f0 on triphthongs might show up with more data.

Chapter 7

General Discussion

The aim of this dissertation was to investigate the extent to which tone language users tease apart the micro and macroprosodic contributions to f0 contours. Our prediction was that tone language users must be very closely attuned to the effects of vowel openness in both production and perception in order for them not to interfere with tone contrasts in complex tone inventories as in Hong Kong Cantonese. If production (intrinsic f0) and perception (intrinsic pitch) of phonetic effects are not well aligned, this could compromise tone contrasts with serious consequences for fields as diverse as speech technology (human-machine interactions) and sound change. Four experiments were designed to investigate this complex relationship.

Experiment 1 (Chapter 2) aimed to confirm the presence of intrinsic f0 in Cantonese tone production and establish whether it is diminished at lower f0. We used the simplest scenario of open and close monophthongs on the three level tones. These three level tones are not distributed equally across the Cantonese tone space: the mid (33) and high (55) level tones are well separated, but the low (22) and mid (33) level tones overlap. As such, we expected an interaction between tone and vowel openness: if intrinsic f0 affects the mid (33) but not the low (22) level tone, there should be less tone overlap on close vowels than on open vowels on this tone contrast. To test this, we carried out two analyses: one looking at the effect of tone and vowel openness on f0 height, and the other looking at machine classification of the tones based on f0 values alone following training with tone but not vowel openness information. Both analyses confirmed a partial effect of vowel openness (intrinsic f0) and greater overlap between the low (22) and mid (33) tones than between the mid (33) and high (55) tones. However, the analyses differed in their assessment of where the intrinsic f0 effect occurred. In the f0 analysis, there were numerical trends for all tones, but especially the low (22) and high (55) tones. The effect was statistically significant on the high (55) tone only. Machine classification, however, showed that intrinsic f0 interfered with correct classification of the low (22) vs. mid (33) tone contrast but not with classification of the mid (33) vs. high (55) tone contrast. We attributed this to the closer proximity of the low (22) and mid (33) tones and the better separation of the high (55) tone from the other two level tones. Thus, while vowel openness had the greatest effect on the high (55) tone, it seemed that its most damaging effect was on the low (22)

vs. mid (33) tone contrast.

Experiment 2 (Chapter 3) was carried out to examine whether the cross-linguistically less robust phenomenon of intrinsic pitch applies to Cantonese. We used the same tones and vowels as in Experiment 1 and hypothesised that intrinsic pitch should occur where intrinsic f0 applies in production in order for vowel openness not to interfere with tone contrasts. That is, according to our own f0 analysis in Experiment 1 as well as the previous studies indicating that intrinsic f0 is often diminished at lower f0, intrinsic pitch should also be smaller or non-existent at lower f0. Alternatively, in line with the machine classification data from Experiment 1, it would be reasonable to expect that intrinsic pitch is greater at lower f0 in Cantonese because this is where intrinsic f0 would be most problematic for the low (22) vs. mid (33) tone contrast. Indeed, this is what we found. A trend towards intrinsic pitch was present in our Cantonese data, but in terms of statistical significance it was restricted to the low (22) vs. mid (33) tone contrast.

In Chapter 4, we examined the relationship between production (intrinsic f0) and perception (intrinsic pitch) in more detail. Our goal was to establish to what extent intrinsic f0 and intrinsic pitch are matched both at a global level (across all participants) and at a more individual level (within participants). Data from participants who took part in both of Experiments 1 and 2 (Chapters 2 and 3) were examined. It is unclear how exactly to quantify the relationship and what would constitute a good match between production and perception, so our analysis must be considered only as a first attempt. After all, Experiment 2 contained speech from one speaker only, and there is no reason to expect that participants' perception of intrinsic pitch should be identical for all speakers. Nonetheless, our analysis revealed an interesting paradox. It appeared that at a global level, intrinsic pitch and intrinsic f0 are matched where the tone contrast depends on it: that is, for the (overlapping) low (22) vs. mid (33) level tone contrast. Yet for the mid (33) vs. high (55) level tone contrast, which is clearly distinguished in Cantonese, perception appears to lag behind production, in that the intrinsic pitch effect is slightly smaller than the intrinsic f0 effect. Interestingly, there is huge between participant variation, and we find that the relationship between production and perception differs greatly by participant. Overall, we do not find that participants who produced large intrinsic f0 differences also perceived large intrinsic pitch effects and vice versa. There appears to be no correlation between the two modes at an individual level. This might be an artefact due to the intrinsic f0 values being based on a group of speakers whereas the intrinsic pitch is based on (numerous listeners') perception of just one speaker's speech. It would be an entirely new undertaking (outside of the scope of this dissertation) to investigate the extent to which intrinsic pitch (and speech perception in general) varies across speech from different model speakers in laboratory-style perception experiments. At present, it is common practice in laboratory phonology to use speech from just one or two model speakers in perception experiments, but interpretations of these studies might be restricted to perception of those model speakers only and not the general population. Indeed, it may be that this methodological problem is responsible for the very different results relating to intrinsic pitch cross-linguistically (e.g. Pape, 2009).

Experiment 3 (Chapter 5) was inspired by a previous study indicating that intrinsic pitch is dynamic and thus applies to diphthongs as well: perceived pitch was found to fall

with decreasing vowel openness and vice versa (Niebuhr, 2004). As the Cantonese tone system includes a number of contour tones with phonologically falling and rising pitch, falling and rising intrinsic pitch might complicate contour tone perception. Our aim was to extend Niebuhr's findings by replicating the result in a tone language. We created an f0 continuum between the low-falling (21) and mid-rising (23) tones and superimposed it on opening and closing diphthongs. Participants chose between the low-falling (21), low level (22) and mid-rising (23) tones in a three-alternative forced choice experiment. However, we were unable to replicate Niebuhr's result for Cantonese. At the category boundary between the low-falling (21) and low level (22) tones, we instead found the reverse result: opening diphthongs sounded more falling than closing diphthongs. At the category boundary between the low level (22) and mid-rising (23) tones, there was no effect of diphthong whatsoever. There are several plausible explanations for these results. From a theoretical perspective, we suggested that either intrinsic pitch is not dynamic, in which case there must be another explanation for Niebuhr's results, or that it can be suppressed as necessary - in this case in order to maintain a clear tone contrast. From a methodological perspective, insufficiently small step sizes may have prevented any small effect of diphthong from showing, and unbalanced lexical or frequency effects might have led to a bias toward one of our diphthongs (especially in the low-falling (21) vs. low level (22) condition). Importantly, these results motivated us to see what is happening in Cantonese diphthong production and whether intrinsic f0 applies there.

Experiment 4 (Chapter 6) aimed to rectify this gap in the literature. To our knowledge, studies on intrinsic f0 in production have been restricted to monophthongs. There, the f0-raising effect of close vowels and/or f0-lowering effect of open vowels seems to be universal. We therefore predicted that intrinsic f0 should change dynamically with changing vowel openness: opening diphthongs should have falling intrinsic f0 and closing diphthongs rising f0. But if this is so, it would indicate that production (intrinsic f0) is not aligned with perception (intrinsic pitch, Chapter 5) for Cantonese diphthongs, with potential consequences for Cantonese tones. We tested the same three tones (low-falling (21), low level (22) and mid-rising (23)) and two diphthongs (opening and closing) as in Experiment 3. In addition, a close monophthong was included as a control, and a near-triphthong rounded out the dataset. Contrary to our hypothesis, we found no evidence of intrinsic f0. The only significant effect of diphthong was an f0-lowering effect of closing diphthongs in combination with the low-falling (21) tone, for which we were unable to find a fitting theoretical explanation. Thus, contrary to the effect we found for monophthongs in Experiment 1 (Chapter 2), there was no visible effect of intrinsic f0 on diphthongs in Cantonese. We pointed out that this does not necessarily prove that intrinsic f0 as reported for monophthongs does not transfer to diphthongs, as it is possible that the effect is suppressed in Cantonese in order to maintain the tone contrast. Either way, the lack of an intrinsic f0 effect in production appears to fit in well with the missing intrinsic pitch in Experiment 3 (Chapter 5). Incidentally, the curious finding that opening diphthongs sounded more falling at the low-falling (21) vs. low level (22) tone boundary in perception also matches quite well with the equally unexpected falling intrinsic f0 on closing diphthongs in combination with low-falling (21) tones in production. It could be argued that these two opposing effects cancel each other

out to some extent and thus pose no problem for the tone contrast. However, although these two effects were both statistically significant, it is entirely possible that they are a coincidence or, in our opinion more likely, an artefact of our experimental design (especially in the case of Experiment 3).

Results are presented in terms of their relevance for the different explanations for vowel-intrinsic f0 and vowel-intrinsic pitch as well as for general tone processing and sound change theory.

7.1 Intrinsic f0: automatic or controlled?

In our opinion, the overwhelming majority of prior evidence points towards vowel-intrinsic f0 being a natural phenomenon arising automatically from vowel production. In this section our results are discussed against this background with special focus on one area of our study that is problematic for this explanation for intrinsic f0. In addition, we discuss our results in light of theories that disagree with the automatic account.

Critically, we found no evidence of falling or rising intrinsic f0 in Experiment 4 (Chapter 6). An automatic account would require that the effect apply to all types of vowels. On the face of it, this might be seen as evidence of intrinsic f0 as controlled. However, there are also problems with this interpretation. If intrinsic f0 is controlled, why would it be implemented on monophthongs but not diphthongs? We concluded that this explanation is not likely and instead suggested that any one (or a combination) of a number of other factors might apply. Firstly, this study looked at the low f0 region, where previous studies have shown diminished intrinsic f0 effects. It is possible that the physiological mechanisms underlying intrinsic f0 simply are much reduced in this region. In Experiment 1 (Chapter 2) for monophthongs, there was no effect of monophthong on f0 in the low-mid tone regions; the effect we found there was in machine classification of these tones[1]. In future, the experiment could be replicated in the upper tone space of a more suitable language such as Guangzhou Cantonese or Mandarin. Secondly, it is possible that any intrinsic f0 effect was (actively) suppressed specifically because of the close proximity of the tones tested. In Section 6.4, we suggested testing this theory by investigating the involvement of the muscles involved in active f0 control (thought to be the cricothyroid and sternohyoid muscles) as a function of diphthong and tone. Thirdly, the articulation and spectral characteristics of diphthongs might be temporally and/or spatially compressed compared with monophthongs, which could conceivably decrease or possibly eliminate the automatic mechanisms behind intrinsic f0. It would therefore be worthwhile examining the spectral characteristics of Cantonese diphthongs. Finally, our analysis procedure might need to be refined in order for very small effects of intrinsic f0 to come to light. In particular, manual segmentation of segment boundaries (and perhaps correction of miscalculated f0 contours) would enable us to effectively analyse the entire voiced portion of the recorded tokens rather than just the central 80% of tokens segmented by forced alignment, as in our study.

[1] although this was also based on f0, so there must have been at least some effect of vowel on f0 for our monophthongs.

We now turn to the relevance of our results for theoretical accounts of intrinsic f0 as a controlled phenomenon.

Kingston (2007) proposes that the absence of intrinsic f0 on low tones in some tone languages is best accounted for by the auditory enhancement theory. He proposes that:

> [...high] tones differ from [low] tones in another way that motivates permitting [vowel-intrinsic f0] differences at the top of the speaker's range, while limiting them at the bottom. Speakers are far freer to vary F0 at the top when pronouncing a [high] tone (Liberman and Pierrhumbert 1984), in particular they can raise F0 more without pushing that tone's F0 target into the range of another tone. But when pronouncing a [low] tone at the bottom, speakers run up against a rather hard floor that prevents further F0 lowering, and they cannot raise F0 much without raising that tone's F0 target into the range of a higher tone's target (Kingston, 2007, 177).

Firstly, the results Kingston refers to in M. Liberman & Pierrehumbert's (1984) study are based on pitch accent in English intonation, which in our view is not comparable with lexical tone. Secondly, while there may be some physiological basis for more controlled f0 variation at higher f0 (the cricothyroid, which is commonly associated with controlled f0, has been linked more with f0 raising than lowering, cf. Hoole and Honda (2011) and references cited therein), we suspect that a given tone language's phonology (i.e., its tone space) would play a much stronger role in the distribution of f0 variation.

For the Cantonese tone space, which is rife with overlap in the low to mid tone regions, it would indeed be logical not to enhance vowel contrasts in this region using intrinsic f0 in order to avoid increased tone overlap. Yet machine classification of tones depended on vowel openness in this region in our Experiment 1 (Chapter 2), hinting that intrinsic f0 occurs here in spite of the crowded tone space (and the "hard floor"). Furthermore, most tone languages do not suffer the tone crowding that Cantonese does. According to Maddieson (2013), the number of languages with simple tone systems (defined by this source as a two-way contrast) far outweigh the number of languages with complex tone systems, so complex tone systems such as Cantonese are rather the exception than the rule (especially when we consider that many complex tone languages rely strongly on phonation as a further cue). In tone systems restricted to just two or three tones, a considerable amount of f0 variation should be possible on each tone without encroaching on the space of another. We demonstrate this with the help of Hombert's (1977a) study of Yoruba, a Benue-Congo language spoken in Nigeria and Benin.

Figure 7.1 (Hombert, 1977a, p. 181) shows average vowel-intrinsic f0 for each of the three contrastive tones in Yoruba, described as high, mid and low. Hombert found an effect of intrinsic f0 on the mid and high tones but not the low tone. According to Kingston's hypothesis, this is likely because an effect of vowel openness on the f0 of the low tone would lead to so much f0 variation that the low tone would encroach on the mid tone target. However, judging from Figure 7.1 this looks highly unlikely. The data points are averaged across two speakers and cannot reflect the amount of variation found, but

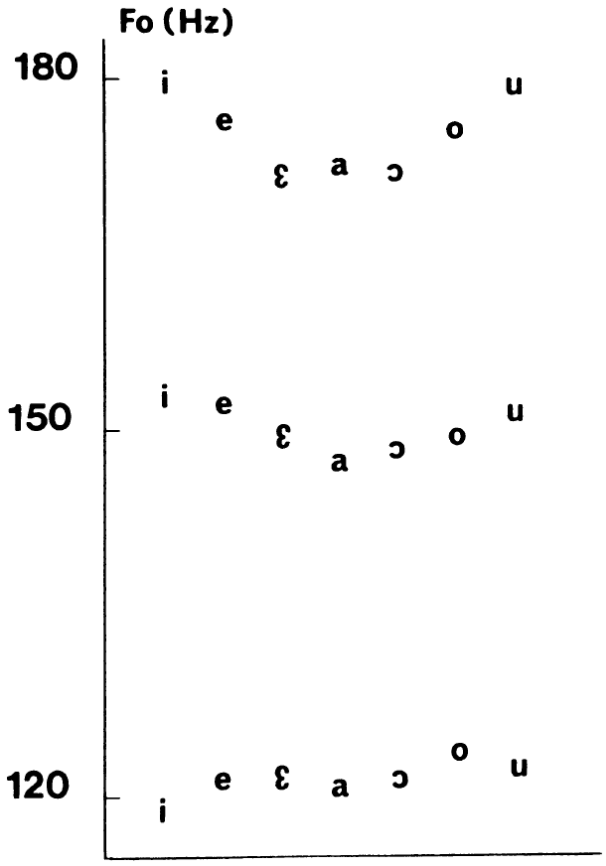


Figure 7.1: Average f0 (two speakers) as a function of vowel and tone (high, mid, low) in Yoruba (from Hombert, 1977a, Figure 2, p. 181)

even so it seems that if anything, it is the mid and high tones that are closer together (and ironically this looks to be at least partly a result of no close vowel f0-raising on the low tone). The low and mid tones are well separated and have considerable room for variation, so it appears highly improbable that any effect of vowel-intrinsic f0 would endanger this contrast. Thus, in our view it is unlikely that the absence of intrinsic f0 on the low tone in Hombert's data is because tonal overlap restricts auditory enhancement[2]. Instead, the most likely explanation in our view is the hybrid account of Hoole and Honda (2011). According to this account, vowel-intrinsic f0 is an automatic, mechanical result of contraction of the posterior genioglossus muscle during close vowel production. However, the effect may be counteracted by activation of the strap (or infrahyoid) muscles, especially

[2]By no means do we intend to imply that auditory enhancement is generally implausible. To the contrary: the model would provide elegant explanations for a number of phenomena. However, we strongly dispute the hypothesis that it is the underlying motivation behind vowel-intrinsic f0. If anything, the origin of intrinsic f0 must be automatic in nature and, once the listener picks up on this extra cue to vowel openness, they may use it to actively enhance the existing contrast (see Hoole & Honda, 2011).

the sternohyoid muscle, when a speaker actively lowers their f0, for example on a low tone. Auditory enhancement, on the other hand, results from exaggeration of the intrinsic f0 effect by actively controlling the cricothyroid muscle. Automatic intrinsic f0 might thus be restricted to the mid-upper f0 regions, while enhancement in the form of f0-raising on close vowels should be possible throughout a speaker's f0 space. Detailed discussions of the physiology behind biomechanical vs. active control of f0 can be found in Honda (1983, 2004); Hoole (2006); Hoole and Honda (2011).

Of course, this does not explain why the two tone languages cited above (i.e. Yoruba and Cantonese) differ in terms of the occurrence of intrinsic f0 on the low tone. Differing experimental methodologies and in particular sample sizes (our study used 20 speakers, Hombert's just two) are just one possibility. However, it is worth noting that with just seven vowels (Hombert, 1977a), Yoruba has a much less crowded vowel inventory than Cantonese (11 monophthongs alone according to Table 1.2.3 and an equal number of diphthongs in Figure 1.4, although the exact number of each varies by source). It is therefore possible that intrinsic f0 is fully automatic in both languages but additionally enhanced in Cantonese in order to aid vowel recognition[3]. According to this explanation, the absence of intrinsic f0 on the low tone in Yoruba would be a result of weakening biomechanical functions responsible for (automatic) intrinsic f0 while the presence of the effect in Cantonese could be attributed to active control; that is, enhancement of the intrinsic effect. A mixed physiological-enhancement approach might also explain other cross-linguistic variation in the size of intrinsic f0 effects (e.g. Jacewicz & Fox, 2015; Van Hoof & Verhoeven, 2011). In our opinion, however, there is one major problem with this account: intrinsic pitch at the low (22) vs. mid (33) tone boundary in Cantonese tone perception essentially means that the intrinsic f0 effect on the low (22) tone is factored out. As we argue in the next section, it is difficult to account for intrinsic pitch in a theory proposing that intrinsic f0 is actively controlled in order to enhance perceptual contrasts.

7.2 Intrinsic pitch: automatic or controlled?

While it was not the aim of this study to investigate the mechanisms behind vowel-intrinsic pitch, our data nevertheless add to a greater body of literature showing that the effect is inconsistent. This finding is, in itself, problematic for any account of intrinsic pitch as an automatic side effect of speech or general auditory processing. Specifically, in this section we focus on the relevance of our results for the virtual pitch model, gestural theories of speech perception and auditory enhancement theory.

7.2.1 Virtual pitch model

According to the virtual pitch model (Terhardt, 1974), vowel-intrinsic pitch may not be a phenomenon unique to speech perception. In this model, it is explained as a "pitch

[3]However, this explanation would be surprising given that Whalen and Levitt (1995) found no effect of vowel inventory size on intrinsic f0 in their meta-analysis of a number of languages.

shift" or "pitch deviation" on account of the spectral properties of complex tones. This model proposes that a pure tone and a complex tone with the same "spectral" pitch (approximately equivalent to fundamental frequency) may have different underlying "virtual" pitches. The virtual pitch of a pure tone matches its spectral pitch, but the virtual pitch of the complex tone is inferred from the subharmonic structure of the complex tone[4]. The difference between the two virtual pitches is the pitch shift or deviation. It is said to be affected by factors such as intensity, spectral structure, co-occurring sounds and inter aural pitch differences. In an early vowel perception study, Stoll (1984) played listeners a series of synthetic vowels as well as a pure tone with identical f0 and found that open vowels were judged as being higher in pitch than the close vowels, while intensity had a negligible effect on perceived pitch. Stoll noted a strong negative correlation between the intrinsic pitch effect in his data and intrinsic f0 as documented in previous studies. However, he argues that the effect cannot be attributed to compensation for coarticulation because it is small compared to the much larger pitch variation in everyday speech and should not require compensation. While he admits that the effect "can significantly exceed the pitch discrimination threshold" (Stoll, 1984, p. 138), his own data show that pitch shift due to vowel-intrinsic pitch roughly matches the JNDs found by Klatt (1973) and are largely predicted by the virtual pitch algorithm (Terhardt, Stoll & Seewann, 1982, in Stoll, 1984). Thus, he attributed the effect to pitch shift rather than compensation for coarticulation.

An explanation for intrinsic pitch as automatic based on the virtual pitch model is attractive. However, it does not exclude the possibility of a co-occurring compensation mechanism, which might help to explain the between-listener variation found in Stoll's pitch shift effect. Furthermore, there are other problems with this account. For example, as a general auditory phenomenon, pitch shift should not be restricted to certain f0 regions, vowel types or languages. Our data show the effect in connection with monophthongs at the low (22) vs. mid (33) level tone boundary only but not at the mid (33) vs. (55) high tone boundary or on diphthongs. Pape's (2009) cross-linguistic investigation found that intrinsic pitch applied to some languages but not others. As it stands, the virtual pitch model cannot account for the multitude of different findings on vowel-intrinsic pitch.

7.2.2 Gestural theories of speech perception

From a strict gestural perspective in the tradition of motor theorists and direct realists, the object of speech perception in this case would be the articulatory gestures responsible for the intrinsic f0 effect. As such, intrinsic pitch would be an automatic side effect of normal vowel perception and the size of the perceived intrinsic pitch effect should be equal to the intrinsic f0 effect found in production. As gesturalists have pointed out themselves (e.g. Fowler & Brown, 1997), this does not appear to be the case (see Hombert et al., 1979; Silverman, 1987; Fowler & Brown, 1997; Pape, 2009). Where intrinsic pitch does occur, the effect measured is smaller than that found for intrinsic f0. Even more troubling for this

[4]In this way, the model also accounts for successful extraction of the missing fundamental in perception of complex tones.

account is the conflicting evidence as to whether intrinsic pitch is universal: the effect has been found in some languages but not others (cf. Pape, 2009). In Chapter 4, we found that the relationship between intrinsic f0 and intrinsic pitch was quite well-matched for the low (22) vs. mid (33) level tones in Cantonese, but for the mid (33) vs. high (55) tone contrast the perceptual effect was only partial compared with intrinsic f0 in production. The "incomplete" parsing of intrinsic f0 at the higher end of the Cantonese tone space thus does not fit well in a gestural theory of speech perception. We agree that intrinsic pitch may be seen as perceptual normalisation for intrinsic f0 (or "compensation for coarticulation"). However, this does not require that the effect be complete or even automatic. Just as the speaker adapts to the requirements of the speech setting in Lindblom's theory of hyper- and hypospeech (Lindblom, 1990), we can conceive of a situation in which the listener is equally adaptive. Thus, where coarticulation would be detrimental to a phonological contrast but is unavoidable (even in hyperarticulated speech), the listener may take particular care in compensating (more) for coarticulation. This would explain the presence of intrinsic pitch at the low (22) vs. mid (33) level tone boundary but not at the mid (33) vs. high (55) level tone boundary in Experiment 2 (Chapter 3), as it is the low (22) vs. mid (33) tone contrast that is most ambiguous. It would also be compatible with the lack of any intrinsic pitch effect in Experiment 3 (Chapter 5), because in this case there was no "coarticulation" (in the form of intrinsic f0) in production (Chapter 6) requiring compensation. One might wonder, then, why intrinsic pitch should occur at all in, for example, non-tone languages in which there is no tone contrast to maintain. In fact, f0 is known to be a cue to vowel openness (Glidden & Assmann, 2004; Reinholt Petersen, 1986)[5], and thus intrinsic pitch may also play a role in distinguishing phonological vowel contrasts. From this perspective, it is plausible that intrinsic pitch is a form of context-specific compensation for coarticulation rather than an automatic effect based on perception of articulatory gestures.

7.2.3 Auditory enhancement

Auditory enhancement theory (e.g. Kingston, 1992, 2007; Kingston et al., 2014) envisages a perceptually motivated theory of adaptive dispersion (Liljencrants & Lindblom, 1972). According to a strict interpretation of this theory, vowel-intrinsic f0 is not a phonetic effect of vowel openness but a controlled method of enhancing the vowel openness contrast for maximal perceptual distance. It is not the object of this section to list the arguments for

[5]Incidentally, Reinholt Petersen considers the vowel-intrinsic pitch effect to be "too small to be of any importance to speech perception" but contradicts himself in the very next sentence by stating that intrinsic f0 "can function as a cue both to segment identity and to the identity of prosodic categories, depending on the actual demands of the speech perception system" (1986, p. 31). If intrinsic f0 acts as a perceptual cue, then in our view it is difficult to argue that it is not relevant for speech perception. Furthermore, the interdependence between tone and vowel openness in Experiment 2 (Chapter 3) and in Zheng (2014) for Mandarin adds additional weight to the argument that vowel-intrinsic pitch is indeed important for speech perception. Klatt (1973) conducted a thorough investigation of just-noticeable differences (JNDs) in the f0 of synthetic vowels and found that differences of just 0.3-2Hz were detectable. These differences are smaller than those generally reported for vowel-intrinsic f0 (e.g. Whalen & Levitt, 1995).

and against auditory enhancement in detail[6]. To our knowledge auditorists have often used vowel-intrinsic f0 (the effect in production) as evidence of auditory enhancement (Kingston, 2007) without addressing vowel-intrinsic pitch (the perceptual effect).

We address this oversight here because we believe cross-linguistic intrinsic pitch patterns have important consequences for this theory. In the above, we have shown that our data support neither a general auditory process nor a speech-specific, gesturally motivated explanation for vowel-intrinsic pitch. As such, so far we have concluded that intrinsic pitch is probably language-specific and controlled. Having said that, our data also do not seem to support the auditory enhancement account, and in fact our view is that any evidence of vowel-intrinsic pitch is, by definition, evidence against auditory enhancement theory.

Vowel-intrinsic pitch is negatively correlated with vowel-intrinsic f0. All else being equal, open vowels tend to have a lower f0 than close vowels, while at equal f0 open vowels tend to sound higher in pitch than close vowels. As we have established, both in our own data and in previous studies, intrinsic pitch and intrinsic f0 are seldom well-matched. However, where intrinsic pitch does occur, it undeniably "cancels out" intrinsic f0, at least to some degree. If vowel-intrinsic f0 is perceptually motivated and its goal is to enhance vowel openness contrasts, why would the listener then implement intrinsic pitch and, in doing so, "filter out" the speaker's enhancement efforts? It makes no sense to enhance a contrast in production that will only be un-enhanced in perception.

In Experiment 1 (Chapters 2), we found a clear effect of vowel openness on the f0 of high (55) level tone tokens as well as an effect of vowel openness on machine classification of low (22) level tone tokens. In Experiment 2 (Chapter 3), intrinsic pitch applied to the low (22) vs. mid (33) level tone contrast only. We concluded in Chapter 4 that the intrinsic pitch effect at this low-mid tone contrast appeared roughly equal to the intrinsic f0 effect, which probably reflects normalisation for the intrinsic f0 effect on the low (22) tone and not on the high (55) tone. Given this result, and considering the Cantonese tone system with its considerable crowding in the lower tone regions, auditorists might argue that the enhancement effect (intrinsic f0) endangers tone contrasts in the lower tone regions and thus needs to be filtered out (intrinsic pitch) there, whereas it does not endanger any tone contrasts in the upper tone regions. Yet this explanation seems overly complicated to us - surely if an enhancement does not facilitate communication to begin with, then we should have found neither an effect of intrinsic f0 nor of intrinsic pitch on the low-mid tones rather than evidence of the two phenomena cancelling each other out. In our view, this is evidence of intrinsic f0 being an automatic, uncontrolled, phonetic side effect of vowel articulation, while intrinsic pitch reflects a more context-specific, controlled mechanism.

To summarise, while we believe it is quite likely that auditory enhancement occurs and is very useful in many speech situations, intrinsic pitch works against intrinsic f0 and is therefore incompatible with any active enhancement of intrinsic f0.

[6]For an overview, see for example Kingston (2007) and references cited therein.

7.2.4 Intrinsic pitch in tone languages

Our data certainly indicate that intrinsic pitch can occur in tone languages, which to our knowledge has only been documented so far by Zheng (2014) for Mandarin. However, contrary to Niebuhr's finding (2004), the effect was restricted to monophthongs and did not transfer to diphthongs in our experiment. Because intrinsic f0 was also restricted to monophthongs in our data, this could be seen as evidence for intrinsic pitch as compensation for coarticulation only where necessary and hence for intrinsic pitch as a controlled phenomenon.

7.3 Microprosody, macroprosody and the phonetics-phonology interface

Overall, we find interactions between microprosody and macroprosody in Cantonese, but these are context-specific. Compensation for coarticulation appears to occur where phonological contrasts would otherwise be confused.

7.3.1 Sound change

According to models of sound change in which compensation for coarticulation plays a central role (Ohala, 1981, 1993; Beddor, 2009; Harrington et al., 2008), we find that for Cantonese, compensation appears to be relatively complete within the speech community as a whole, if not for individual speaker-listeners. Thus, despite asymmetries in the effects of intrinsic f0 on Cantonese level tones, the low (22) vs. mid (33) tone contrast in Cantonese does not appear to be at risk of tonal sound change due to vowel openness. If anything, the relationship between production (intrinsic f0) and perception (intrinsic pitch) is mismatched for the mid (33) vs. high (55) tone contrast. We argued that full compensation for coarticulation was not necessary in order to main this tone contrast, but this does not exclude that tiny shifts in speaker-listener's exemplars due to vowel openness do not occur. As such, if anything we would predict tonal sound change due to vowel openness on the high (55) level tone, although in the absence of similar sound changes this seems unlikely.

We concluded that intrinsic f0 is most likely a true coarticulatory side effect rather than a type of enhancement and additionally that intrinsic pitch is clearly audible for monophthongs. Although others have argued that in order for phonetically biased sound change to occur the sound pattern needs to be both universal and perceptible (Hombert, 1977a; Hombert et al., 1979), our data show that both of these criteria apply to vowel-intrinsic f0 despite the lack of hard diachronic evidence of this type of sound change. This indicates that there must be further mechanisms behind phonetically-biased sound change that prevent vowel openness from causing tonal sound change.

7.4 Tone processing

Crucially, our evidence from both production and perception studies reveals the importance of integrating vowel-intrinsic f0 and vowel-intrinsic pitch into accounts of speech and specifically tone processing. Even complex tone languages that manipulate f0 for linguistic distinctions are susceptible to intrinsic effects of vowel openness on f0, as demonstrated in Chapter 2. Thus, f0 in such languages is simultaneously influenced by both the macroprosody (tone) and the microprosody (vowel-intrinsic f0). It is therefore crucial that listeners distinguish between these two determinants of f0 in order to parse tonal contrasts successfully. Our listeners were aware of vowel-intrinsic pitch on Cantonese monophthongs and adjusted their perception of level tones accordingly, at least for the low (22) vs. mid (33) level tone contrast, which is inherently less distinct than other contrasts in Cantonese. For the mid (33) vs. high (55) level tones, where the f0 contrast is inherently more distinct, compensation for intrinsic f0 effects is less important for successful tone processing, because the magnitude of the effect of vowel on f0 is negligible compared with the effect of tone on f0.

This is also valid for improvement of speech recognition as well as synthesis of natural-sounding (and understandable) speech. In particular, the results of machine classification in Chapter 2 highlighted the importance of training speech recognition technology with vowel-intrinsic f0 information in order for tones to be identified correctly. This is especially important for tone contrasts that are not clearly distinctive, such as the Cantonese low (22) vs. mid (33) level tone contrast. In addition, given that listeners adjusted for vowel-intrinsic pitch in tone perception (cf. Chapter 3), it follows that integrating vowel-intrinsic f0 effects into text-to-speech systems would improve naturalness and possibly even intelligibility ratings.

7.5 Further work

This study investigated an interaction between segmental (vowel) and suprasegmental (tonal) determinants of f0. The results demonstrate that, at least for monophthongs, both vowels and tones affect f0 in speech production, and that listeners are able to successfully distinguish between these two sources in most cases. As a next step, it would be equally important to establish whether segmental determinants of f0 influence suprasegmental patterns at a more global level; that is, whether they influence intonation as well as tone; and whether listeners are equally good at distinguishing between segmental and suprasegmental sources of f0 in such contexts.

One of the most perplexing findings of this study was the lack of statistically significant intrinsic f0 and intrinsic pitch effects on diphthongs and triphthongs. It is difficult to comment on whether this (non) effect is specific to our data or to Hong Kong Cantonese or whether it represents a phonetic universal. Further studies on intrinsic f0 in diphthongs of other languages, especially those that have both true closing and true opening diphthongs, are in order. If our results can be replicated for other languages, it will have consequences

for theories of the mechanism(s) underlying intrinsic f0. If dynamic vowels such as diphthongs have no intrinsic f0 of their own, why not? What is it about their production that suppresses an effect that is otherwise so robust on monophthongs? Diphthongs may well pose an interesting test case for improving our understanding of intrinsic f0. For more reliable results, is vital that any follow-up studies collect data from a considerably larger group than our production experiment in Chapter 6. Although somewhat more difficult to collect, phsyiological data comparing laryngeal and articulatory configurations of diphthongs with monophthong sequences would be a more direct method of examining the underlying causes of intrinsic f0 effects.

Appendices

Appendix A

Translations of stimuli

A.1 Experiment 1

Stimulus	Prompt	English translation
tɐŋ$_{55}$	燈	*lamp*
tɐŋ$_{33}$	凳	*chair*
tɐŋ$_{22}$	鄧	(common family name)
taːn$_{55}$	丹	*pill, powder*
taːn$_{33}$	誕	*birth*
taːn$_{22}$	但	*but*
tin$_{55}$	顛	*upside down*
tim$_{33}$	店	*shop*
tin$_{22}$	電	*electricity*
tyn$_{55}$	端	*beginning*
tyn$_{33}$	鍛	*exercise*
tyn$_{22}$	段	*section, piece*

Table A.1: English translations of the stimuli used for Experiment 1 in Chapter 2 (cf. Table 2.1).

A.2 Experiment 4

Stimulus	Prompt	English translation
maːi$_{23}$	買	*buy*
maːi$_{22}$	賣	*sell*
maːi$_{21}$	埋	*bury*
maːu$_{23}$	牡	*peony*
maːu$_{22}$	貌	*look (n.)*

ma:u$_{21}$	矛	*spear*
ŋɐi$_{23}$	蟻	*ant*
ŋɐi$_{22}$	偽	*false*
ŋɐi$_{21}$	危：危險	*danger*
ŋɐu$_{23}$	偶：偶然	*accidentally*
ŋɐu$_{22}$	吽：發吽逗	*stare blankly, lost in one's own thought*
ŋɐu$_{21}$	牛	*cow*
jɐn$_{25}$	讔: 讔喻	*hidden meaning, metaphor*
jɐn$_{23}$	蚓：蚯蚓	*earthworm*
jɐn$_{22}$	刃：刀刃	*knife edge*
jɐn$_{21}$	人	*people*
wɐn$_{25}$	穩: 穩定	*stable*
wɐn$_{23}$	尹	*family name "Wan"*
wɐn$_{22}$	運	*transport*
wɐn$_{21}$	雲	*cloud*
wa:n$_{23}$	挽: 挽回	*pull back*
wa:n$_{22}$	幻	*illusion, fantasy*
wa:n$_{21}$	鬟: 丫鬟	*maid*
ji$_{55}$	醫	*doctor*
ji$_{25}$	綺	*colourful*
ji$_{33}$	意	*meaning*
ji$_{23}$	耳	*ear*
ji$_{22}$	貳	*two*
ji$_{21}$	移	*move*
jɐu$_{55}$	幽	*quiet, dark*
jɐu$_{25}$	柚	*pumelo,grapefruit*
jɐu$_{33}$	幼	*infant*
jɐu$_{23}$	友	*friend*
jɐu$_{22}$	又	*again*
jɐu$_{21}$	油	*oil*

Table A.2: English translations of the stimuli used for Experiment 4 in Chapter 6 (cf. Table 6.1).

Appendix B

By-listener pre-examination of perception data in Chapter 3

In Chapter 3, Section 3.2.4, we checked data for three types of errors. For full disclosure, the pre-examined data is depicted here.

We determined a listener to have an A error when one or both endpoints of a continuum was not firmly in a specific tone category, i.e. they showed either no categorical perception of tone or they might have benefited from a wider continuum. B errors reflected an inability to calculate slopes, which occurred for speakers who perception was extremely categorical. Finally, C errors reflected deviations of 10% (i.e. one click in the experiment) from 1.0 or 0.0 at one or both endpoints of a continuum, interpreted as being mistakes.

Figures B.1 to B.4 show each of the four combinations of Tone Condition and Vowel Openness, and the errors identified for each model are described in the captions. The three types of errors were handled differently; see Section 3.2.4 for details. Each contour represents the average of one speaker.

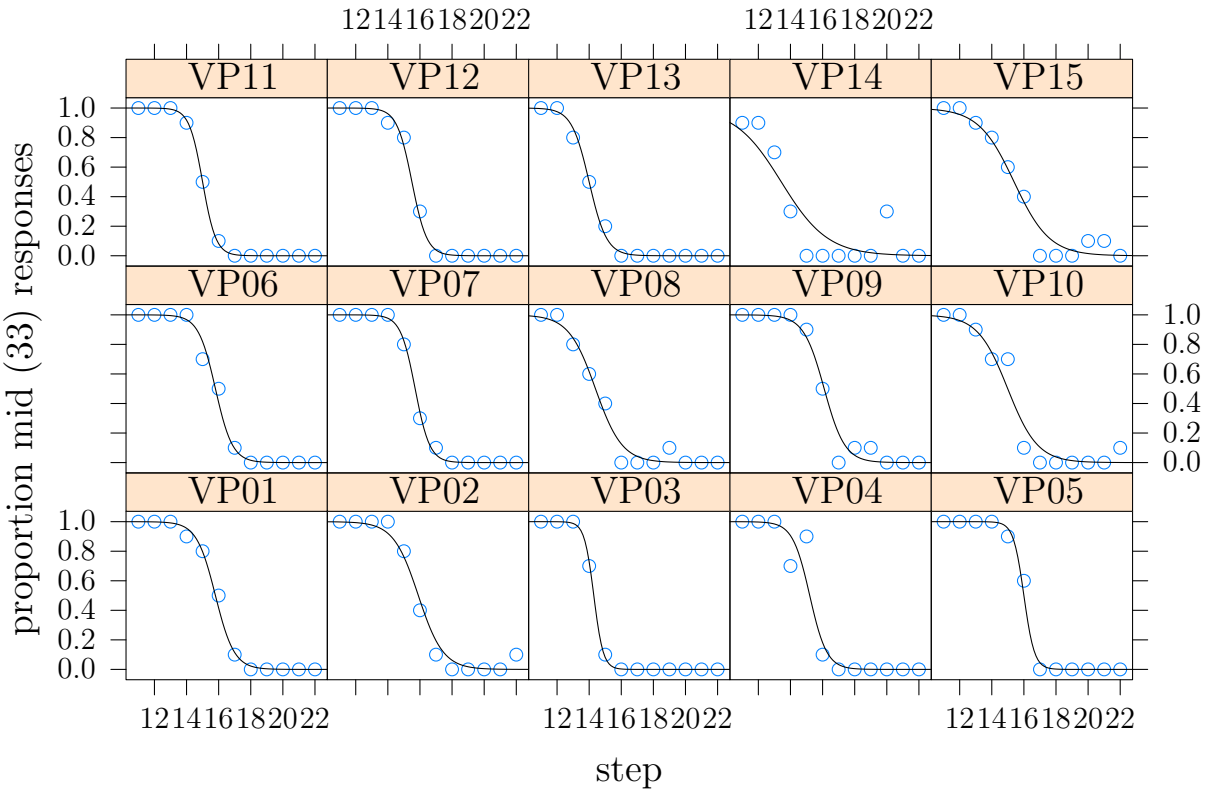


Figure B.1: Mid (33) vs. high (55) contrast on close vowels: proportion of mid (33) tone responses for steps 12 to 22 of the continuum. An A error was identified for speaker VP14, who was excluded from the analysis for this condition. No B errors were identified for this model. C errors occurred (and were corrected) at step 22 of the continuum for speakers VP02 and VP10.

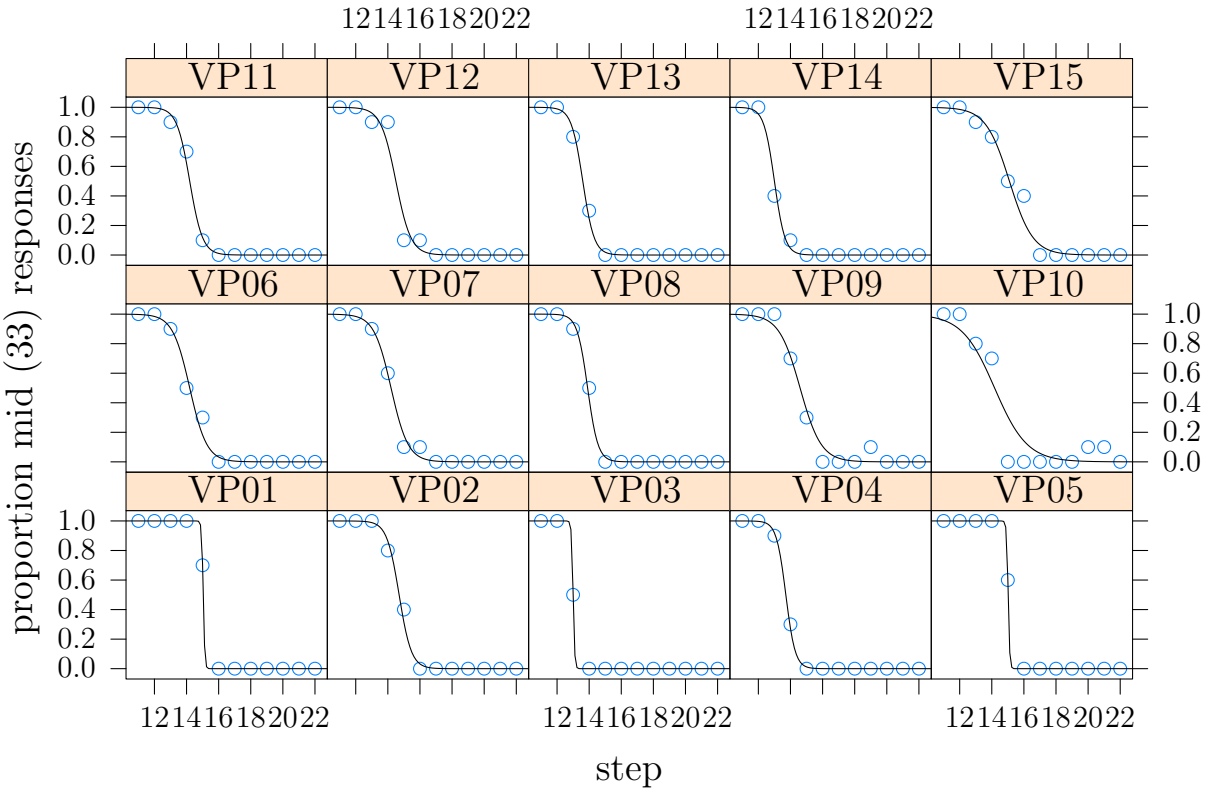


Figure B.2: Mid (33) vs. high (55) contrast on open vowels: proportion of mid (33) tone responses for steps 12 to 22 of the continuum. A errors were identified for speakers VP01, VP03 and VP05, who were excluded from the analysis for this condition. No B or C errors were identified for this model.

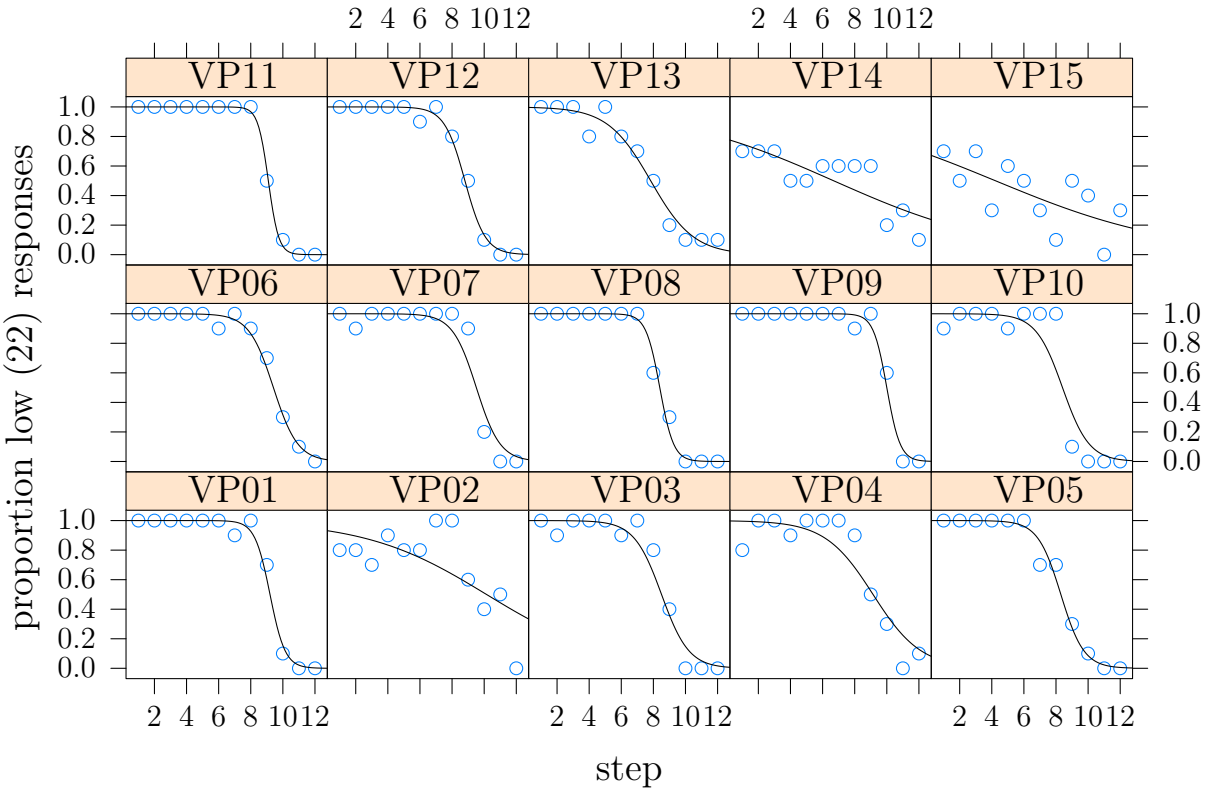


Figure B.3: Low (22) vs. mid (33) contrast on close vowels: proportion of low (22) tone responses for steps 1 to 12 of the continuum. A errors were identified for speakers VP02, VP13, VP14 and VP15, who were excluded from the analysis for this condition. No B errors were identified for this model. C errors occurred (and were corrected) at step 1 of the continuum for speakers VP04 and VP10 and at step 12 for speaker VP04.

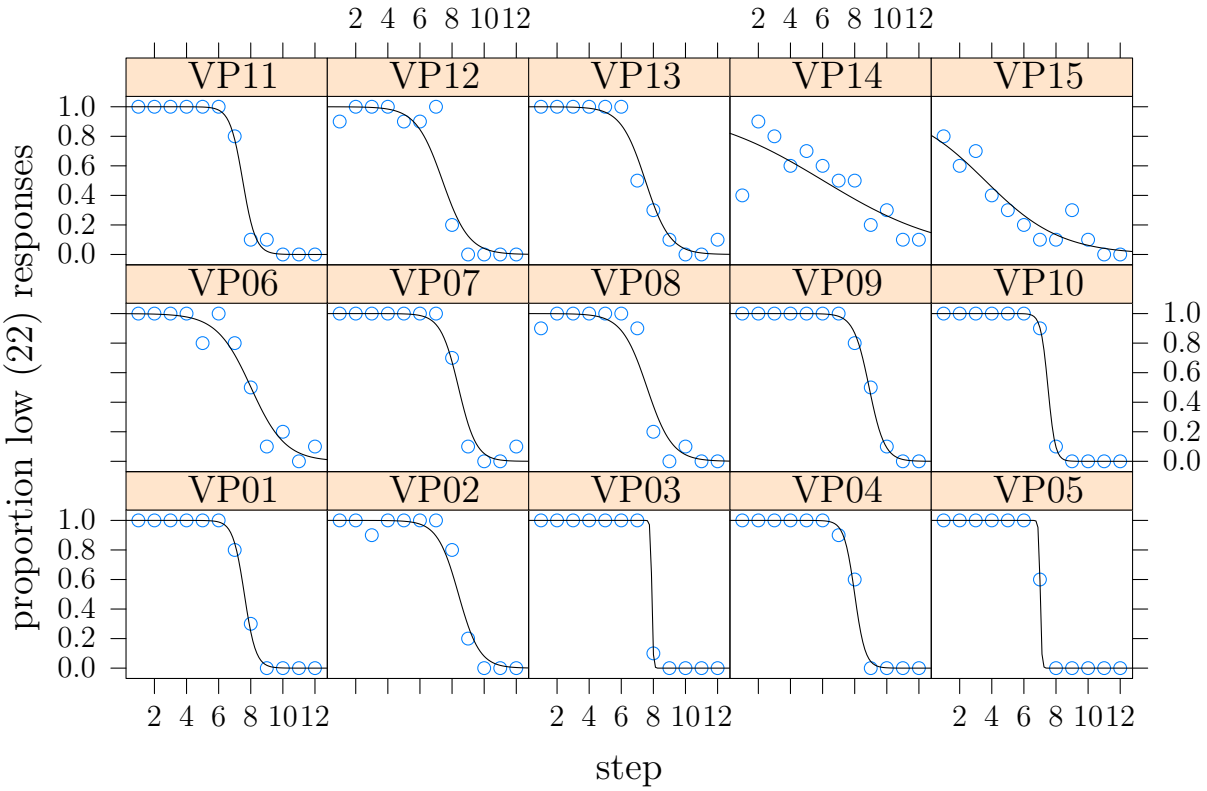


Figure B.4: Low (22) vs. mid (33) contrast on open vowels: proportion of low (22) tone responses for steps 1 to 12 of the continuum. A errors were identified for speakers VP03 and VP05, who were excluded from the analysis for this condition. B errors were identified for speakers VP14 and VP14. C errors occurred (and were corrected) at step 1 of the continuum for speakers VP08 and VP12 and at step 12 for speakers VP06, VP07 and VP13.

Appendix C

Spectrograms of perception stimuli from Chapter 5

The spectrograms included here are of the resynthesised stimuli used for Experiment 3 described in Chapter 5. The blue lines reflect the f0 contours, with f0 (Hz) plotted on the x-axis for guidance. Figures C.1 to C.10 show the 10 steps on the f0 continuum (from falling to rising, respectively) on the /maːu/ target. Figures C.11 to C.20 show the same 10 steps on the /waːn/ target.

For details about the resynthesis, see Chapter 5, Section 5.2. Figures 5.1 and 5.3, also in Section 5.2, show the spectrograms and overlaid f0 contours for the natural productions used as the basis for the resynthesised stimuli displayed here.

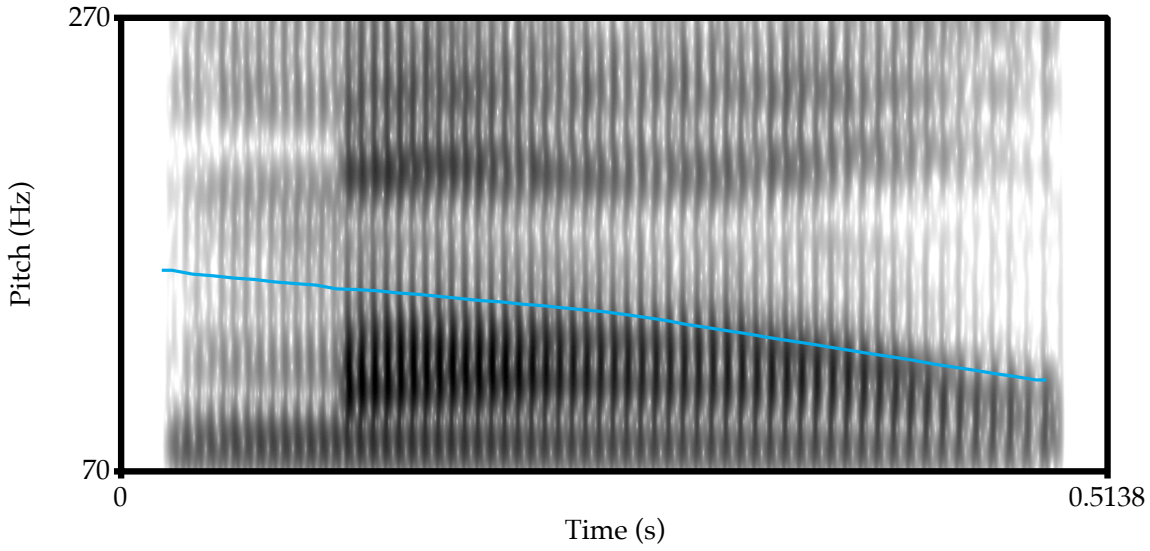


Figure C.1: /maːu/ step 1

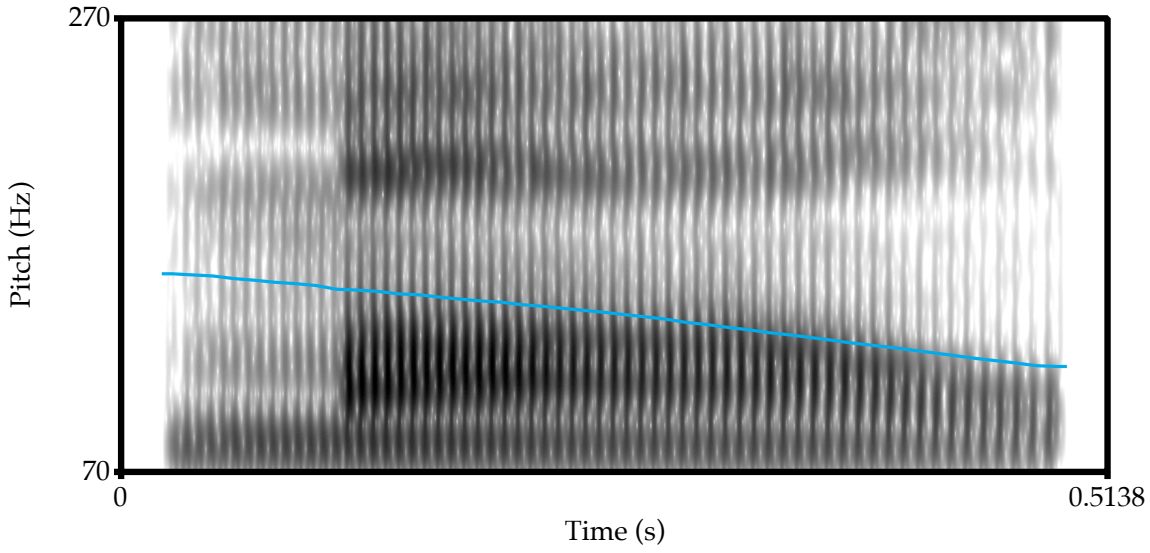


Figure C.2: /maːu/ step 2

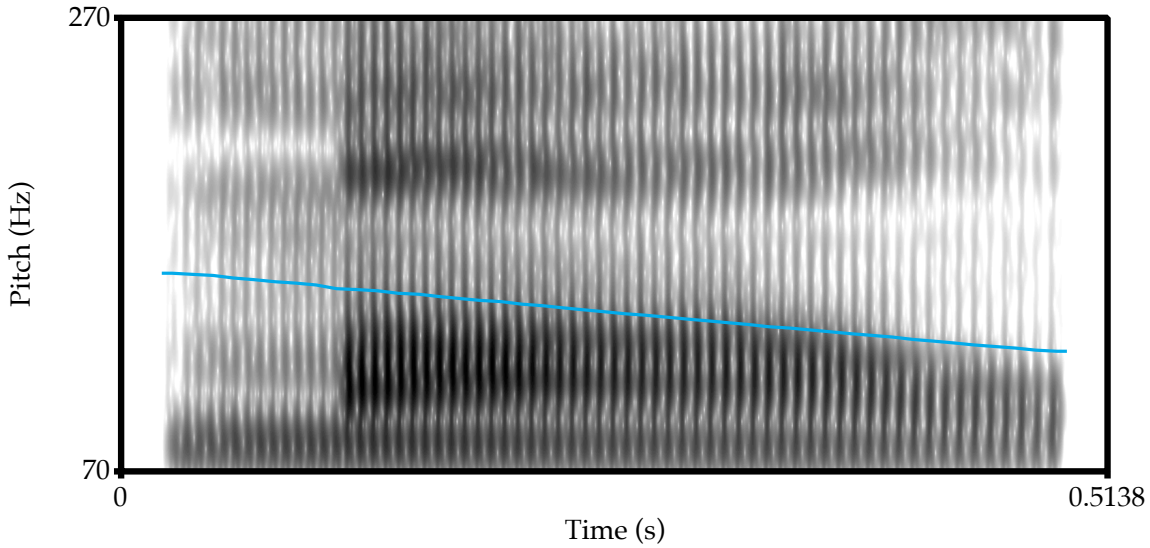


Figure C.3: /maːu/ step 3

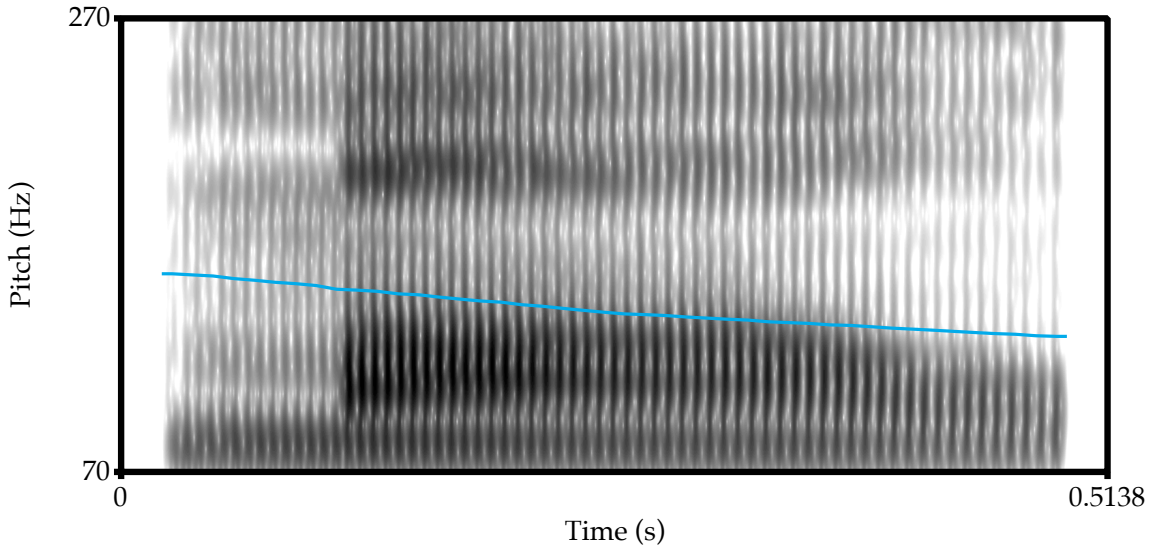


Figure C.4: /maːu/ step 4

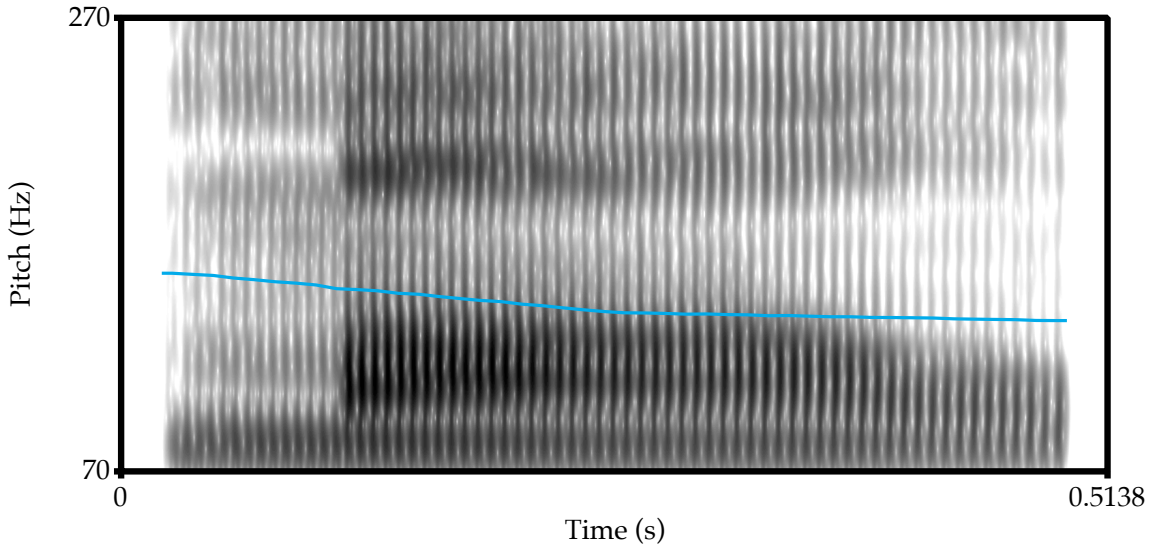


Figure C.5: /maːu/ step 5

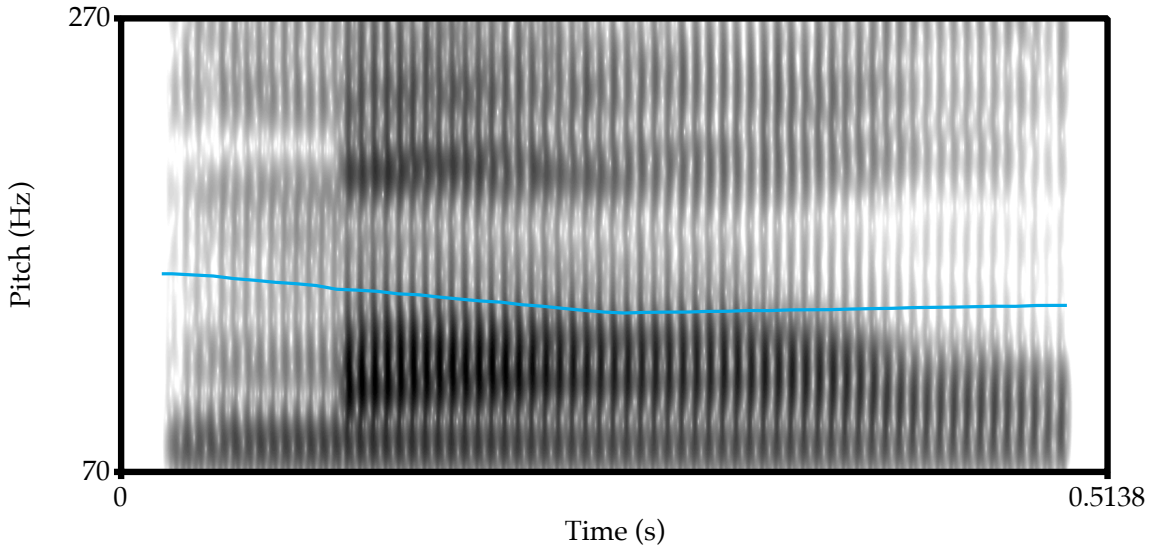


Figure C.6: /maːu/ step 6

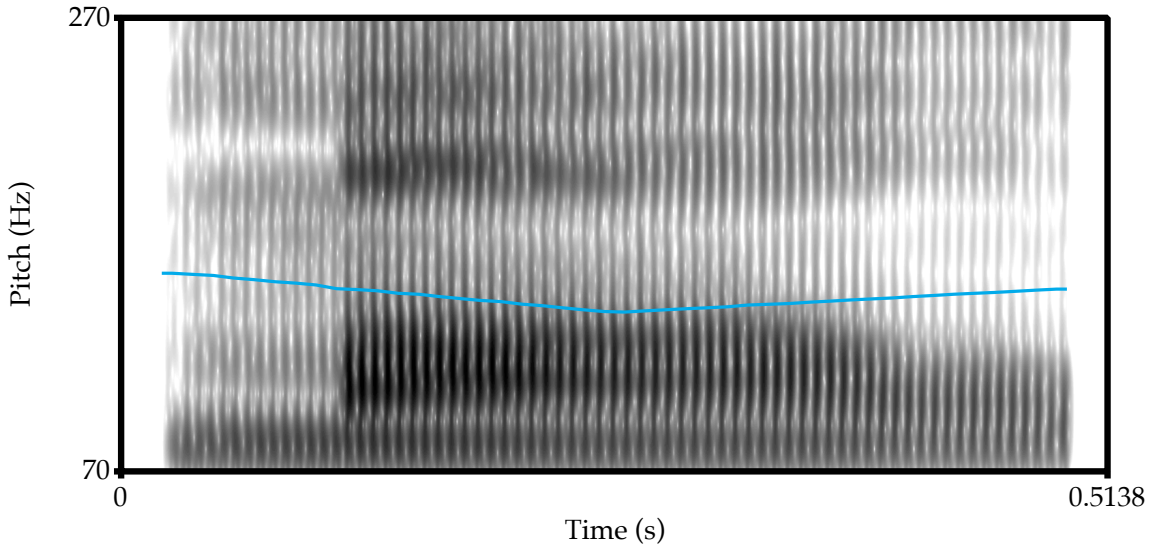


Figure C.7: /maːu/ step 7

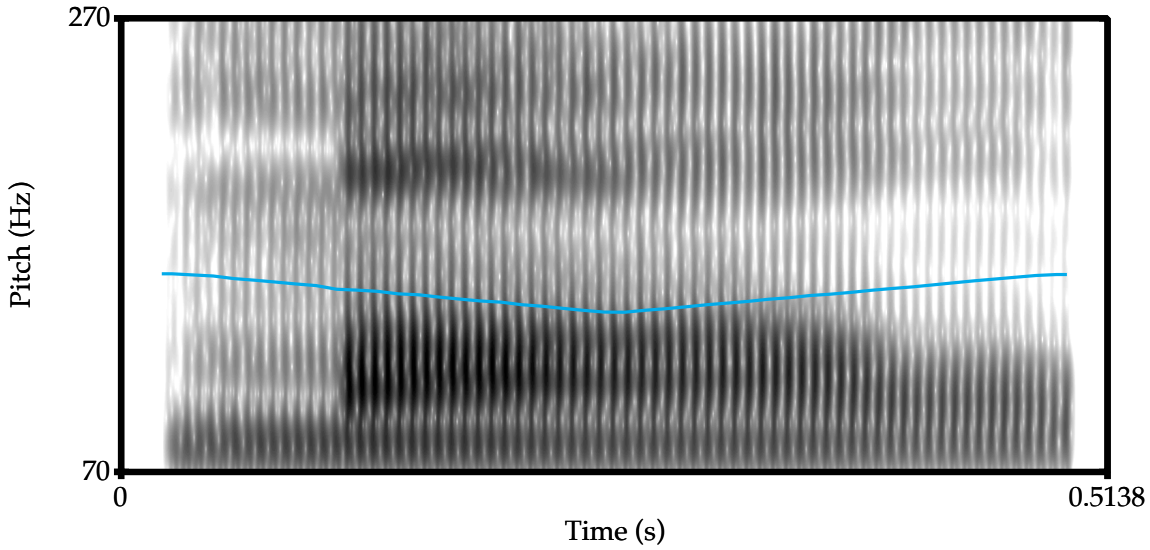


Figure C.8: /maːu/ step 8

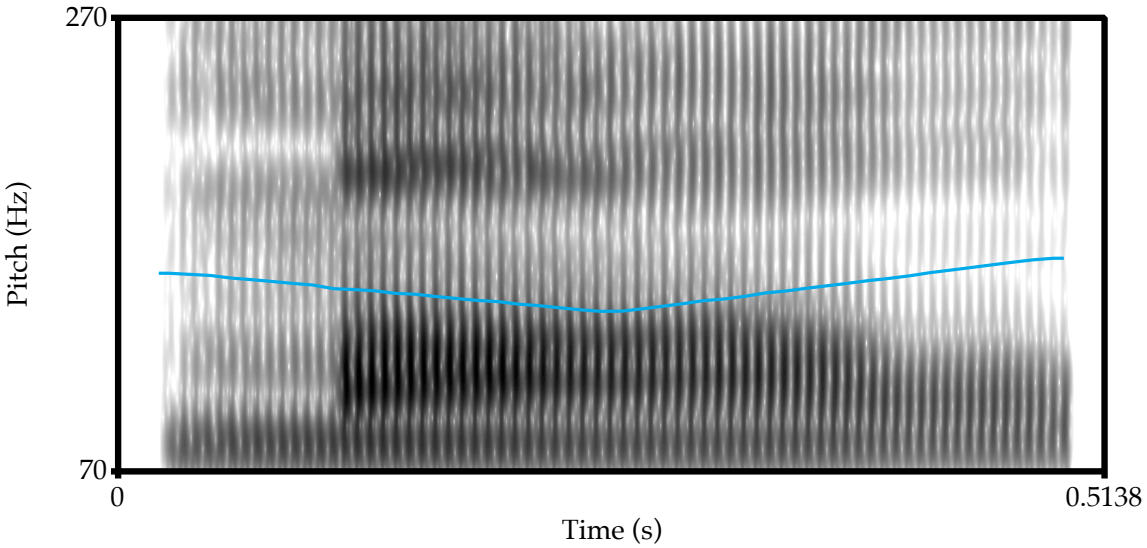


Figure C.9: /maːu/ step 9

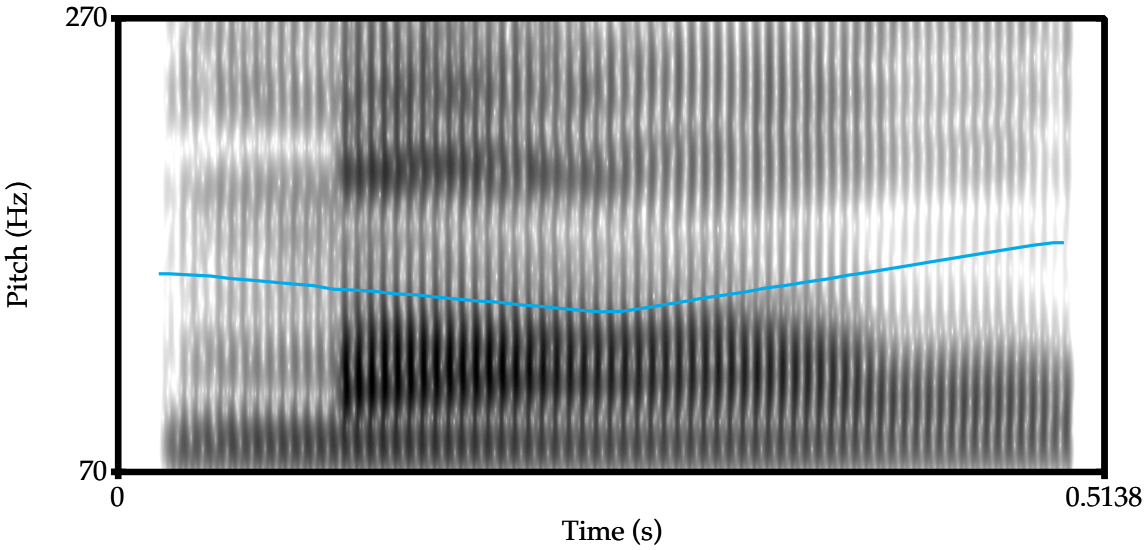


Figure C.10: /maːu/ step 10

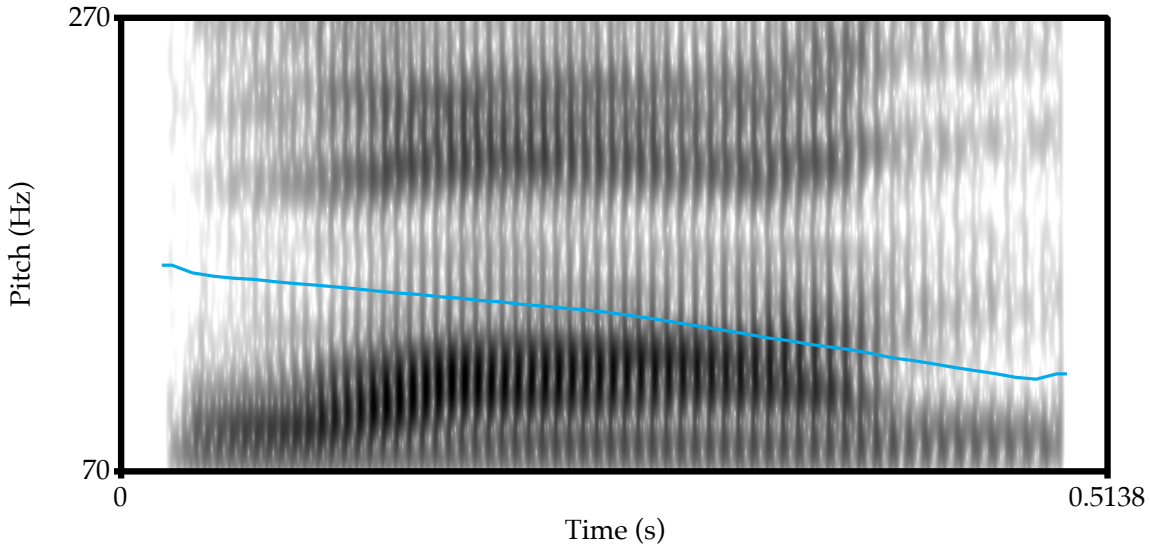


Figure C.11: /waːn/ step 1

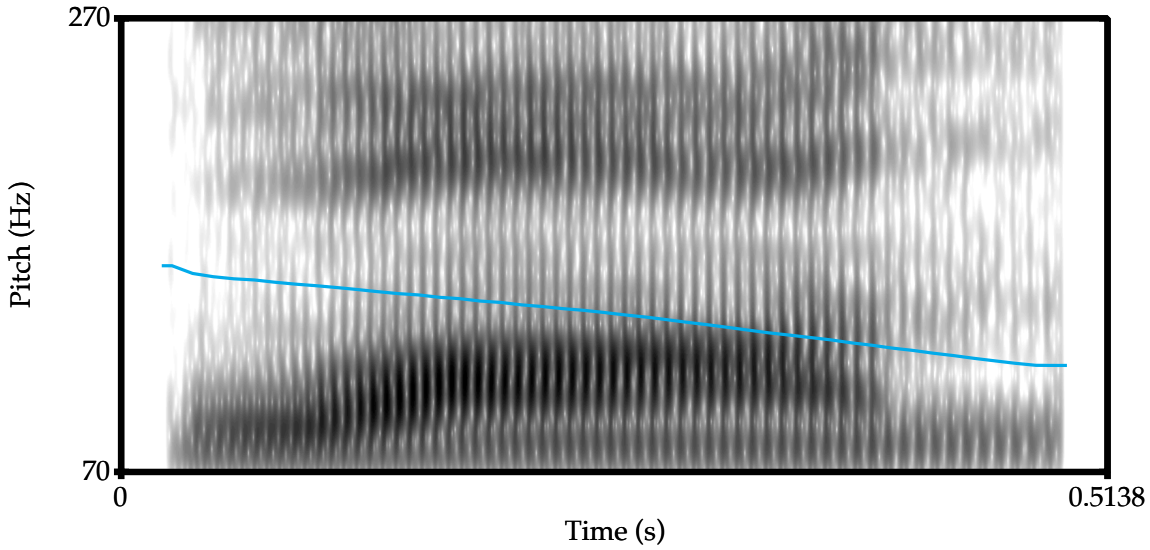


Figure C.12: /waːn/ step 2

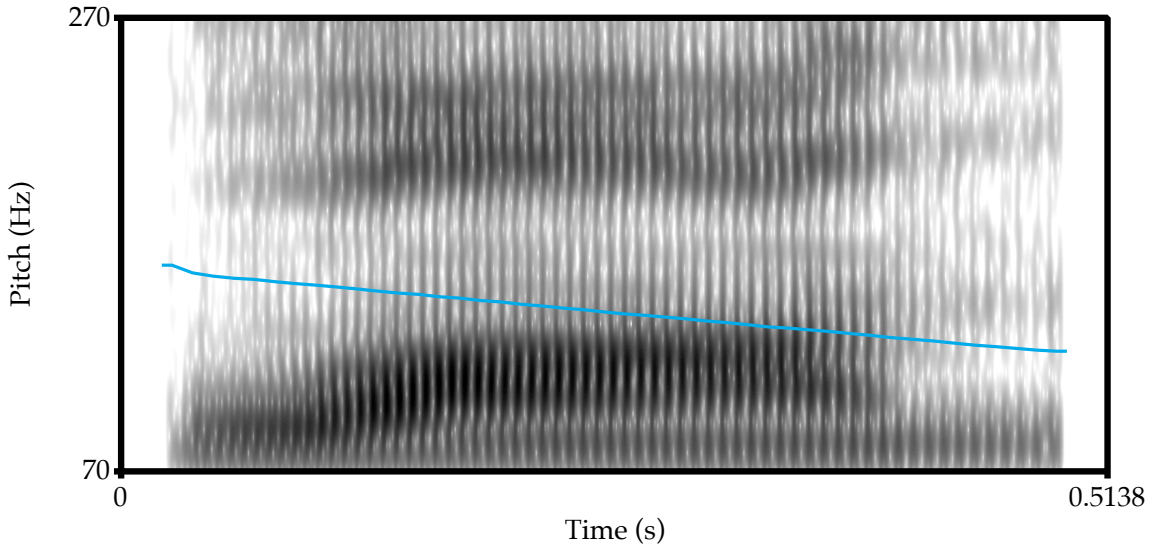


Figure C.13: /waːn/ step 3

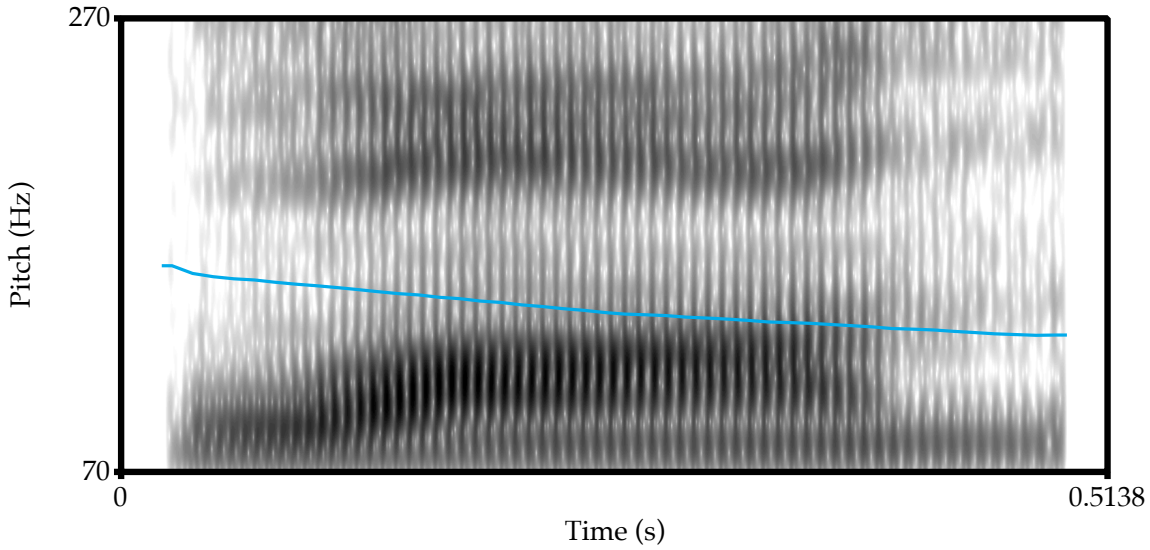


Figure C.14: /waːn/ step 4

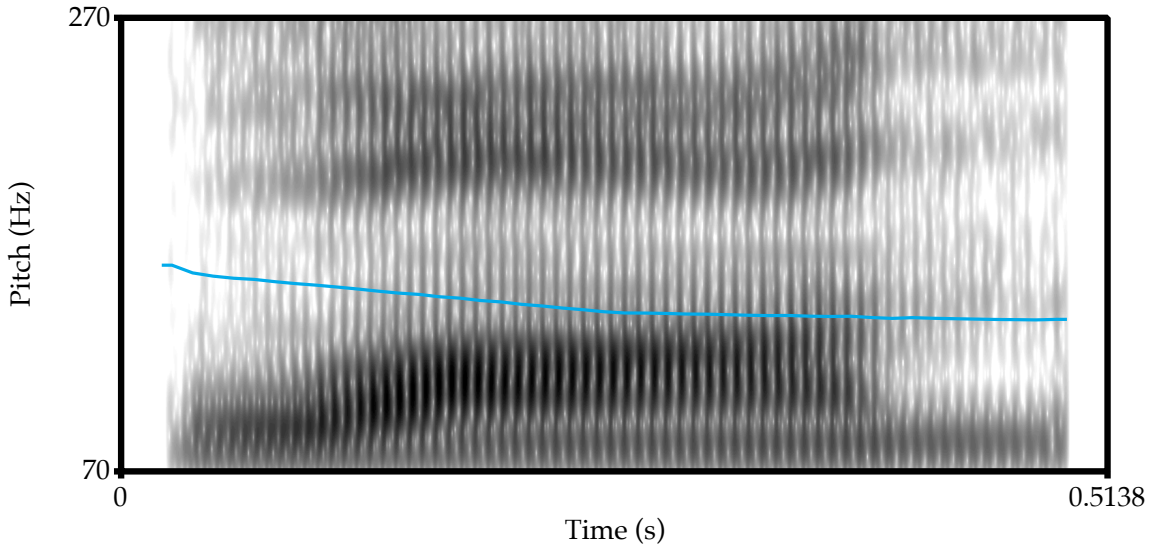


Figure C.15: /waːn/ step 5

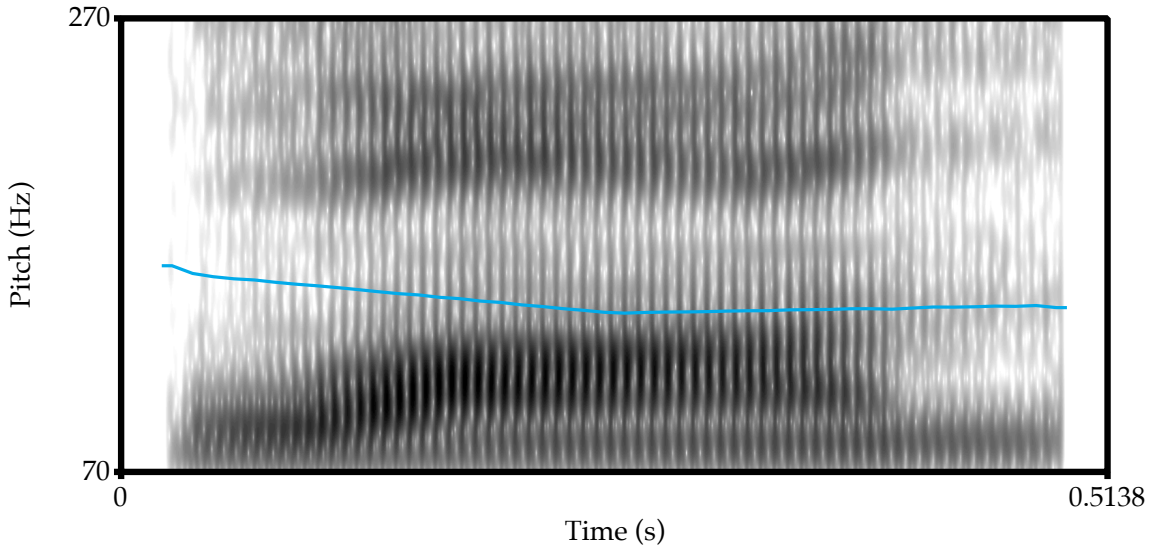


Figure C.16: /waːn/ step 6

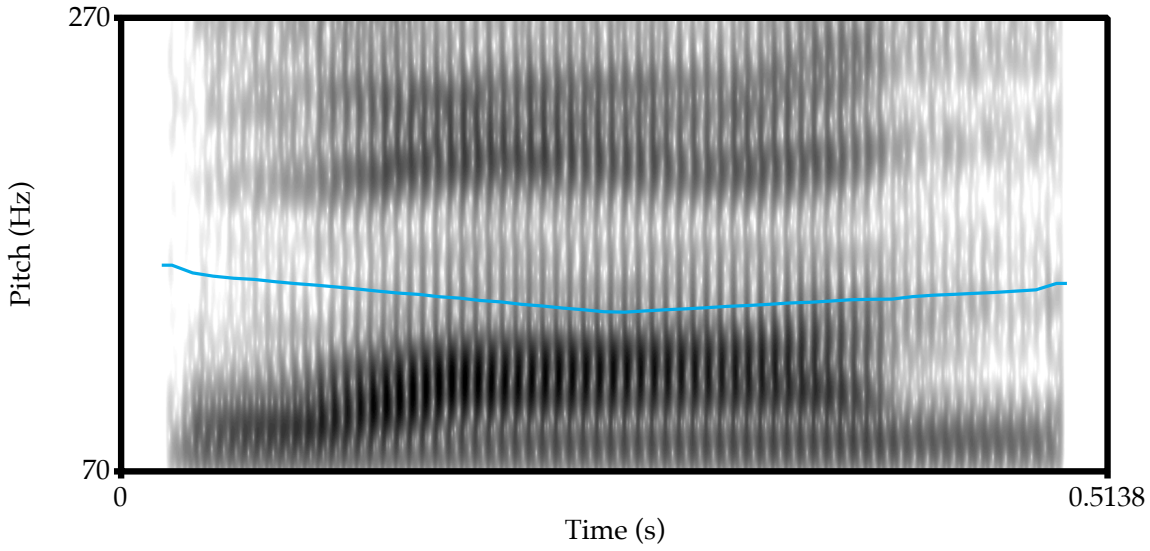


Figure C.17: /waːn/ step 7

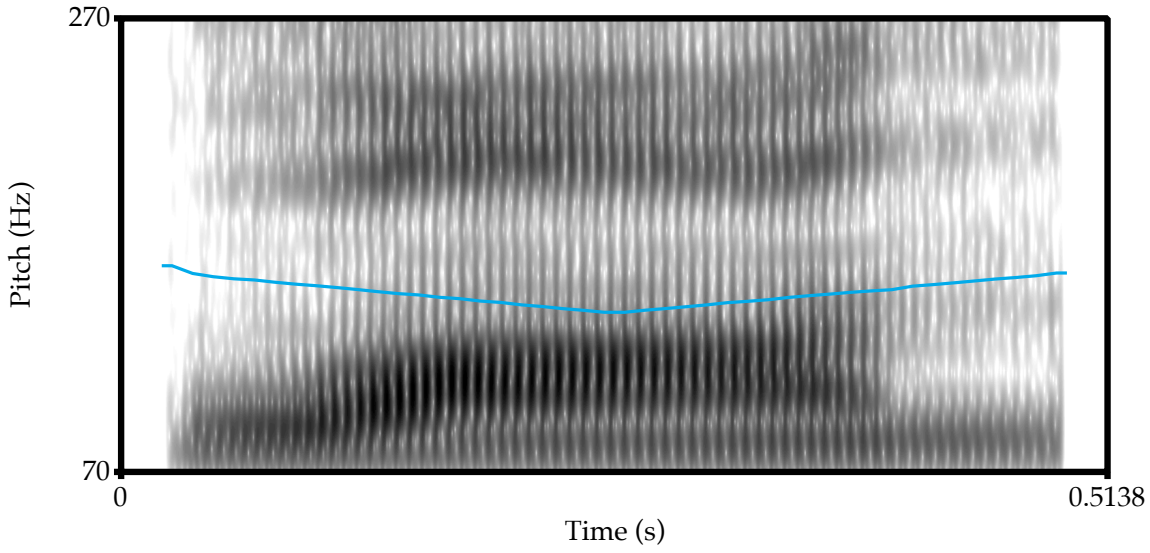


Figure C.18: /waːn/ step 8

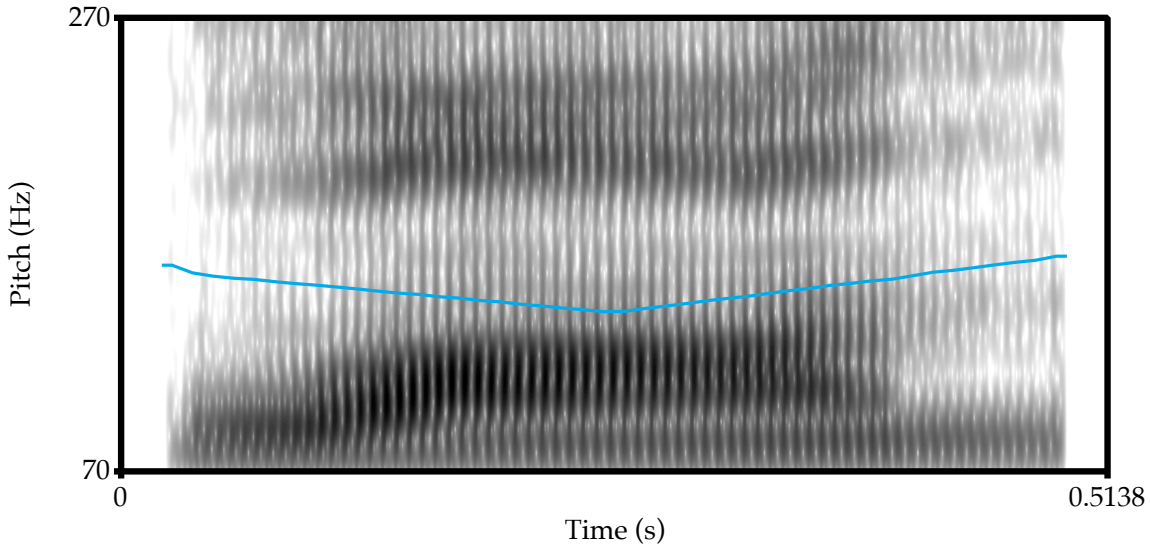


Figure C.19: /waːn/ step 9

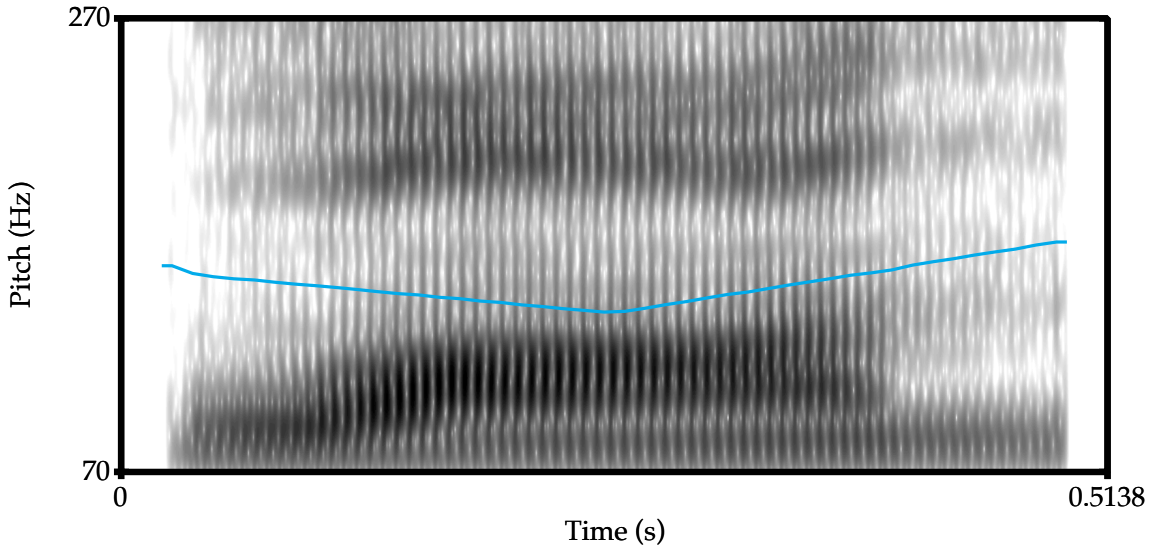


Figure C.20: /waːn/ step 10

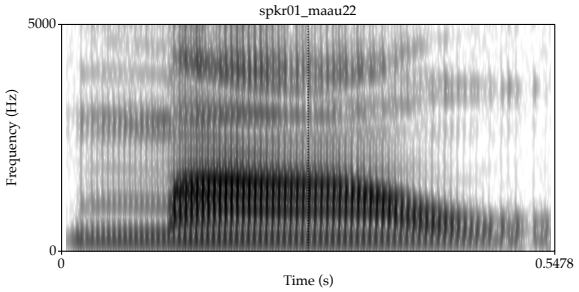
Appendix D

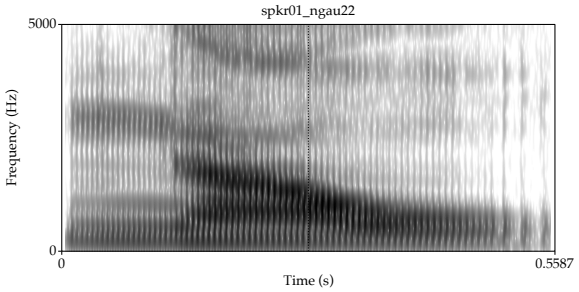
Example spectrograms from production data in Chapter 6

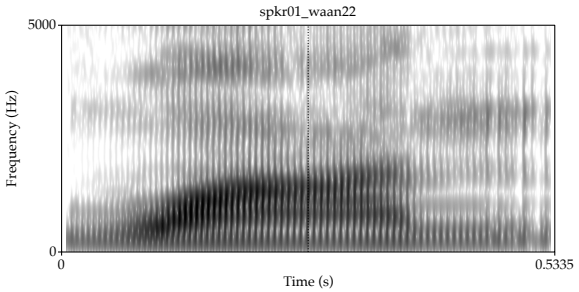
For transparency, example spectrograms of a range of the stimuli in Chapter 6 are reproduced here. It is thus possible to examine the realisation of the triphthong and some of the diphthongs, in particular in light of their variation in length (e.g. /maːu/ vs. /ŋɐu/) and composition (the closing diphthongs were true diphthongs, but the opening diphthongs instead consisted of a diphthong-like semi-vowel+open monophthong sequence). The figures below display for each speaker analysed one randomly selected repetition from each of tokens /maːu/, /ŋɐu/, /wɐn/, /waːn/ and /jɐu/, each on the low (22) level tone. Note that for the /ŋɐu/ token, initial /ŋ/ is undergoing a sound change in which it is increasingly pronounced as /ʔ/ (see Section 1.2.3). As a result, some speakers begin this token with a nasal, while others begin it with a stop.

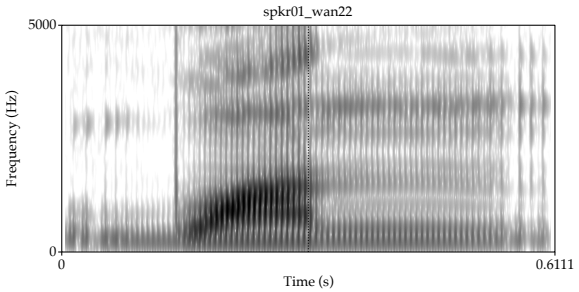
For each speaker, the first two spectrograms depict the long and short closing diphthongs, respectively, the next two rows the long and short opening diphthongs, respectively, and the last the triphthong. The speaker code and phonological target are printed at the top of each spectrogram for reference. Data from Speaker 4 were removed prior to analysis because the f0 could not be tracked reliably; thus, they are not depicted here.

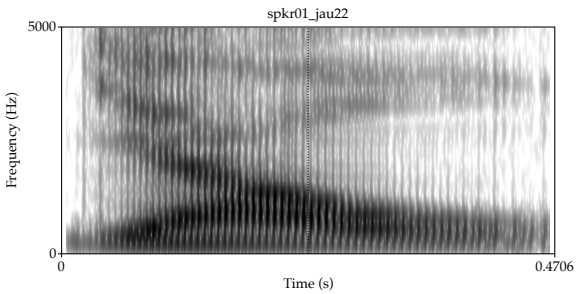
D.1 Speaker 1


spkr01_maau22
5000
Frequency (Hz)
0
0
0.5478
Time (s)

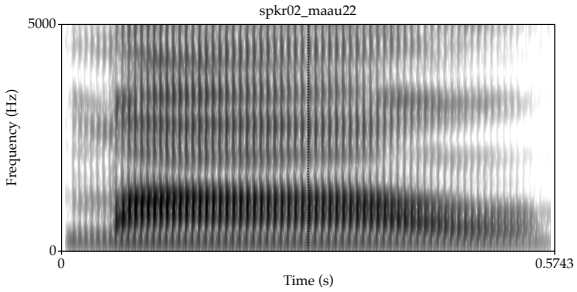

spkr01_ngau22
5000
Frequency (Hz)
0
0
0.5587
Time (s)

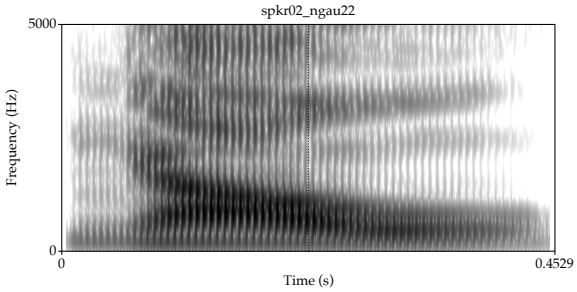

spkr01_waan22
5000
Frequency (Hz)
0
0
0.5335
Time (s)

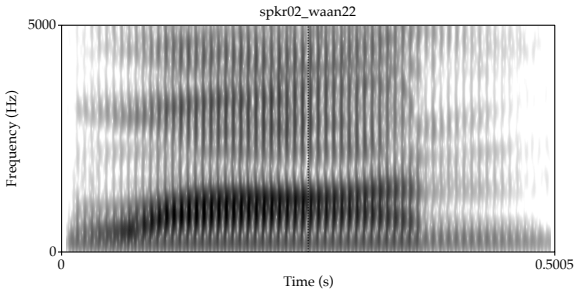

spkr01_wan22
5000
Frequency (Hz)
0
0
0.6111
Time (s)

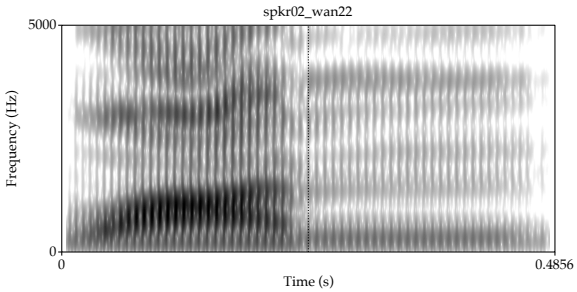

spkr01_jau22
5000
Frequency (Hz)
0
0
0.4706
Time (s)

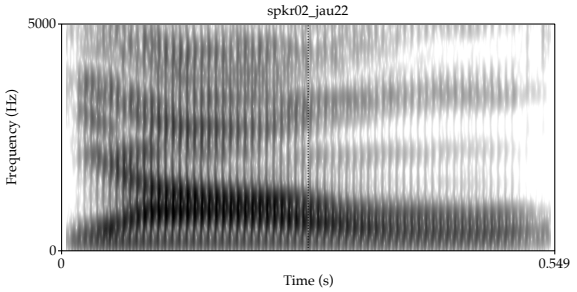
D.2 Speaker 2


spkr02_maau22
5000
Frequency (Hz)
0
0
0.5743
Time (s)

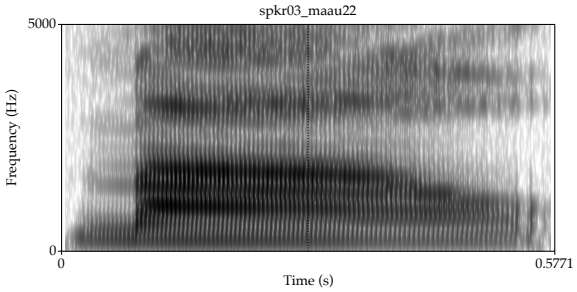

spkr02_ngau22
5000
Frequency (Hz)
0
0
0.4529
Time (s)

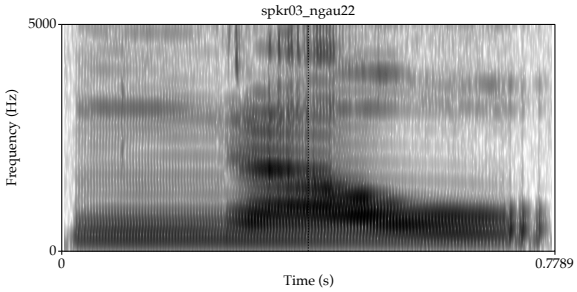

spkr02_waan22
5000
Frequency (Hz)
0
0
0.5005
Time (s)

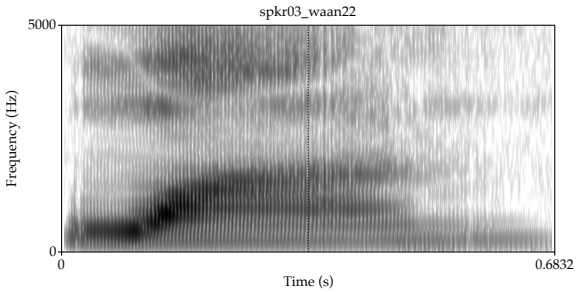

spkr02_wan22
5000
Frequency (Hz)
0
0
0.4856
Time (s)

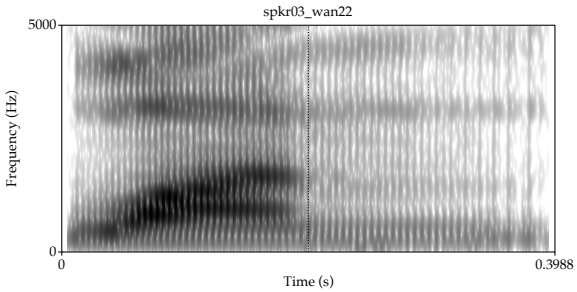

spkr02_jau22
5000
Frequency (Hz)
0
0
0.549
Time (s)

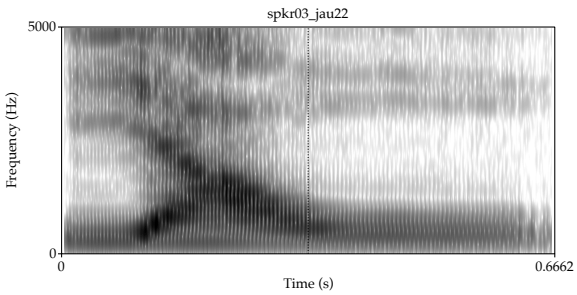
D.3 Speaker 3


spkr03_maau22
5000
Frequency (Hz)
0
0
0.5771
Time (s)

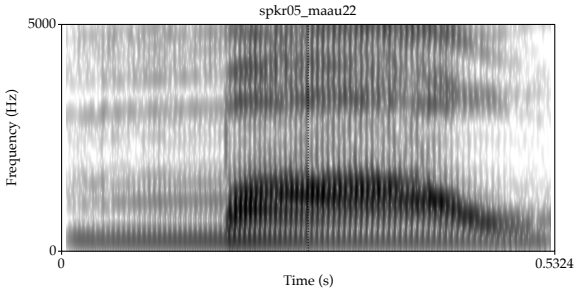

spkr03_ngau22
5000
Frequency (Hz)
0
0
0.7789
Time (s)

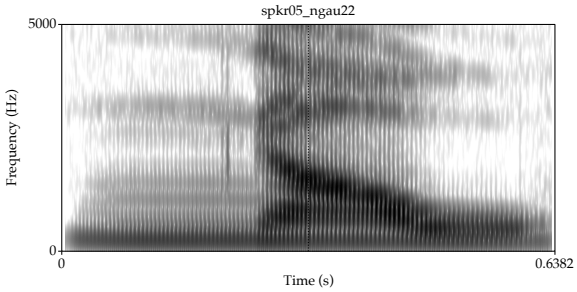

spkr03_waan22
5000
Frequency (Hz)
0
0
0.6832
Time (s)

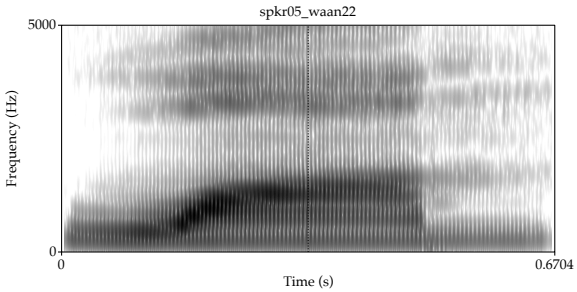

spkr03_wan22
5000
Frequency (Hz)
0
0
0.3988
Time (s)

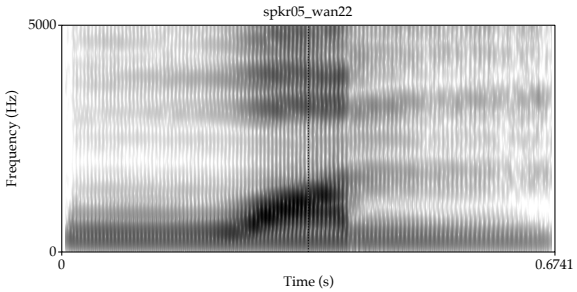

spkr03_jau22
5000
Frequency (Hz)
0
0
0.6662
Time (s)

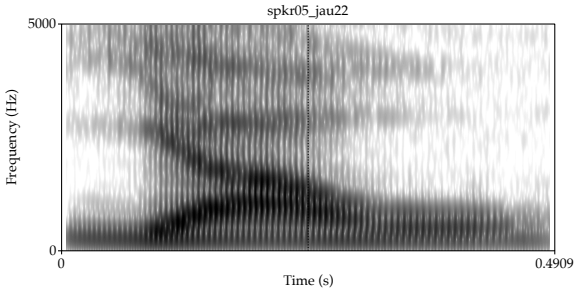
D.4 Speaker 5


spkr05_maau22
5000
Frequency (Hz)
0
0
0.5324
Time (s)

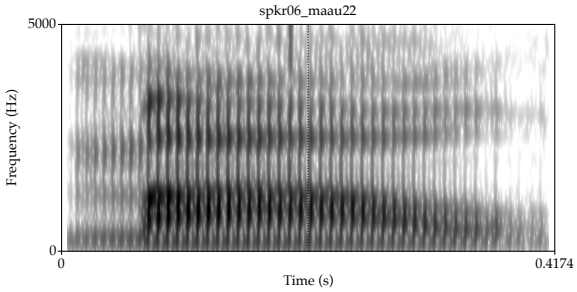

spkr05_ngau22
5000
Frequency (Hz)
0
0
0.6382
Time (s)

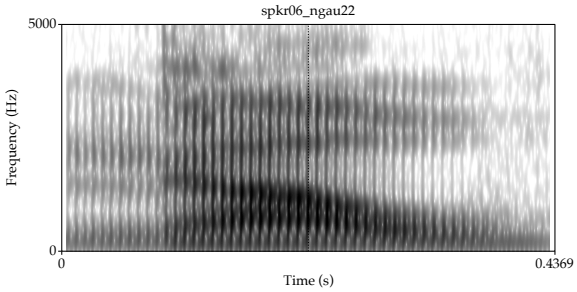

spkr05_waan22
5000
Frequency (Hz)
0
0
0.6704
Time (s)

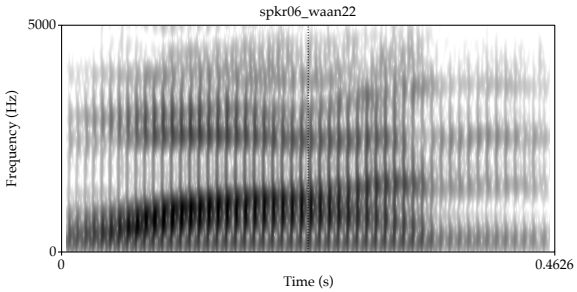

spkr05_wan22
5000
Frequency (Hz)
0
0
0.6741
Time (s)

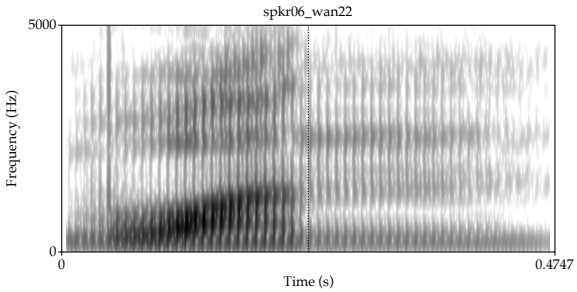

spkr05_jau22
5000
Frequency (Hz)
0
0
0.4909
Time (s)

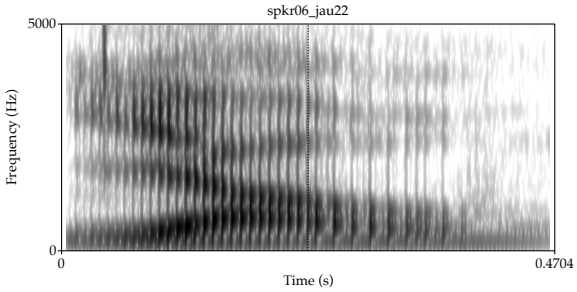
D.5 Speaker 6


spkr06_maau22
5000
Frequency (Hz)
0
0
0.4174
Time (s)

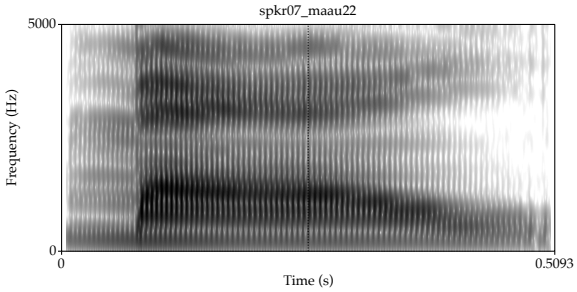

spkr06_ngau22
5000
Frequency (Hz)
0
0
0.4369
Time (s)

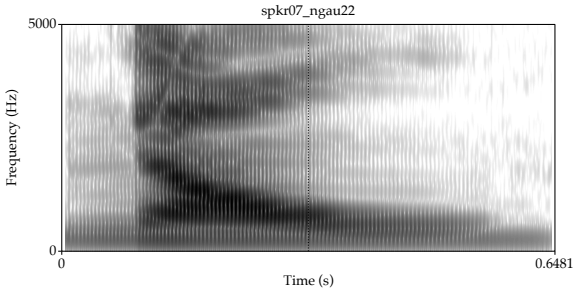

spkr06_waan22
5000
Frequency (Hz)
0
0
0.4626
Time (s)

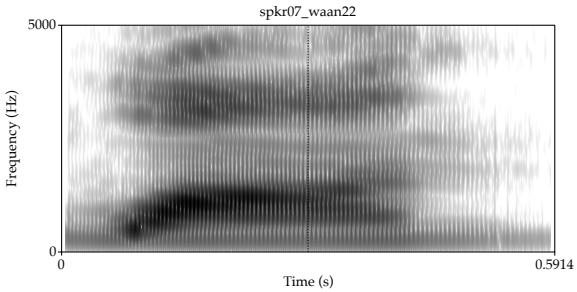

spkr06_wan22
5000
Frequency (Hz)
0
0
0.4747
Time (s)

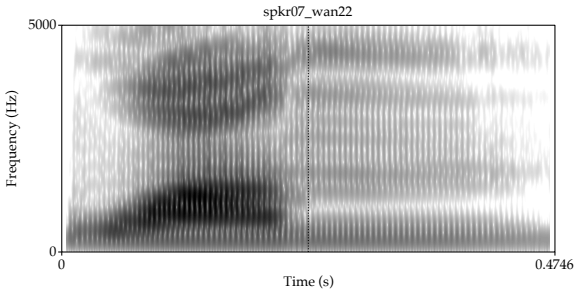

spkr06_jau22
5000
Frequency (Hz)
0
0
0.4704
Time (s)

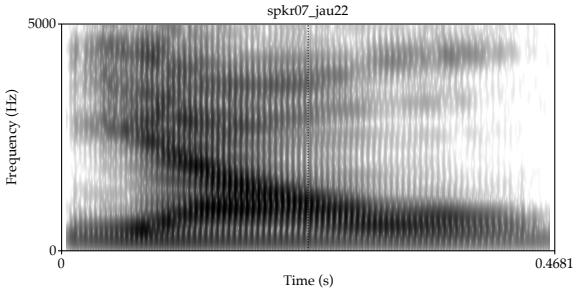
D.6 Speaker 7


spkr07_maau22
5000
Frequency (Hz)
0
0
0.5093
Time (s)

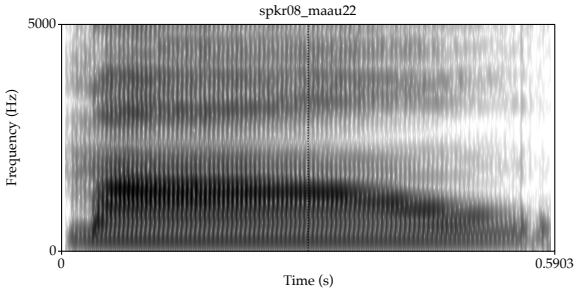

spkr07_ngau22
5000
Frequency (Hz)
0
0
0.6481
Time (s)

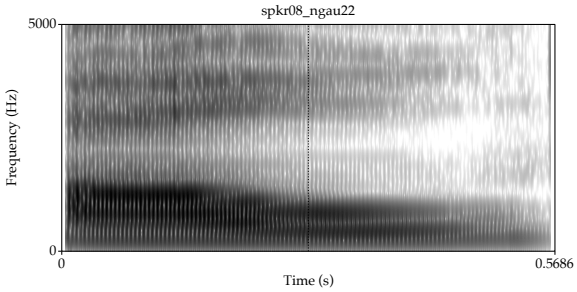

spkr07_waan22
5000
Frequency (Hz)
0
0
0.5914
Time (s)

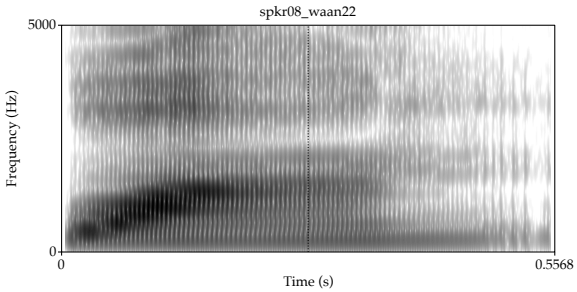

spkr07_wan22
5000
Frequency (Hz)
0
0
0.4746
Time (s)

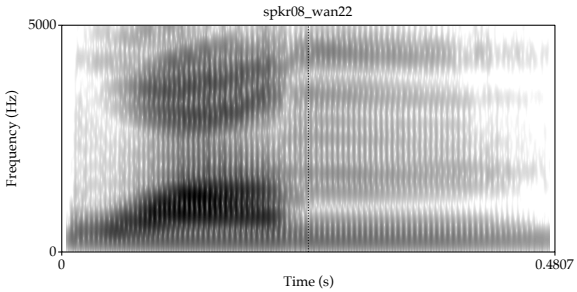

spkr07_jau22
5000
Frequency (Hz)
0
0
0.4681
Time (s)

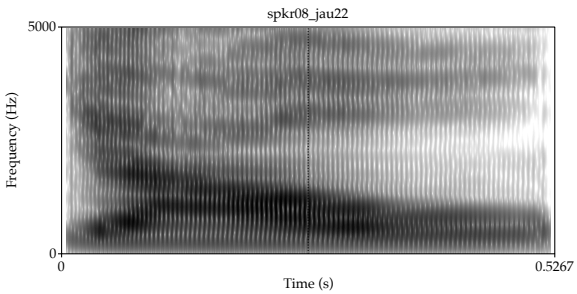
D.7 Speaker 8


spkr08_maau22
5000
Frequency (Hz)
0
0
0.5903
Time (s)

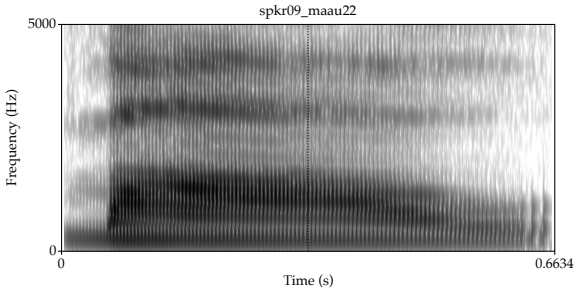

spkr08_ngau22
5000
Frequency (Hz)
0
0
0.5686
Time (s)

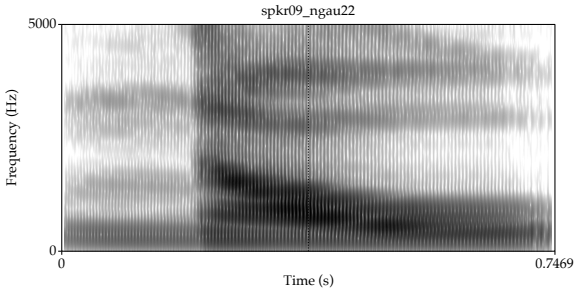

spkr08_waan22
5000
Frequency (Hz)
0
0
0.5568
Time (s)

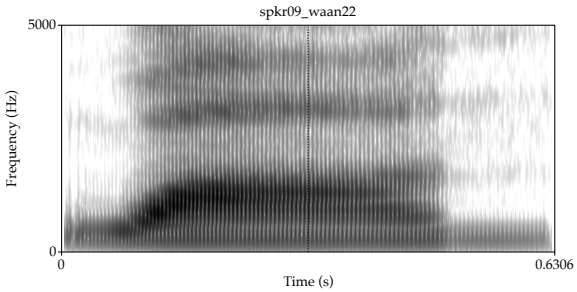

spkr08_wan22
5000
Frequency (Hz)
0
0
0.4807
Time (s)

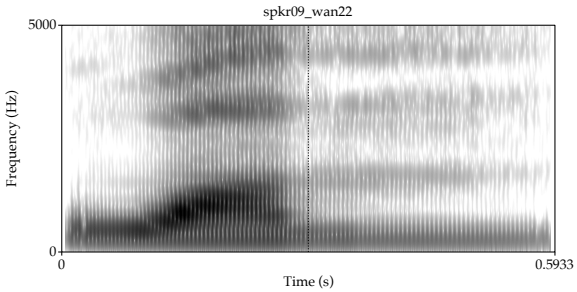

spkr08_jau22
5000
Frequency (Hz)
0
0
0.5267
Time (s)

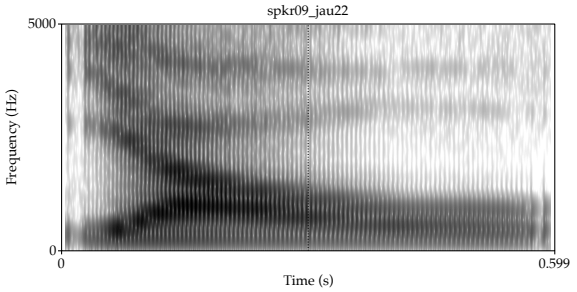
D.8 Speaker 9


spkr09_maau22
5000
Frequency (Hz)
0
0
0.6634
Time (s)

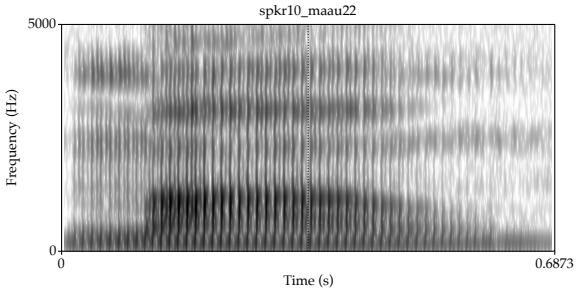

spkr09_ngau22
5000
Frequency (Hz)
0
0
0.7469
Time (s)

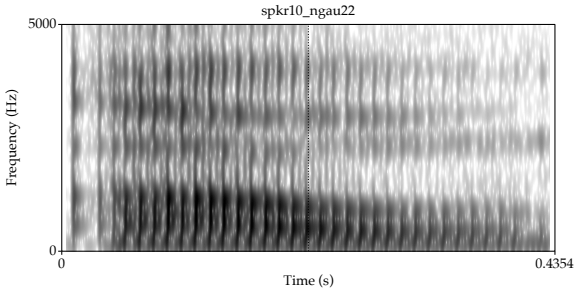

spkr09_waan22
5000
Frequency (Hz)
0
0
0.6306
Time (s)

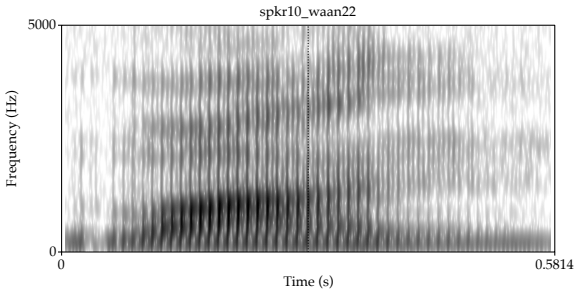

spkr09_wan22
5000
Frequency (Hz)
0
0
0.5933
Time (s)

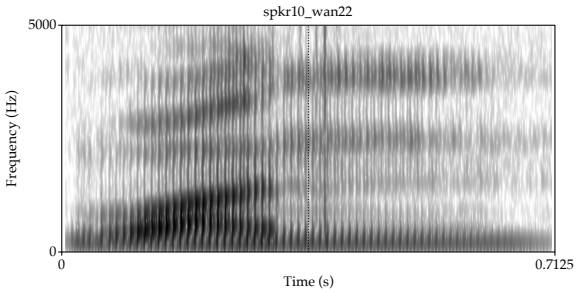

spkr09_jau22
5000
Frequency (Hz)
0
0
0.599
Time (s)

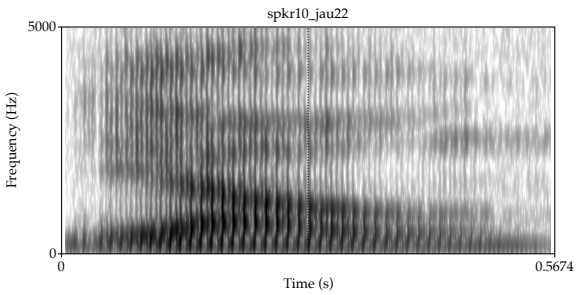
D.9 Speaker 10











Appendix E

Original f0 contours from production data in Chapter 6

Figures E.1 and E.2 show the unsmoothed f0 contours of the data presented in Section 6.3. That is, these are the speaker-normalised mean f0 values before DCT transformation. As such, Figure E.1 is the counterpart to Figure 6.1 in Section 6.3, and Figure E.2 the counterpart to Figure 6.4 in Section 6.3.

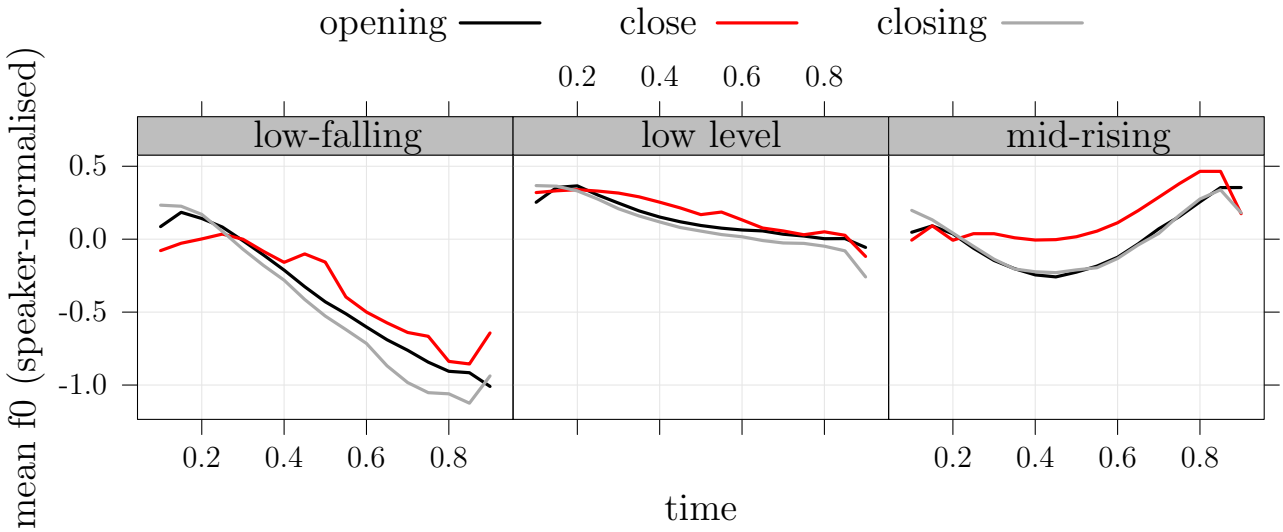


Figure E.1: Slopes analysis, equivalent to DCT-smoothed Figure 6.1 in Section 6.3. Vowel is plotted in colour, whereby red represents the close monophthong control.

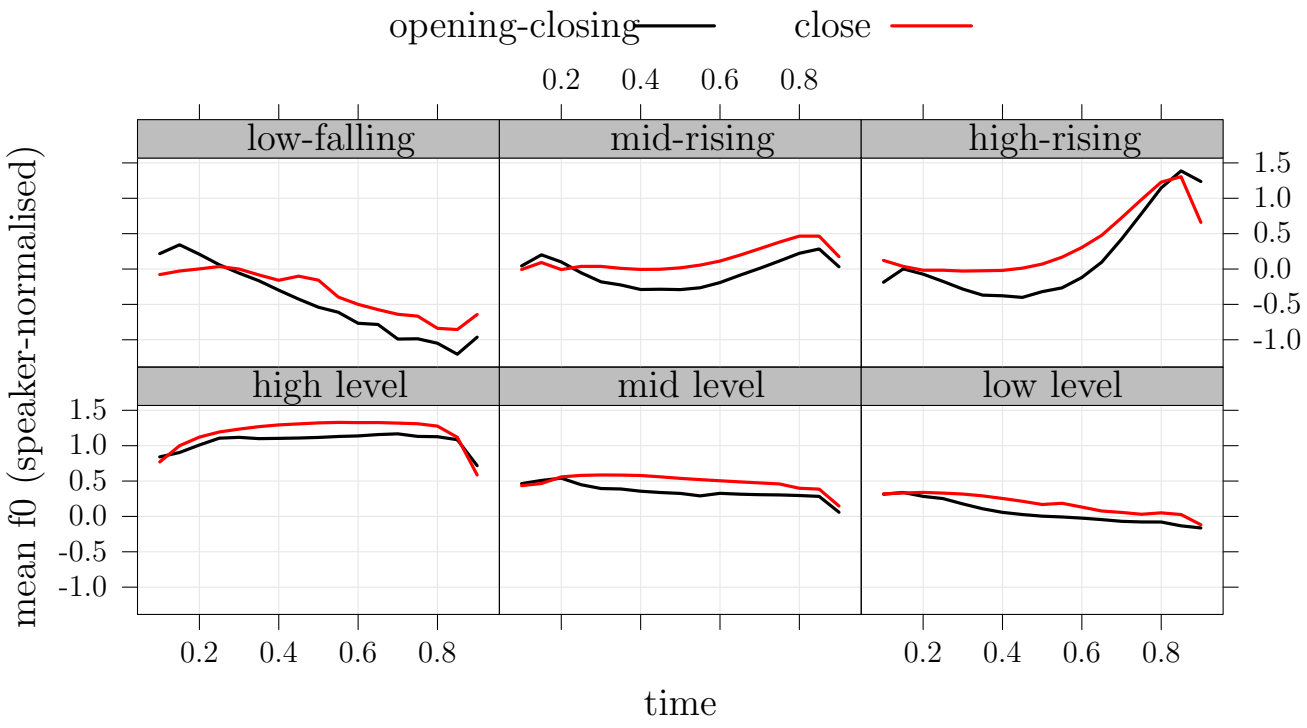


Figure E.2: Curvature analysis, equivalent to DCT-smoothed Figure 6.4 in Section 6.3. Vowel is plotted in colour, whereby red represents the close monophthong control.

Disclosure of pre-published data

For full disclosure, preliminary results from subsets of the data studied in Chapters 2, 3 and 5 were presented at the ICPhS 2015, AMLaP 2015 and LabPhon 2016 conferences, respectively (Siddins & Harrington, 2015a, 2015b; Siddins & Reinisch, 2016).

References

Atkinson, J. E. (1973). Intrinsic f0 in vowels: Physiological correlates. *The Journal of the Acoustical Society of America*, *53*(1), 346-346.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48.

Bauer, R. S. (1984). The Hong Kong Cantonese speech community. *Cahiers de linguistique - Asie orientale*, *13*(1), 57-90.

Bauer, R. S. (1988). Written Cantonese of Hong Kong. *Cahiers de linguistique - Asie orientale*, *17*(2), 245-293.

Bauer, R. S. (2000). Hong Kong Cantonese and the road ahead. In D. C. S. Li, A. Lin, & W. K. Tsang (Eds.), *Language and education in postcolonial Hong Kong* (p. 35-58). Linguistic Society of Hong Kong.

Bauer, R. S., & Benedict, P. K. (1997). *Modern Cantonese Phonology* (No. 102). Berlin: Mouton de Gruyter.

Bauer, R. S., Kwan-Hin, C., & Pak-Man, C. (2003). Variation and merger of the rising tones in Hong Kong Cantonese. *Language Variation and Change*, *15*, 211-225.

Beddor, P. S. (2009). A coarticulatory path to sound change. *Language*, *85*(4), 785-821.

Beddor, P. S. (2015). The relation between language users' perception and production repertoires. In *Proceedings of ICPhS.*

Beddor, P. S., Harnsberger, J. D., & Lindemann, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates. *Journal of Phonetics*, *30*(4), 591–627.

Boersma, P., & Weenink, D. (2012). Praat: doing phonetics by computer (Version 5.3.04 ed.) [Computer software manual]. Retrieved 12. Januar 2012, from `www.praat.org`

Brunelle, M., Lim, J., & Chow, D. (2010, September). Tone identification and confusion in Cantonese. In *Conference on the typology of tone and intonation (TIE4).* Stockholm.

Campbell, L. (2013). *Historical linguistics: An introduction* (3rd ed.). Malta: Edinburgh University Press.

Chinese Language Education Section, H. S. (2008). *Hong Kong Chinese lexical lists for primary learning.* Retrieved 30.09.2016, from `http://www.edbchinese.hk/lexlist_en/index.htm`

Coetzee, A., Beddor, P. S., & Wissing, D. (2014). The emergence of tonogenesis in Afrikaans. *Journal of the Acoustical Society of America*, *135*, 2421.

Connell, B. (2002). Tone languages and the universality of intrinsic F0: evidence from Africa. *Journal of Phonetics*, *30*, 101-129.

de Cheveigne, A., & Kawahara, H. (2001). Comparative evaluation of f0 estimation algorithms. In *Proceedings of Eurospeech 2001.*

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology*, *1*(2), 121-144.

Dong, H. (2014). *A history of the Chinese language.* Routledge.

Draxler, C., & Jänsch, K. (2004). SpeechRecorder - a Universal Platform Independent Multi-Channel Audio Recording Software. In *Proc. of the IV. International Conference on Language Resources and Evaluation* (p. 559-562). Lisbon, Portugal.

Duanmu, S. (1990). *A formal study of syllable, tone, stress and domain in Chinese languages* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R.* Great Britain: SAGE Publications Ltd.

Flanagan, J. L. (1965). Recent studies in speech research at Bell Telephone Laboratories. In *Proceedings of the 5th International Congress on Acoustics.* Liege.

Fok Chan, Y.-Y. (1974). *A perceptual study of tones in Cantonese.* Hong Kong: Centre for Asian Studies, University of Hong Kong.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*(14), 3-18.

Fowler, C. A. (2005). Parsing coarticulated speech in perception: effects of coarticulation resistance. *Journal of Phonetics*, *33*(2), 199-213.

Fowler, C. A., & Brown, J. M. (1997). Intrinsic f0 differences in spoken and sung vowels and their perception by listeners. *Perception & Psychophysics*, *59*(5), 729-738.

Francis, A. L., Ciocca, V., & Ng, B. K. C. (2003, October). On the (non)categorical perception of lexical tones. *Perception & Psychophysics*, *65*(7), 1029-1044.

Francis, A. L., Ciocca, V., Wong, V. K. M., & Chan, J. K. L. (2006, November). Is fundamental frequency a cue to aspiration in initial stops? *Journal of the Acoustical Society of America*, *120*(5), 2884-2895.

Fung, R., & Wong, C. S. P. (2011). Acoustic analysis of the new rising tone in Hong Kong Cantonese. In *Proveedings of ICPhS XVII.* Hong Kong.

Gandour, J. (1979). Perceptual dimensions of Cantonese tones: a multidimensional scaling reanalysis of Fok's tone confusion data. *South-east Asian Linguistic Studies*, *4*, 415-429.

Gay, T. (1968). Effect of speaking rate on diphthong formant movements. *The Journal of the Acoustical Society of America*, *44*(6), 1570-1573.

Gelfand, S. A. (1998). *Hearing: an introduction to psychological and physiological acoustics* (3rd ed.). New York, USA: Marcel Dekker Inc.

Glidden, C. M., & Assmann, P. F. (2004). Effects of visual gender and frequency shifts on vowel category judgments. *Acoustics Research Letters Online*, *5*(4), 132-138.

Gu, W., Hirose, K., & Fujisaki, H. (2004). Analysis and synthesis of Cantonese f0 contours based on the command-response model. In *2004 International Symposium on Chinese Spoken Language Processing.*

Harrington, J. (2010). *Phonetic Analysis of Speech Corpora*. Wiley Publishing.

Harrington, J., Kleber, F., & Reubold, U. (2008). Compensation for coarticulation, /u/-fronting, and sound change in Standard Southern British: an acoustic and perceptual study. *Journal of the Acoustical Society of America*, *123*, 2825-2835.

Hombert, J.-M. (1975). Phonetic motivations for the development of tones from postvocalic [h] and [ʔ] : Evidence from contour tone perception. *Proceedings of ICPhS 1975*.

Hombert, J.-M. (1977a). Consonant types, vowel height and tone in Yoruba. In *Studies in African linguistics* (Vol. 8, p. 173-190).

Hombert, J.-M. (1977b). Development of tones from vowel height? *Journal of Phonetics*, *5*, 9-16.

Hombert, J.-M., Ohala, J. J., & Ewan, W. G. (1979). Phonetical explanations for the develoment of tones. *Language*, *55*(1).

Honda, K. (1983). Relationship between pitch control and vowel articulation. In D. M. Bless & J. H. Abbs (Eds.), *Vocal fold physiology* (p. 286-297). San Diego: College-Hill Press.

Honda, K. (2004). Physiological factors causing tonal characteristics of speech: from global to local prosody. In *Proceedings of Speech Prosody* (p. 739-744).

Hoole, P. (2006). *Experimental studies of laryngeal articulation.* (Unpublished habilitation thesis, Ludwig-Maximilians-Universität Munich)

Hoole, P., & Honda, K. (2011). Automaticity vs. feature-enhancement in the control of segmental f0. In G. N. Clements & R. Ridouane (Eds.), *Where do phonological features come from?* (p. 131-171). Benjamins Publishing Company.

Hsieh, F.-f. (2005). Tonal chain-shifts as anti-neutralization-induced tone sandhi. *University of Pennsylvania Working Papers in Linguistics*, *11*(1), 9.

Hu, F. (2013). Falling diphthongs have a dynamic target while rising diphthongs have two targets: Acoustics and articulation of the diphthong production in Ningbo Chinese. *The Journal of the Acoustical Society of America*, *134*(5), 4199-4199.

Jacewicz, E., & Fox, R. A. (2015). Intrinsic fundamental frequency of vowels is moderated by regional dialect. *The Journal of the Acoustical Society of America*, *138*(4), EL405-EL410.

Jun, S.-A. (1996). Influence of microprosody on macroprosody: a case of phrase initial strengthening. *UCLA Working Papers in Phonetics*, 97–116.

Kao, D. L. (1971). *Structure of the syllable in Cantonese* (No. 78). The Hague: Mouton & Co N. V.

Kataoka, R. (2011). *Phonetic and cognitive bases of sound change* (Unpublished doctoral dissertation). UC Berkeley.

Keating, P., Garellek, M., & Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice. In *Proceedings of the 18th International Congress of Phonetic Sciences.* Glasgow, Scotland.

Khouw, E., & Ciocca, V. (2007). Perceptual correlates of Cantonese tones. *Journal of Phonetics*, *35*(1), 104-117.

Kingston, J. (1992). The phonetics and phonology of perceptually motivated articulatory covariation. *Language and Speech*, *35*(1-2), 99-113.

Kingston, J. (2007). Segmental influences on f0: automatic or controlled? In C. Gussenhoven & T. Riad (Eds.), *Tones and tunes.* (Vol. 2: Experimental studies in word and sentence prosody, p. 171-210). Berlin: Mouton de Gruyter.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, *70*(3), pp. 419-454.

Kingston, J., Kawahara, S., Chambless, D., Key, M., Mash, D., & Watsky, S. (2014). Context effects as auditory contrast. *Attention, Perception, & Psychophysics*, *76*(5), 1437–1464.

Kisler, T., Schiel, F., & Sloetjes, H. (2012). Signal processing via web services: the use case WebMAUS. In *Digital humanities* (p. 30-34). Hamburg, Germany.

Klatt, D. H. (1973). Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. *The Journal of the Acoustical Society of America*, *53*(1), 8-16.

Kuang, J., & Cui, A. (2016). Relative cue weighting in perception and production of a sound change in progress. In *Proceedings of the 15th Conference on Laboratory Phonology.* Ithaca, NY.

Ladefoged, P. (1964). *A phonetic study of West African languages.* Cambridge University Press.

Lehiste, I. (1970). *Suprasegmentals.* Cambridge, Massachusetts: The MIT Press.

Leung, M.-T., & Law, S.-P. (2001). HKCAC: The Hong Kong Cantonese adult language corpus. *International Journal of Corpus Linguistics*, *6*, 305-326.

Leung, M.-T., Law, S.-P., & Fung, S.-Y. (2004). Type and token frequencies of phonological units in Hong Kong Cantonese. *Behavior Research Methods, Instruments & Computers*, *36*(3), 500-505.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1-36.

Liberman, M., & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff & R. T. Oehrle (Eds.), *Language, sound, structure: Studies in phonology* (p. 157-233). Cambridge, Ma: MIT Press.

Lieberman, P. (1970). *A study of prosodic features* (Tech. Rep.). Haskins Laboratories Status Reports on Speech Research, SR-23, 179-208.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 839–862.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, *35*(11), 1773-1781.

Lindblom, B. (1967). Vowel duration and a model of lip mandible coordination. *Speech Transmission Laboratory Quarterly Progress Status Report*, *4*, 1-29.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modeling* (p. 403-439). Dordrecht: Kluwer.

Lindblom, B., Guion, S., Hura, S., Moon, S.-J., & Willerman, R. (1995). Is sound change adaptive? *Rivista Di Linguistica*, *7*(1), 5-37.

Lindblom, B., Lubker, J., Gay, T., Lyberg, B., Branderud, P., & Holmgren, K. (1987). The concept of target and speech timing. In R. Channon & L. Shockey (Eds.), *In*

honnor of Ilse Lehiste (p. 161-182). De Gruyter.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (coturnix coturnix japonica). *The Journal of the Acoustical Society of America*, *102*(2), 1134–1140.

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). USA: Massachusetts Institute of Technology.

Maddieson, I. (1976). The intrinsic pitch of vowels and tones in Foochow. In *UCLA working papers in phonetics* (p. 191-202).

Maddieson, I. (1978). Universals of tone. In C. A. Ferguson & E. A. Moravcsik (Eds.), *Universals of human language* (Vol. 2: Phonology). Stanford, California: Stanford University Press.

Maddieson, I. (2013). Tone. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `http://wals.info/chapter/13`

Mann, V. A., & Repp, B. H. (1980, September). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception & Psychophysics*, *28*(3), 213-228.

Matisoff, J. A. (1973). Tonogenesis in Southeast Asia. In L. M. Hyman (Ed.), *Consonant types and tones.* Southern California Occasional Papers in Linguistics.

Matthews, S., & Yip, V. (2013). *Cantonese: A comprehensive grammar.* Routledge.

Menzerath, P., & de Lacerda, A. (1933). *Koartikulation, Steuerung und Lautabgrenzung: Eine experimentelle Untersuchung.* F. Dümmler.

Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition*, *109*(1), 168–173.

Mok, P., Zuo, D., & Wong, P. W. Y. (2013, 10). Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language Variation and Change*, *25*, 341-370.

Niebuhr, O. (2004, 23-26 March). Intrinsic pitch in opening and closing diphthongs of German. In *Proceedings of Speech Prosody.* Nara, Japan.

Nitta, T. (2001). The accent systems in the Kanazawa dialect: the relationship between pitch and sound segments. In S. Kaji (Ed.), *Proceedings of the symposium cross-linguistic studies of tonal phenomena: tonogenesis, Japanese accentology, and other topics: held Tokyo 12-14 December 2000* (p. 153-185). Tokyo: Institute for the study of languages and cultures of Asia and Africa (ILCAA), Tokyo University.

Odden, D. (2010). Features impinging on tone. In J. A. Goldsmith, E. Hume, & W. L. Wetzels (Eds.), *Tones and features: Phonetic and phonological perspectives* (p. 81-107). Berlin: De Gruyter.

Ohala, J. J. (1981). The listener as a source of sound change. In C. S. Masek, R. A. Hendrick, & M. F. Miller (Eds.), *Papers from the parasession on language and behavior* (p. 178-203). Chicago: Chicago Linguistic Society.

Ohala, J. J. (1993). The phonetics of sound change. In C. Jones (Ed.), *Historical Linguistics: Problems and Perspectives* (p. 237-278). London: Longman.

Ou, J. (2012). *Tone merger in Guangzhou Cantonese* (M.Phil). The Hong Kong Polytechnic University.

Panfili, L. (2016). The physiological underpinnings of vowel height and voice quality. *The Journal of the Acoustical Society of America*, *139*(4), 2221-2221.

Pape, D. (2008). The native language influence on perceptual intrinsic pitch: Crosslinguistic data from German, Italian, Portuguese, and Spanish. In *Proceedings of Speech Prosody.*

Pape, D. (2009). *Microprosodic differences in a cross-linguistic vowel comparison of speech production and speech perception.* Weißensee-Verlag.

Pape, D., & Mooshammer, C. (2008, 30 June - 2 July). Intrinsic pitch is not a universal phenomenon: evidence from Romance languages. In *Proceedings of the 11th Conference on Laboratory Phonology* (p. 109-110). Wellington, New Zealand.

Pape, D., Mooshammer, C., Fuchs, S., & Hoole, P. (2005). Intrinsic pitch differences between german vowels /i:/, /i/ and /y:/ in a cross-linguistic perception experiment. In *Proceedings of the isca workshop on plasticity in speech perception (psp2005).* London, UK.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8-13.

Pilszczikowa-Chodak, N. (1972). Tone-vowel height correlation and tone assignment in the patterns of verb and noun plurals in Hausa. *Studies in African Linguistics*, *3*, 399-422.

R Development Core Team. (2011). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/` (ISBN 3-900051-07-0)

Reinholt Petersen, N. (1986). Perceptual compensation for segmentally conditioned fundamental frequency perturbation. *Phonetica*, *43*, 31-42.

Sagart, L., Hallé, P., de Boysson-Bardies, B., & Arabia-Guidet, C. (1986). Tone production in modern standard Chinese: An electromyographic investigation. *Cahiers de linguistique-Asie orientale*, *15*(2), 205-221.

Schiel, F. (1999, August). Automatic Phonetic Transcription of Non-Prompted Speech. In *Proc. of the icphs* (p. 607-610). San Francisco.

Siddins, J., & Harrington, J. (2015a, 10-14 August). Does vowel intrinsic f0 affect lexical tone? In *Proceedings of the 18th International Congress of Phonetic Sciences.* Glasgow.

Siddins, J., & Harrington, J. (2015b). Vowel height affects the perception of lexical tone. In A. Gatt & H. Mitterer (Eds.), *Architectures & mechanisms for language processing* (p. 22). Valletta, Malta.

Siddins, J., & Reinisch, E. (2016). Intrinsic pitch of diphthongs in lexical tone perception. In *Proceedings of LabPhon 15.* Ithaca, NY.

Silverman, K. E. A. (1987). *The structure and processing of fundamental frequency contours* (Unpublished doctoral dissertation). University of Cambridge.

Snow, D. (2004). *Cantonese as written language: The growth of a written Chinese vernacular.* Hong Kong University Press.

Stoll, G. (1984). Pitch of vowels: Experimental and theoretical investigation of its dependence on vowel quality. *Speech Communication*, *3*(2), 137-147.

Terhardt, E. (1972a). Zur Tonhöhenwahrnehmung von Klängen: Ii: Ein Funktionsschema. *Acustica*, *26*, 187-199.

Terhardt, E. (1972b). Zur Tonhöhenwahrnehmung von Klängen: I: Psychoakustische Grundlagen. *Acustica*, *26*, 173-186.

Terhardt, E. (1974). Pitch, consonance, and harmony. *The Journal of the Acoustical Society of America*, *55*(5), 1061-1069.

Van Hoof, S., & Verhoeven, J. (2011). Intrinsic vowel f0, the size of vowel inventories and second language acquisition. *Journal of Phonetics*, *39*(2), 168-177.

Vance, T. J. (1977). Tonal distinctions in Cantonese. *Phonetica*, *34*(2), 93-107.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (Fourth ed.). New York: Springer.

Watson, C. I., & Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *Journal of the Acoustical Society of America*, *106*(1), 458-468.

Whalen, D., Gick, B., & LeSourd, P. S. (1999). Intrinsic f0 in Passamaquoddy vowels. In *Papers of the algonquian conference* (Vol. 30, p. 417).

Whalen, D., & Levitt, A. G. (1995). The universality of intrinsic f0 of vowels. *Journal of Phonetics*, *23*, 349-366.

Whalen, D., Levitt, A. G., Hsiao, P.-L., & Smorodinsky, I. (1994). Intrinsic f0 of vowels in the babbling of 6-, 9- and 12-month-old French- and English-learning infants. *Haskins Laboratories Status Report on Speech Research*, *SR-117/118*, 15-24.

Winn, M. B., Chatterjee, M., & Idsardi, W. J. (2013). Roles of voice onset time and f0 in stop consonant voicing perception: Effects of masking noise and low-pass filtering. *Journal of Speech, Language & Hearing Research*, *56*(4), 1097-1107.

Winter, B. (2013). *Linear models and linear mixed effects models in R with linguistic applications.* Retrieved 07.09.2016, from `http://arxiv.org/pdf/1308.5499.pdf`

Wong, P. C. M., & Diehl, R. L. (1998). Effects of speaking fundamental frequency on the normalization of Cantonese level tones. *The Journal of the Acoustical Society of America*, *104*(3), 1834-1834.

Wong, P. C. M., & Diehl, R. L. (2003). Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *Journal of Speech, Language, and Hearing Research*, *46*, 413–421.

Yu, K. M., & Lam, H. W. (2014). The role of creaky voice in Cantonese tonal perception. *The Journal of the Acoustical Society of America*, *136*(3), 1320-1333.

Yue Hashimoto, O.-K. (1972). *Studies in Yüe dialects 1: Phonology of Cantonese* (Vol. III). Cambridge University Press.

Zee, E. (1999). Chinese (Hong Kong Cantonese). In *Handbook of the International Phonetic Association* (p. 58-60). Cambridge: Cambridge University Press.

Zhao, Y., & Jurafsky, D. (2009). The effect of lexical frequency and lombard reflex on tone hyperarticulation. *Journal of Phonetics*, *37*(2), 231-247.

Zheng, Q. (2014). Effects of vowels on Mandarin tone categorical perception. *Acta Psychologica Sinica*, *46*(9), 1223-1231.

Acknowledgements

Many people made this project possible for me.

In February 2008, I came to the LMU's Open Day in search of anything and everything related to speech and language. It was there that I first stumbled upon phonetics, via Uwe Reichel, who was manning the IPS stand. The subject appealed to me because it looked scientific and measurable, and because Uwe was like an island of calm amongst the other stands aggressively advertising their fields. Soon after, he became one of my favourite lecturers, a valuable colleague and a good friend.

More than anyone, I owe my heartfelt thanks to my supervisor, Jonathan Harrington, for offering me this opportunity. He has created a truly unique working environment and a wonderful team of scientists here at the IPS, and it has been a great honour to learn the tools of the trade in this setting. While it wasn't always easy, I greatly appreciated Jonathan's questioning, criticism and of course encouragement of my work, and words cannot describe how grateful I am for his compassion and understanding when times were tough.

Phil Hoole has been a constant source of encyclopaedic knowledge and inspiration as well as tips and tricks relevant to the topic of my thesis. His door was always open (even when he was nowhere to be found!).

It has been a particularly special honour to work with Emeritus Professor Hans "Tim" Tillmann. Many, many hours have been filled with Tim's hugely entertaining stories of phoneticians past and present, and I learnt things about the history of our subject that I would never have known otherwise. The IPS would not be the same without Tim's grandfatherly presence and sunny demeanour.

I am hugely indebted to Siyi Li as my primary informant and "little helper" with all things Cantonese. Without Siyi, I would never have been able to come up with appropriate stimuli, and recruiting participants would have been much more challenging. I was very sad to see Siyi go but wish her all the best for her future in China. Thanks also go to Ventus Siu for recruitment and testing of the participants of the perception experiment reported in Chapter 5 and to Eva Reinisch for getting the ball rolling on this same experiment.

Florian Schiel was always there to lend a hand with advanced MAuS parameters and talk about good methodological practice. I appreciate the time he took for these discussions and truly admire Florian's calm, thorough and logical manner of problem-solving. Lia Saki Bucar-Shigemori, Felicitas Kleber, Eva Reinisch, Ulrich Reubold and Mary Stevens all deserve a shout-out for "all-round awesomeness". They acted as bouncing boards for

ideas, patient receivers of stupid statistics questions, and tissue holders in moments of despair. In addition, Lia and Eva very kindly provided comments on early versions of some experimental chapters. Klaus Jänsch is the Greatest System Administration and Technical Assistant Ever, without whom so much at the IPS would not be possible. Similarly, thank you to our wonderful secretary, Ulrike Vallender-Kalus, for providing help above and beyond what is asked of her with all things bureaucratic and complicated and annoying!

I am constantly awed by the friendliness and helpfulness of the international phonetics community. Four people stand out in this regard. James Kirby (University of Edinburgh) shared his knowledge of tone and phonation as well as a Cantonese corpus, John Kingston (University of Massachusetts) scanned in and emailed an old, unpublished doctoral dissertation I had given up trying to find, and Carmen Kung (Macquarie University) double-checked stimuli Siyi and I had selected for suitability in Experiment 2 (Chapter 3). Colin Wilson (John Hopkins University), who I had never even met, very generously shared tips, suggestions and even code to help me with statistical procedures I was stuck with (and which unfortunately didn't end up making it into this dissertation, but they will be used eventually!).

On a more personal level, my eternal love and gratitude go to the following people: my grandfather, Des Stevenson, for encouraging me very early on in life to go my own way; my father, Tom Siddins, for his patience, acceptance, and constant faith in me, even when I don't have it myself; and finally, to my family here in Munich: Raphael, Brenda and Anton Winkelmann (and Jacky!). I don't even know what to say! Thank you for taking me in so warmly, looking after me in times of need, and being there every moment of every day. I am so lucky to have you!