The dynamics of Japanese prosody

Kwing Lok Albert Lee

A dissertation submitted in
fulfilment of the requirements for the degree of
Doctor of Philosophy

to

Department of Speech, Hearing and Phonetic Sciences
Division of Psychology and Language Sciences
University College London (UCL)

2015

DECLARATION

I, Kwing Lok Albert Lee, confirm that the work presented in this dissertation is my own. Where information has been derived from other sources, I confirm that this has been indicated in the dissertation.

Kwing Lok Albert Lee

ABSTRACT

This dissertation explores aspects of Tokyo Japanese (Japanese henceforth) prosody through acoustic analysis and analysis-by-synthesis. It 1) revisits existing issues in Japanese prosody with the minimal use of abstract notions and 2) tests if the Parallel Encoding and Target Approximation (Xu, 2005) framework is suitable for Japanese, a pitch accent language.

The first part of the dissertation considers the nature of lexical pitch accent through examining factors that affect the surface F0 realisation of an accent peak (Chapter 2) and establishing the articulatory domain that hosts a tonal target in Japanese (Chapter 3). Next, pitch accent interactions with other communicative functions are considered, specifically in terms of focus (Chapter 4) and sentence type (Chapter 5). Hypotheses using acoustic analyses from the previous Chapters are then verified through analysis-by-synthesis with articulatory synthesisers AMtrainer, PENTAtrainer1, and PENTAtrainer2 (Chapter 6).

Chapter 2 provides conclusive evidence of Japanese as a two-tone language as opposed to bearing three underlying tones in its phonology, previously unresolved in existing literature. Proponents of the two-tone hypothesis gather evidence from perception: when stimuli are played in isolation, native listeners can only distinguish two tone levels (High and Low). On the other hand, production evidence reveals robustly three distinct surface F0 levels. Using a series of linear regression analyses, I show the third tone level could be interpreted as a result of pre-low raising, a common articulatory phenomenon. The F0 of an accent peak is inversely correlated with the F0 of the following low target, being an enhanced peak in preparation for the upcoming L. Interpreted together with native listeners' inability to hear three tones when said in isolation, as repeatedly reported in previous studies, I establish Japanese has only H and L in its tonal inventory.

Chapter 3 establishes the syllable as the tone-bearing unit in Japanese tonal articulation. Often described as a mora-timed language, it has been previously unclear whether articulatory tonal targets are hosted in a mora or a syllable in Japanese. When comparing accented words of various syllable structures I found that the F0 accent peak of CVCV words occurs consistently earlier than that of CVn/CVCV words. CVCV words are longer in total duration so its earlier F0 peak is a result of a shorter tone-bearing unit (i.e. two consecutive short morae/syllables). CVn/CVV words on the other hand have a later peak F0 due to hosting an articulatory target as a long syllable, rather than two short morae. I further verified the syllable hypothesis using two articulatory synthesisers, PENTAtrainer1 and PENTAtrainer2. The syllable as a tone-bearing unit incurs fewer predictors but provides better learning accuracy.

Chapter 4 explores focus prosody in declarative sentences. Using a newly collected corpus of 6251 sentences that controls for accent condition, focus condition, sentence type, and sentence length, I challenge the widely held idea that post-focus compression of F0 range is accent-independent. Currently it is generally accepted that regardless of the accent condition of the focused word, the excursion size of 'initial rise' that marks the beginning of the first word

after focus is shrunken. However, confining the notion of post-focus compression to initial-rise (usually extending across only two morae) sets Japanese apart from other languages like English or Mandarin, where such compression is robust across the entire post-focus domain. I show that when F0 range is measured across a wider domain, compression is absent. Where post-focus compression is absent, the F0 trajectory appears to be a result of articulatory carryover effects. This will be interpreted as a result of weak articulatory strength on the post focus domain, explaining the difference in F0 trajectories in long and short utterances.

Chapter 5 builds on the previous Chapter to consider in addition the focus prosody in yes/no questions. I investigate what marks a yes/no question, and how focus prosody differs in declarative and interrogative utterances. Acoustic analyses show that questions are marked by a final rise, but the exact shape of such a rise depends on the accent condition of the sentence-final word. When compared to declarative sentences, the key differences in yes/no questions include: a higher F0 level; the absence of post-focus compression even in contexts otherwise observed in statements; and on-focus F0 raising as the only robust focus marker. These findings point to the fact that interrogative focus prosody is not an amalgamation of focus markers and question markers, and bear implication on the representation of Japanese intonation.

Chapter 6 verifies observations established thus far through analysis-by-synthesis. I demonstrate comparative modeling as a means to adjudicate between competing theories using PENTAtrainer2, PENTAtrainer1 and AMtrainer. In terms of local fitting accuracy, AMtrainer yielded comparable synthesis accuracy to the PENTAtrainers. Finally, I further demonstrate the compatibility of PENTA with Japanese prosody showing highly accurate F0 predictive analysis (when trained with Chapter 2 production data), and highly satisfactory speaker-dependent synthesis accuracy (when trained with Chapter 4 and 5 sentential data). Naturalness judgment ratings show that the natural stimuli sound as natural as the synthetic stimuli, though questions generally sound less natural than statements. Reasons for this discrepancy are discussed with reference to the design of the stimuli.

ACKNOWLEDGEMENTS

This long journey is finally over! Along the way, I have been helped by lot of people, who I am going to thank below.

First and foremost, I would like to express my most sincere gratitude and appreciation to my principal supervisor, Yi Xu. Without any background at all, I was initially unsure of doing a PhD in phonetics, but Xulaoshi told me ‘don’t worry, we are going to do this together’. To help me catch up, he involved me in a number of side projects, from which I have learnt a lot. Needless to say, Xulaoshi's expertise in this area is godsent; I only hope I could have learnt more from him. On a personal level, I don’t think I would have been as comfortable with any other supervisor. He has also been very kind to allow me to practically do a third of my PhD from home; thanks to him I started my programme with troubling illness and now I am leaving without it. 许老师, 谢谢您!

I have benefitted tremendously from the regular meetings with Mark Huckvale, my secondary supervisor. Mark has provided many useful insights that have helped shape a lot of my arguments. I was also a teaching assistant in some of Mark’s courses, in which I gained invaluable teaching experience.

Bronwen Evans and Shinichiro Ishihara, my internal and external examiners, have provided extensive and constructive comments on this dissertation. My viva exam with them was a great learning experience, and I enjoyed every minute of it.

This journey has been great fun with my PhD buddies, including (Dr./Mr./Miss/Ms./Mrs.) Al-Sari, Al-Shangiti, Atria, Brint, Chiu, Figueroa, Granlund, Hackman, Hsu, Kennedy-Higgins, Liu, Pereira, Redey-Nagy, Schoof, Shinohara, Song, Steinmetzger, Stringer, Tome, and Zhang. Many of these people have been and will continue to be close friends, and I thank them all collectively.

Outside UCL, Mark Shek has been my life coach, medical advisor, and best friend in England since 2002. Ricky Wong is the closest thing to a brother, supplying lovely homemade dinners and patient ears. These two friends had persistently persuaded me to return to England to pursue a PhD — I would not be writing this dissertation but for them.

Big thanks go to Santitham Prom-on (KMUTT) who has offered extensive help on my F0 modelling work; to Candide Simard (SOAS) for hiring me as her research assistant, and exposing me to the area of endangered languages; and to Peggy Mok (CUHK) for giving me employment, networking opportunities, career advice, and encouragement.

The financial awards from UCL Overseas Research Scholarship and Japan Foundation Endowment Committee Research Grant are gratefully acknowledged.

Finally, I thank Debbie for all the loving support that has lasted me through this PhD, and mum and dad who have always encouraged me to pursue my dreams.

TABLE OF CONTENTS

LIST OF FIGURES **9**

LIST OF TABLES **12**

CHAPTER 1: INTRODUCTION **13**

1.1 BACKGROUND 13
1.2 AUTOSEGMENTAL-METRICAL THEORY 14
1.3 THE FUJISAKI MODEL 16
1.4 PENTA MODEL 18

CHAPTER 2: ACOUSTIC PROPERTIES OF LEXICAL ACCENT **22**

2.1 BACKGROUND 22
2.1.1 The phonology of Japanese pitch accent *22*
2.1.2 F0 scaling *23*
2.1.3 F0 alignment *25*
2.1.4 Final accented vs. unaccented words *26*
2.1.5 Other acoustic cues *27*
2.2 METHODOLOGY 28
2.2.1 The stimuli *28*
2.2.2 Recording procedures *30*
2.2.3 Annotation *31*
2.3 GENERAL ACOUSTIC ANALYSIS 32
2.3.1 F0 scaling *32*
2.3.2 Peak alignment *36*
2.3.3 Formant analysis *39*
2.3.4 Discussion *42*
2.4 PITCH ACCENT AS PRE-LOW RAISING 43
2.4.1 Background *43*
2.4.2 Methodology *46*
2.4.3 Results *46*
2.4.4 Discussion *51*
2.5 CHAPTER CONCLUSION 54

CHAPTER 3: SYLLABLE AS THE TONE-BEARING UNIT FOR JAPANESE **55**

3.1 INTRODUCTION 55
3.2 METHODOLOGY 57
3.3 RESULTS AND DISCUSSION 58
3.3.1 Evidence 1: F0 peak timing *58*
3.3.2 Evidence 2: Global articulatory targets from PENTAtrainer2 ..... *60*
3.3.3 Evidence 3: Learning accuracy with restricted target search *61*
3.4 DISCUSSION AND CHAPTER CONCLUSION 63

CHAPTER 4: DECLARATIVE FOCUS PROSODY **64**

4.1 INTRODUCTION 64
4.1.1 Basic facts about Japanese focus prosody *64*
4.1.2 Autosegmental-Metrical representation *65*
4.1.3 Post-focus compression (PFC) *66*
4.1.4 Research questions *68*
4.2 METHODOLOGY 69
4.2.1 Stimuli *69*

4.2.2 Recording procedure ... *71*
4.2.3 Processing of data ... *71*
4.3 RESULTS ... 71
4.3.1 Initial focus ... *72*
4.3.2 Medial focus ... *74*
4.3.3 Final focus ... *76*
4.4 DISCUSSION ... 78
4.4.1 Conditional realisation of PFC ... *78*
4.4.2 Indistinguishable cases? ... *81*
4.4.3 No pre-focus reduction ... *82*
4.4.4 Non-F0 focus markers ... *82*
4.5 CHAPTER CONCLUSION ... 83

CHAPTER 5: INTERROGATIVE FOCUS PROSODY ... 84

5.1 INTRODUCTION ... 84
5.2 METHODS ... 84
5.3. RESULTS ... 85
5.3.1 Initial focus ... *85*
5.3.2 Medial focus ... *87*
5.3.3 Final focus ... *89*
5.3.4 Effect of question ... *90*
5.4 DISCUSSION ... 92
5.5 CHAPTER CONCLUSION ... 93

CHAPTER 6: SYNTHESIZING WORD AND SENTENTIAL PROSODY ... 95

6.1 INTRODUCTION ... 95
6.2 SYNTHESIZING WORD PROSODY ... 96
6.2.1 Introduction ... *96*
6.2.2 Methods ... *97*
6.2.3 Results ... *101*
6.2.4 Discussion ... *104*
6.2.5 Interim conclusion ... *105*
6.3 SYNTHESIZING SENTENTIAL PROSODY ... 106
6.3.1 Introduction ... *106*
6.3.2 Methods ... *106*
6.3.3 Results ... *108*
6.3.4 Discussion ... *112*
6.4 CHAPTER CONCLUSION ... 113

CHAPTER 7: GENERAL CONCLUSIONS ... 115

BIBLIOGRAPHY ... 118

APPENDICES ... 128

APPENDIX 1: F0 CONTOURS - STATEMENT VS. QUESTIONS ... 128
1.1 Description of data ... *128*
1.2 Initial focus ... *129*
1.3 Medial focus ... *132*
1.4 Final focus ... *135*
APPENDIX 2: GLOBAL ARTICULATORY PARAMETERS OF LEXICAL CORPUS ... 138
2.1 Description of data ... *138*
2.2 Moraic segmentation ... *138*
2.3 Syllabic segmentation ... *138*
APPENDIX 3: GLOBAL ARTICULATORY PARAMETERS OF SENTENTIAL CORPUS ... 139

3.1 Description of data *139*
3.2 Moraic segmentation *139*
3.3 Syllabic segmentation *140*
3.4 Syllabic segmentation (revised for §6.3.3.4) *141*
APPENDIX 4: ON-FOCUS VS. NEUTRAL FOCUS WORD DURATION 143
APPENDIX 5: F0 CONTOURS – EFFECT OF PRE-LOW RAISING 144
5.1 Description of data *144*
5.2 One-mora words *145*
5.3 two-mora words *145*
5.4 Three-mora words *146*
5.5 Four-mora words *147*

LIST OF FIGURES

FIGURE 1. PROSODIC AND TONAL STRUCTURE OF A PHRASE AKAI **SE**ETAA WA 'RED SWEATER' (FROM PIERREHUMBERT & BECKMAN, 1988, P. 21). 14
FIGURE 2. THE FUJISAKI MODEL (FUJISAKI & HIROSE, 1984, P. 235)..... 17
FIGURE 3. FOUR REPETITIONS OF THE NOUN PHRASE SAKIHODO-NO `MANGˆO "THE MANGO FROM BEFORE". VERTICAL LINES MARK THE START OF THE ACCENTED MORA /MA/ IN `MANGO (VENDITTI & VAN SANTEN, 2000, P. 606). 18
FIGURE 4. THE TARGET APPROXIMATION (TA) MODEL. THE VERTICAL LINES REPRESENT SYLLABLE BOUNDARIES. THE DASHED LINES ARE THE UNDERLYING TARGETS. THE THICK CURVE REPRESENTS THE F0 CONTOUR THAT RESULTS FROM ASYMPTOTIC APPROXIMATION OF THE TARGETS. THE DASHED CURVE SIMULATES THE EFFECT OF WEAK EFFORT. ADAPTED FROM XU & WANG (2001). 19
FIGURE 5. TONAL PATTERN OF THE WORD OYASUMINA**SA**I 'GOOD NIGHT'. 23
FIGURE 6. F0 CURVES OF ATA**MA**-WA VS. MIYAKO-WA (WARNER, 1997, P. 45)..... 25
FIGURE 7. A SCREENSHOT OF ANNOTATION USING PROSODYPRO. 32
FIGURE 8. AVERAGED F0 CONTOURS OF 3-MORA CVCV ACCENTED WORDS. FOR THIS DIAGRAM, THE CAPITALISED SYLLABLE BEARS PITCH ACCENT. X-AXIS SHOWS NORMALISED TIME, WHEREAS Y-AXIS IS F0 IN HZ. 33
FIGURE 9. BOXPLOTS SHOWING ACCENT PEAK F0 UNDER VARIOUS CONDITIONS 34
FIGURE 10. EFFECTS OF ACCENT CONDITION, SPEECH RATE, AND PEAK-TO-END DISTANCE ON THE F0 OF (I) MINIMUM F0 OF PROSODIC WORD, (II) EXCURSION SIZE OF ACCENTUAL FALL, AND (III) MINIMUM F0 VELOCITY IN ACCENTUAL FALL. 36
FIGURE 11. BARPLOT SHOWING MEAN PEAK DELAY RATIO UNDER DIFFERENT PEAK-TO-END DISTANCE AND WORD LENGTH CONDITIONS. 39
FIGURE 12. FORMANT CONTOURS AVERAGED ACROSS 40 REPETITIONS FROM EIGHT SPEAKERS. X-AXIS SHOWS NORMALISED TIME, WITH VERTICAL LINES REPRESENTING MORA BOUNDARIES; Y-AXIS SHOWS FORMANT FREQUENCY IN HZ. 41
FIGURE 13. AN ACCENTED WORD AND AN UNACCENTED WORD. F0 CONTOURS OF MONOMANE (UNACCENTED) 'MIMICKING' AND MI**NA**MINA (ACCENTED) 'EVERYONE' AVERAGED ACROSS 40 REPETITIONS FROM EIGHT SPEAKERS..... 44
FIGURE 14. MEAN MAX F0 OF UNACCENTED WORDS BY WORD LENGTH (MORAE)..... 47
FIGURE 15. TIME-NORMALISED AVERAGE F0 CONTOUR OF FOUR 4-MORA WORDS. X-AXIS SHOWS NORMALISED TIME, WHILE Y-AXIS SHOWS F0 IN HZ. THE FOUR INTERVALS ARE PARTS OF THE CARRIER SENTENCE, THE TARGET WORD, AND THE PARTICLE –MO. 48
FIGURE 16. THE EFFECT OF PEAK-TO-END DISTANCE (NUMBER OF MORAE BETWEEN THE ACCENTED MORA AND WORD END) ON MAXF0 IN SEMITONES (Y-AXIS). BAR COLORS REPRESENT THE LENGTH OF PEAK-TO-END DISTANCE. 48
FIGURE 17. SCATTERPLOT OF MAXF0~MINF0A (N = 2640, R = -0.198, P < 0.001)..... 50
FIGURE 18. CONCEPTUAL REPRESENTATION OF THE UNDERLYING TARGETS OF MEMO-MO 'MEMO ALSO' (HIGH-LOW-LOW) IN THE LEFT PANEL, AND MEN-MO 'FACE ALSO' (FALLING-LOW) IN THE RIGHT PANEL. THE DASHED LINES REPRESENT UNDERLYING TARGETS. 56
FIGURE 19. SNAPSHOT OF ANNOTATION WITH PENTATRAINER1. 58
FIGURE 20. MEAN ACTUAL PEAK DELAY (MS) OF HYPOTHESISED FALLING AND HIGH TARGETS. 59
FIGURE 21. F0 CONTOURS OF THE MINIMAL TRIPLET **ME**N 'FACE' VS. **ME**I 'MAY' VS. **ME**MO 'MEMO' SPOKEN AT NORMAL SPEED, AVERAGED ACROSS 40 REPETITIONS. X-AXIS SHOWS NORMALISED TIME, AND Y-AXIS F0 IN SEMITONS. THE FIRST TWO INTERVALS REPRESENT THE TARGET WORDS, AND THE THIRD INTERVAL IS THE FOLLOWING PARTICLE –MO. VERTICAL LINES ARE MORA BOUNDARIES. 60
FIGURE 22. DISTRIBUTION OF ARTICULATORY TARGETS (M = SLOPE, B = HEIGHT) IN WORD-INITIAL ACCENTS (LEFT PANEL) AND WORD-MEDIAL ACCENTS (RIGHT PANEL). EACH POINT REPRESENTS ONE SPEAKER. 61
FIGURE 23. F0 CONTOURS OF THE SENTENCE O**GA**TA-GA **KU**RU-MADE **NI**NKI-O **KA**ITERU 'OGATA WROTE "NINKI" UNTIL HE CAME' SPOKEN IN FOUR FOCUS CONDITIONS AVERAGED ACROSS 7 SPEAKERS. (FROM A. LEE & XU, 2012) 64
FIGURE 24. F0 CONTOURS OF TWO VERSIONS OF **YA**MANO-WA O**YO**IDERU. THE GRAY DASHED LINE REPRESENTS NEUTRAL FOCUS, WHEREAS THE SOLID BLACK LINE SHOWS NARROW FOCUS ON O**YO**IDERU (FROM VENDITTI ET AL., 2008, P. 465). 66
FIGURE 25. F0 CONTOURS FROM ISHIHARA (2011A) ILLUSTRATING THE EFFECT OF UNACCENTED FOCUS ON PFC (MY ANNOTATION)..... 67
FIGURE 26. AVERAGED F0 CONTOURS (ACROSS 50 REPETITIONS EACH) OF FOUR SENTENCES: **ME**I-GA **MO**MO-O **MI**TA 'MAY LOOKED AT THE THIGH' (111S), **ME**I-GA MOMO-O **MI**TA 'MAY LOOKED AT THE PEACH' (121S), MEI-GA **MO**MO-O **MI**TA

'THE NIECE LOOKED AT THE THIGH '(211S), AND MEI-GA MOMO-O **MI**TA 'THE NIECE LOOKED AT THE PEACH' (221S). X-AXIS SHOWS NORMALISED TIME, WHEREAS VERTICAL LINES REPRESENT WORD BOUNDARIES. LINE COLOUR REPRESENTS DIFFERENT FOCUS CONDITIONS. 73
FIGURE 27. AVERAGED INTENSITY PROFILE OF **ME**I-GA **MO**MO-O **MI**TA 'MAY LOOKED AT THE THIGH' (LEFT) AND MEI-GA MOMO-NI NITA 'THE NIECE RESEMBLED A PEACH'(RIGHT). X-AXIS IS NORMALISED TIME, WHILE Y-AXIS SHOW INTENSITY (DB). 74
FIGURE 28. AVERAGED F0 CONTOURS OF **ME**I-GA **MO**MO-NI NITA 'MAY RESEMBLED THE THIGH' (112S), **MU**UMIN-GA **BU**DOU-NI NITA 'MOOMIN WATCHED MARTIAL ARTS' (332S), **MU**UMIN-GA BUDOU-O **MI**TA 'MOOMIN LOOKED AT THE GRAPES' (341S) AND **MU**UMIN-GA BUDOU-NI NITA 'MOOMIN RESEMBLED THE GRAPES' (342S) IN FOUR FOCUS CONDITIONS. 75
FIGURE 29. AVERAGED INTENSITY PROFILE OF MEI-GA **MO**MO-O **MI**TA 'THE NIECE LOOKED AT THE THIGH' (LEFT), MEI-GA MOMO-O **MI**TA 'THE NIECE LOOKED AT THE PEACE' (RIGHT)..... 76
FIGURE 30. AVERAGED F0 CONTOURS OF MEI-GA **MO**MO-NI NITA 'THE NIECE RESEMBLED THE THIGH' (212S), AND NOUMIN-GA **BU**DOU-NI NITA 'THE FARMER RESEMBLED MARTIAL ARTS' (432S) IN FOUR FOCUS CONDITIONS. 77
FIGURE 31. AVERAGED INTENSITY PROFILE OF **ME**I-GA MOMO-O **MI**TA 'MAY LOOKED AT THE PEACH' (LEFT) AND **ME**I-GA MOMO-NI NITA 'MAY RESEMBLED A PEACH' (RIGHT). 77
FIGURE 32. AVERAGED F0 CONTOURS OF NOUMIN-GA **BU**DOU-O **MI**TA 'THE FARMER WATCHED MARTIAL ARTS' (431S), AND NOUMIN-GA BUDOU-O **MI**TA 'THE FARMER LOOKED AT THE GRAPES' (441S) IN FOUR FOCUS CONDITIONS. 78
FIGURE 33. AVERAGED F0 CONTOURS OF MEI-GA MOMO-NI NITA 'THE NIECE RESEMBLED A PEACH' (222S) AND NOUMIN-GA BUDOU-NI NITA 'THE FARMER RESEMBLED GRAPES' (442S) IN FOUR FOCUS CONDITIONS. 79
FIGURE 34. AVERAGED F0 CONTOUR OF **ME**I-GA MOMO-NI NITA 'MAY RESEMBLED A PEACH' (122S), AND **MU**UMIN-GA **BU**DOU-O **MI**TA 'MOOMIN WATCHED MARTIAL ARTS' (331S) IN FOUR FOCUS CONDITIONS. 81
FIGURE 35. AVERAGED F0 CONTOURS OF **ME**I-GA **MO**MO-O **MI**TA? 'DID MAY LOOK AT THE THIGH?' (111Q), **ME**I-GA MOMO-O **MI**TA? 'DID MAY LOOK AT THE PEACH?' (121Q), MEI-GA MOMO-NI NITA? 'DID THE NIECE RESEMBLE A PEACH?' (222Q), **MU**UMIN-GA **BU**DOU-O **MI**TA? 'DID MOOMIN WATCH MARTIAL ARTS?' (331Q), AND **MU**UMIN-GA BUDOU-O **MI**TA? 'DID MOOMIN LOOK AT THE GRAPES' (341Q), AND NOUMIN-GA BUDOU-NI NITA? 'DID THE FARMER RESEMBLED GRAPES?' (442Q) IN FOUR FOCUS CONDITIONS..... 86
FIGURE 36. AVERAGED F0 CONTOURS OF **ME**I-GA MOMO-NI NITA? 'DID MAY RESEMBLE A PEACH?' (122Q), MEI-GA **MO**MO-O **MI**TA? 'DID THE NIECE LOOK AT THE THIGH? '(211Q), MEI-GA **MO**MO-NI NITA? 'DID THE NIECE RESEMBLE THE THIGH?' (212Q), **MU**UMIN-GA BUDOU-NI NITA? 'DID MOOMIN RESEMBLE THE GRAPES' (342Q), NOUMIN-GA **BU**DOU-O **MI**TA? 'DID THE FARMER WATCH MARTIAL ARTS?' (431Q), AND NOUMIN-GA **BU**DOU-NI NITA? 'DID THE FARMER RESEMBLE MARTIAL ARTS' (432Q) IN FOUR FOCUS CONDITIONS. 88
FIGURE 37. AVERAGE F0 CONTOURS OF **ME**I-GA **MO**MO-NI NITA? 'DID MAY RESEMBLE THE THIGH?' (112Q), MEI-GA MOMO-O **MI**TA? 'DID THE NIECE LOOK AT THE PEACH?' (221Q), **MU**UMIN-GA **BU**DOU-NI NITA? 'DID MOOMIN WATCH MARTIAL ARTS?' (332Q), AND NOUMIN-GA BUDOU-O **MI**TA? 'DID THE FARMER LOOK AT THE GRAPES?' (441Q) IN FOUR FOCUS CONDITIONS. 90
FIGURE 38. AVERAGED F0 CONTOURS OF **ME**I-GA **MO**MO-O **MI**TA 'MAY LOOKED AT THE THIGH' (IIIQ) AND MEI-GA MOMO-NI NITA 'THE NIECE RESEMBLED THE PEACH' (222Q) SPOKEN IN QUESTION VS. STATEMENT AND INITIAL VS. NEUTRAL FOCUS. 91
FIGURE 39. ON-FOCUS MAXF0 (IN HZ) ACROSS SENTENCE TYPES, ACCENT CONDITIONS AND FOCUS CONDITIONS. 92
FIGURE 40. EXTRACTION OF MODEL PARAMETERS FOR EACH LABELED INTERVAL BY PENTATRAINER1 (JITEN-NI ME**MA**I-MO NOTTEMASU 'THE WORD "MEMAI" TOO IS FOUND IN THE DICTIONARY'). IN ORDER, THE SECOND TO THE FIFTH TIERS SHOW SLOPE, HEIGHT, STRENGTH, AND DURATION OF THE LABELED INTERVALS. THE PARAMETER NUMBERS IN TIERS 2-5 ARE EXTRACTED RATHER THAN MANUALLY ENTERED..... 98
FIGURE 41. FUNCTIONAL ANNOTATION IN PENTATRAINER2 (SAME SENTENCE AS IN **FIGURE 40**). THE LABELED FUNCTIONS ARE TONE AND DEMARCATION..... 99
FIGURE 42. TOBI ANNOTATION FOR AMTRAINER (SENTENCE AS **FIGURE 40**) 100
FIGURE 43. MEAN PEARSON'S R OF THE THREE TRAINING TOOLS..... 102
FIGURE 44. MEAN PEARSON'S R OF SPEAKER-DEPENDENT SYNTHESIS BY BOTH PENTATRAINERS. 103
FIGURE 45. MEAN PEARSON'S R OF SPEAKER-INDEPENDENT PREDICTIVE SYNTHESIS (JACKKNIFE PROCEDURE). 104
FIGURE 46. FUNCTIONAL ANNOTATION IN PENTATRAINER2. THE LABELED FUNCTIONS ARE TONE, SENTENCE TYPE, DEMARCATION AND FOCUS. THE FIFTH TIER (GLOSS) IS NOT INCLUDED IN ACTUAL ANALYSIS. 107

FIGURE 47. AN INTERFACE OF PENTATRAINER2 FOR VISUAL INSPECTION OF SYNTHESIS ACCURACY. THE BLUE CURVE IS THE F0 CONTOUR OF A NATURAL UTTERANCE WHEREAS THE RED DOTTED CURVE REPRESENTS THE CORRESPONDING RESYNTHESIS. THE TARGET SENTENCE IN THIS EXAMPLE IS **MU**UMIN-GA **BU**DOU-O **MI**TA 'MOOMIN WATCHED MARTIAL ARTS', WITH FOCUS ON THE FIRST WORD. 109
FIGURE 48. MEAN STRENGTH VALUES UNDER DIFFERENT COMMUNICATIVE FUNCTION CONDITIONS (REVISED ANNOTATION). 112

LIST OF TABLES

Table 1. F0 scaling rules in Beckman and Pierrehumbert (1986b). 24
Table 2. List of stimuli used and corresponding English gloss. For simplicity, tonal representation used in the first column of this Table follows Haraguchi (2002), which comprises only H and L. 29
Table 3. Average mora duration of each speaker (ms). 30
Table 4. Information of participants in Chapters 2 and 3 30
Table 5. Variables used in **Table 6** 37
Table 6. Pearson's correlations of peak delay ratio (PDR) vs. numerous variables. 37
Table 7. Variables used in **Table 8**. 46
Table 8. Pearson's correlations of normalised data (converted into semitones using utterance-initial F0 value). Non-significant (p>0.05) correlations are not displayed. Data of unaccented words have been removed. 49
Table 9. Pearson's r of RiseSize~MinF0A (converted into semitones using utterance-initial F0 value). Data were subsetted into four groups according to peak-to-end distance and word length. 50
Table 10. Overall learning accuracy in RMSE and r using PENTAtrainer1 under two imposed targets. 62
Table 11. Information of participants in Chapters 4 and 5 69
Table 12. Corpus used in Chapters 4 and 5 70
Table 13. Functional labels used in the three annotation schemes for PENTAtrainer1 and PENTAtrainer2. 99
Table 14. Definition of the functional labels in **Table 13**. 99
Table 15. Label extraction criteria for AMtrainer 100
Table 16. Learning accuracy of AMtrainer, PENTAtrainer1 and PENTAtrainer2. 101
Table 17. Accuracies of speaker-dependent synthesis by both PENTAtrainers. 103
Table 18. Accuracies of speaker-independent predictive synthesis 104
Table 19. Information of participants in natural judgment test 108
Table 20. Accuracy of speaker-dependent synthesis. 109
Table 21. Synthesis accuracy of PENTAtrainer2 under Jackknife procedure by sentence type and focus condition. 110
Table 22. Mean naturalness ratings by focus condition and sentence type. 110

CHAPTER 1: INTRODUCTION

1.1 Background

Tokyo Japanese (Japanese henceforth) has always been a testing ground for phonological and phonetic theories because of its accentual nature which resembles both tonal and non-tonal languages. Even after over half a century (Arisaka, 1941) since it began to catch the attention of linguists, Japanese prosody continues to interest scholars in such fields as theoretical linguistics, speech synthesis, and phonetics. But after so many years of hard work, rival theories continue to co-exist. One reason for such a plight is the 'lack of reference' problem[1] (Xu, 2011b), whereby there is no unambiguous lexical anchor to prosodic correlates like there is word meaning to segments. The lexical function of Japanese pitch accent helps to some extent, but there are many other factors that cause surface F0 variations. Circularity ensues when researchers attribute such acoustical differences to hypothetical prosodic categories (e.g. tone, communicative functions) which do not exist. The study of prosody is thus challenging and requires more stringent methodology.

The complexity of Japanese prosody further compounds this challenge. One example is the oft-questioned size of its tonal inventory. As will be elaborated in Chapter 2, there is a discrepancy between how many tones native listeners can discriminate and how many distinct tone levels phoneticians have repeatedly identified in this language. Whereas in a tonal language like Chinese, tonal categories can be easily identified by a native speaker linguist, the same is not as straightforward in the case of Japanese[2]. Even more complicated is the interaction among multiple prosodic categories. As will be shown in Chapters 4 and 5, when sentence type and focus interact with lexical pitch accent, the resulting F0 contour is often more than an amalgamation of their respective prosodic markers. The correct interpretation of such complex patterns, therefore, requires a multifaceted strategy, which starts with a careful experiment design. Ideally, a prosodic analysis should offer a mechanistic explanation about how prosody works, through evidence from multiple angles such as acoustic analysis and analysis-by-synthesis, which the present work aims to offer (see Xu, 2006 for a related set of research principles that this dissertation strives to adhere to).

Since the publication of Pierrehumbert and Beckman's (1988) seminal work, the Autosegmental-Metrical (AM) Theory has dominated the study of Japanese pitch accent and

[1] Here 'reference' means a pivot that serves as both a starting point of inquest and a point that one can comfortably fall back on. For segments, word identity serves as such a reference, to which native speakers have no difficulty accessing (Xu, 2011b).

[2] As will be discussed in more details in Chapter 2, whereas native speakers of Japanese have no difficulty distinguishing accented and unaccented words, it remains an open question to linguists whether there are two High tones or one, especially in terms of articulatory planning.

intonation in laboratory phonology. On the other hand, in speech synthesis, the Fujisaki Model (e.g. Fujisaki & Hirose, 1984) remains a highly influential approach. Nevertheless, both models suffer from limitations in their theoretical construct that lead to restrictions of their applicability. There is thus a need for an approach that can synthesise a wide range of intonation types, and preferably with predictive power.

This dissertation aims to find an alternative to AM and the Fujisaki Model. The Parallel Encoding and Target Approximation (Xu, 2005; PENTA henceforth) Model is used as a theoretical framework. While these three models differ from one another in their intended application, AM and the Fujisaki Model have been widely used in previous studies on Japanese prosody, and thus provide a context for the adoption of PENTA in this project.

1.2 Autosegmental-Metrical Theory

Autosegmental-Metrical (AM) Theory was first proposed for English intonation (Beckman & Pierrehumbert, 1986a; Pierrehumbert, 1980), based on the insights from Autosegmental Theory (Goldsmith, 1976). This framework was soon extended to Japanese intonation (Pierrehumbert & Beckman, 1988), and has since dominated intonational phonology. In AM, tones (either H or L) form a tier separate from segments (hence 'Autosegmental'), and are aligned with stressed syllables on the basis of the 'Metrical' pattern of the text. Below is the representation of prosodic units in AM.

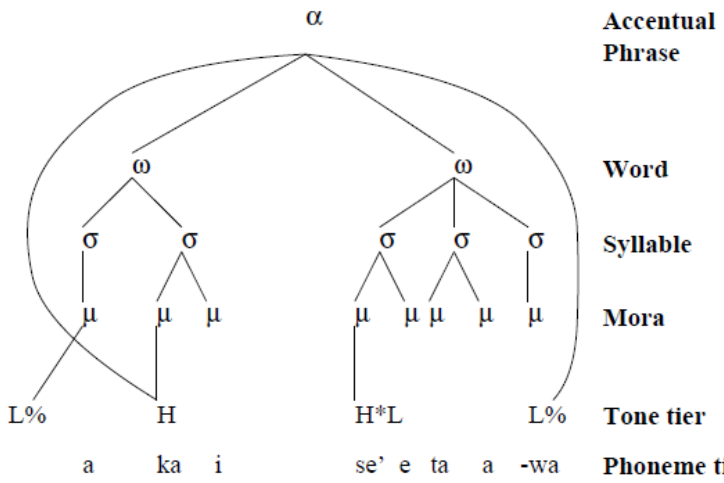


Figure 1. *Prosodic and tonal structure of a phrase* akai **se**etaa wa[3] *'red sweater'* (from Pierrehumbert & Beckman, 1988, p. 21).

[3] Throughout this dissertation, the syllable that bears lexical pitch accent is boldfaced and underlined.

Note a number of special features in AM. First, a tone can be associated with any node, i.e. mora or the utterance. Thus there is a distinction between tones that are linked to certain tone-bearing units of a word (e.g. pitch accent) and boundary tones that are affiliated to phrase edges (e.g. end of an utterance). Second, as a phonological theory it is not designed to capture phonetic features that are not deemed relevant to the grammar, such as anticipatory tonal dissimilation (see Chapter 3).

J-ToBI[4] (Campbell & Venditti, 1995; Venditti, 2005) is an intonation labeling convention that is based on the principles of AM Theory. In AM, the accentual phrase (often a word, and AP henceforth) is marked by an initial rise to a high target around the second mora of the phrase, then F0 gradually falls to a low at the right edge of the phrase (see Appendix 5 for examples of the typical F0 contour of a Japanese word). Thus, a Japanese word like *anoma**ma*** (LHHH) where the pitch accent falls on the final mora would be transcribed as %LH- H*+L(L%). An initial rise is labeled as %L H- (where % denotes a boundary), of which the phrasal tone H- typically falls on the second mora. This phrasal H- is not to be confused with the following H*, which is aligned with the accent peak (second ***ma***). Together with the tail L, the H*+L label is used to annotate an accent peak and the ensuing accentual fall. The boundary tone L% is found at the right edge of an AP. Note that in AM and J-ToBI, not all tone bearing units receive tone specification. The phonetic realisation of tonally-unspecified units is a result of interpolation of the specified targets. Moreover, some argue that the alignment of tones with segments is a phonologically conditioned factor. For example Arvaniti and Ladd (2009) reported that the location of a F0 peak was determined by the proximity of adjacent tonal targets (i.e. tonal crowding), and by its relative importance over other tones with regards to phonetic realisation. In other frameworks like the Fujisaki and the PENTA models, peak timing is an end result of various tonal processes combined and is thus not specified in the input.

Attempts have been made to extend the application of ToBI (prototype of J_ToBI) to F0 contour synthesis. In particular, Pierrehumbert has proposed a rule for determining the shape of interpolation between sparse tones: 'when one target is at the bottom of the pitch range… the transition is monotonic… when neither target is near the baseline, a sagging transition is used' (Pierrehumbert, 1981, p. 988). These rules, however, are generalised from her observations of actual F0 production, and are not motivated by any physiological or mechanistic reasons. Subsequent work on Japanese F0 synthesis includes Beckman and Pierrehumbert (1986b), where explicit rules of F0 scaling were stated. Considering the properties of ToBI, one would expect that the accuracy of synthesis would to a large extent depend on the labeler's annotation proficiency. In other words, because ToBI is a linear system whose input is surface acoustical landmarks, i.e. part of the output, precise labeling would directly contribute to the accuracy of the synthesised output.

[4] I am referring to the original version of J-ToBI. There is also an extended version (X-JToBI) for spontaneous speech, see Maekawa et al (2002).

There is no doubt that ToBI is widely popular as an annotation scheme, but whether the number of tonal labels equals the number of linguistic and discoursal contrasts in a given language remains a doubt. One obvious concern is the size of tonal inventory, where the number of labels can exceed the number of contrasts an average speaker is able to conceive. To quote Calhoun, 'people perceive a particular pitch variation as being categorical only if that change signals a meaning change in the given acoustic and discourse context' (Calhoun, 2004, p. 1). If there are more labels than perceivable contrasts, an annotation system would be a phonetic transcription rather than a phonological one, but even so ToBI still faces a common problem of inter- and intra-labeler reliability (Beckman, Hirschberg, & Shattuck-Hufnagel, 2005). Examples of labels that do not seem to contrast in meaning abound, for example H* and L+H* in English (Beckman et al., 2005, p. 45). At their own admission, the 'intrinsic meaning' of most intonational contours in AM still remains 'controversial' and 'elusive' (Hirschberg, 2004, p. 533).

In relation to the research question, AM phonologists have identified three tone levels in a Japanese word (L, H-, H*). Note that although the distinction between H- and H* was not originally based on F0 but function, Beckman and Pierrehumbert (1986b, p. 177) stated that '(as) the H tone of the lexical accent is generally higher than the phrasal H of the accentual phrase... we automatically assign it a higher target value within the local pitch range'. Then, in order to motivate an alternative with two tones, it is necessary to show that such surface differences are not perceivable, and are derived through an articulatory phenomenon known as pre-low raising[5]. We shall return to this issue in §2.4.

1.3 The Fujisaki Model

The Fujisaki Model is a superpositional model based on Öhman's (1967) earlier work on word prosody. It assumes that although intonation curves are continuous in time and frequency, they originate in a series of discrete events triggered by the speaker. Unlike other approaches to intonation modeling, many elements in the Fujisaki Model are motivated by the physical or physiological properties behind the generation of surface prosody. Two second-order critically damped linear systems (phrase and accent commands) are superposed on a constant base frequency and result in the surface F0 contour. The phrase command is a set of impulses, whilst the accent command is set of stepwise functions. The baseline is gradually decaying, corresponding to declination which is found across languages. Generally, local humps on a F0 contour are a result of accent control, whereas a change in global slope is due to a response of phrase control mechanism to a positive/negative impulse (e.g. phrase-final lowering).

[5] Pre-low raising is a local articulatory phenomenon where the F0 of a High tone becomes higher when preceding a Low tone. Please refer to §2.4.1 for more details.

Figure 2 shows the workings of the Fujisaki Model. There are phrase commands that express (global) effects from the syntactic phrasing. For example, T03 represents final lowering that marks the end of the utterance. There are also accent commands that express (local) effects from lexical accents (i.e. initial rise and accentual fall, see §2.1.1 for more information on Japanese lexical prosody). The amplitude and timing (beginning and end) of the accent commands reflect the accent condition of a given word. Then, the final logarithmic F0 is controlled in proportion to the sum of these two types of commands.

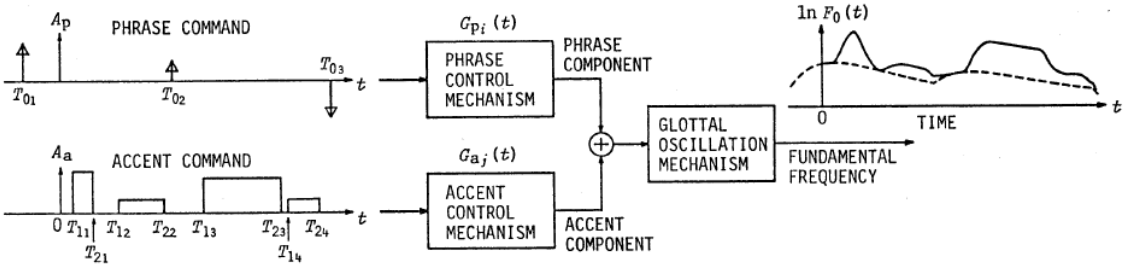


***Figure 2**. The Fujisaki Model (Fujisaki & Hirose, 1984, p. 235)*

One major difference between the Fujisaki Model and AM is that the former is superpositional while the latter is linear. In this regard, the Fujisaki Model is more empirically testable in that the output is the result of a combination of command settings; in AM, on the other hand, part of the input is found in the output (i.e. the acoustical landmarks like peaks observed in the surface contour). That said, both frameworks are rather restricted in the types of contours they are capable of replicating in synthesis. Venditti and van Santen (2000) raised this example that apparently challenges both theories in **Figure 3**. The four panels represent possible surface contours of the same utterance, which consists of an unaccented word *sakihodo-no* followed by an accented word ***ma**ngo*. Notice that in the four panels, although the accented word **ma**ngo consistently undergoes accentual fall, it does not always have an initial rise, especially in panel A. The Fujisaki Model is unable to model this variable height of local peak, except by redefining the scope of the accent command. This shortcoming stems from the fact that this model has only 2 command types, which however are delegated a wide range of tasks (e.g. question, focus, emotions, demarcation). Next, we will consider the PENTA Model which takes a step forward by incorporating communicative functions into its theoretical construct, which as a result can accommodate a wider range of F0 contour types.

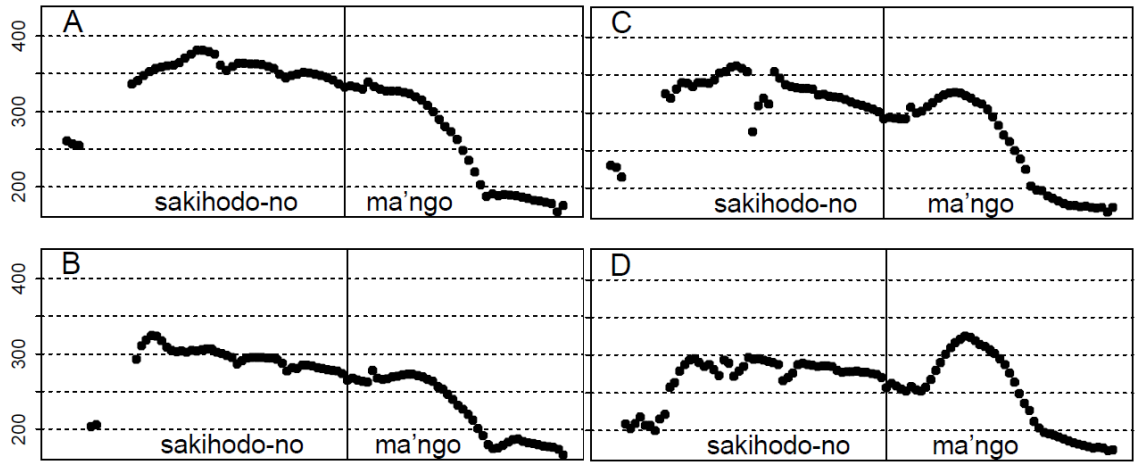


***Figure 3**. Four repetitions of the noun phrase sakihodo-no `mangˆo "the mango from before". Vertical lines mark the start of the accented mora /ma/ in `mango* (Venditti & van Santen, 2000, p. 606).

1.4 PENTA Model[6]

The Parallel Encoding and Target Approximation (PENTA) Model is currently the most comprehensive alternative[7] to AM Theory. It aims to explain speech phenomena with both articulatory mechanisms and communicative functions. It assumes that speech conveys communicative information and is produced by an articulatory system. **Figure 4** shows how F0 is realised under PENTA.

[6] There are also other important theories about intonation such as the IPO Model ('t Hart et al., 1990) and Tilt Model (Taylor, 1994). Because these theories are less influential in Japanese prosody compared to AM and the Fujisaki Model they will not be discussed here. The interested reader is referred to the relevant references for details.

[7] PENTA is an alternative to AM in the sense that it offers a different angle on speech prosody (articulatory rather than phonological).

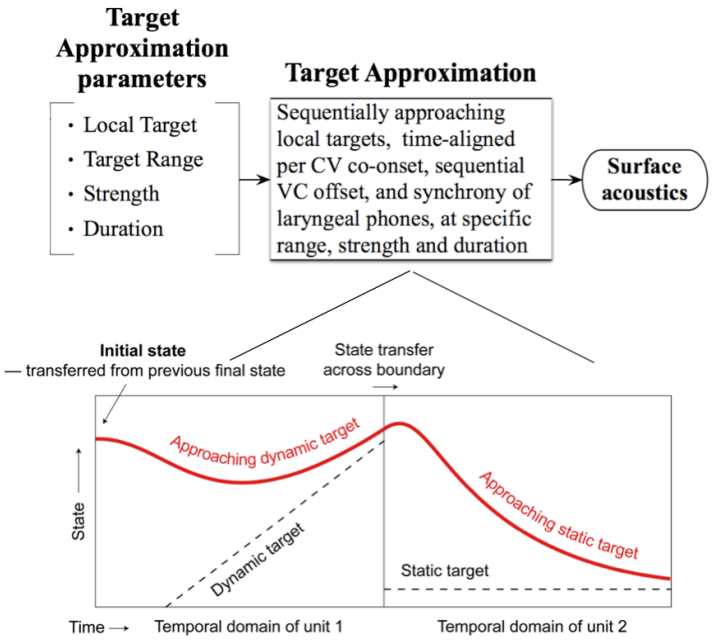


***Figure 4**. The Target Approximation (TA) model. The vertical lines represent syllable boundaries. The dashed lines are the underlying targets. The thick curve represents the F0 contour that results from asymptotic approximation of the targets. The dashed curve simulates the effect of weak effort. Adapted from Xu & Wang (2001).*

Melodic primitives are the parameters that control the Target Approximation (TA) process in PENTA. They consist of local target, target range, strength, and duration. Within the TA process, the values of parameters are specified by distinctive encoding schemes symbolically and numerically. The local target can be specified to be static (i.e. [high], [low] or [mid]), or dynamic (i.e. [rise] or [fall]). These articulatory parameters are used to encode the communicative function of lexical tone and lexical stress. The effect of the target range is its determination of F0 scope across implemented local pitch targets. Its specification is through height (high, low, or mid) as well as span (wide or narrow). Focus, sentence type or modality, new topic or turn taking can encode specifications for target range. Strength (represented as λ in PENTA) refers to articulatory strength, and determines the speed of approaching a local pitch target. Like in other parts of the body, it is related to muscle stiffness. Strength is either strong or weak; the speed of approaching the target is faster when the strength is specified to be strong. Duration gives the period of time interval within which a target is approximated. Usually, this is the length of a syllable.

TA in PENTA implements the articulatory movement that produces acoustic features that convey the desired communicative meaning. The lower panel of **Figure 4** how it realises a F0 contour in Mandarin. Each syllable is fully-specified with an underlying tone target, which is asymptotically approached syllable by syllable. At the end of a syllable, the momentum of approximation is inherited by the next syllable as a carryover effect. In this framework, F0 is therefore not a direct representation of prosodic meaning but an observable output resulting

from an articulatory implementation of the intended meaning (i.e. communicative functions). Thus in PENTA, acoustical landmarks are not seen to have communicative meanings, but are results of communicative functions being encoded in parallel through target approximation.

Finally, there is a set of communicative functions, each of which has its own encoding scheme. These functions are independent of one another, but are encoded in parallel, such that at any point in an utterance there can be multiple functions in play influencing the surface F0. At present a number of communicative functions have been attested, namely Lexical (tone), Sentential (statement vs. question), Focal (focus), Topical, Demarcative (phrasing and syntax), and most recently Emotional (Chuenwattanapranithi, Xu, Thipakorn, & Maneewongvatana, 2008). The workings of parallel encoding will become clearer in Chapters 3 and 6 through its application in PENTAtrainer2.

Note that in PENTA the introduction of any new functional category is not arbitrary, but governed by a set of stringent principles (Xu, 2006). Although not formally part of the model per se, these guidelines are faithfully adhered to by the practitioners of PENTA. These include rigid experimental control by eliciting strictly minimal contrasts as well as an emphasis on looking for the actual mechanism (e.g. physiological, physical) instead of only reporting systematic differences in the acoustic signal. The research principles adopted in PENTA ensures that an otherwise abstract category (communicative function) is motivated with empirical support.

Given the properties of PENTA, it is well suited for circumventing the 'lack of reference' problem in speech prosody. Compared to AM, of which some of the phonological units (i.e. the input, e.g. H* and L) come from the surface realisation (i.e. the output), input in PENTA consists of communicative functions which are empirically established. That the input categories are blind to the output (i.e. an F0 peak is not named by its height, but the function it serves, e.g. a 'Tone 1' under 'narrow focus') thus takes PENTA one step away from running into circularity. Moreover, in AM the course of F0 trajectories is a linear/sagging interpolation (Pierrehumbert, 1981) between sparse tones, which is not an articulatorily motivated mechanism; whereas the phonetic implementation of articulatory targets in PENTA is guided by the TA process, motivated by physical and physiological factors (Xu & Wang, 2001).

Compared to the Fujisaki Model, in which communicative information is encoded in only two types of commands, there is no stipulation in PENTA on how many communicative functions there are in a given language. Though this flexibility means a larger degree of freedom, it also allows the representation of a given communicative function to remain invariant across different contexts. Besides, with convenient tools (PENTAtrainer1 and PENTAtrainer2, see Chapter 6) available for analysis-by-synthesis, the validity of a PENTA analysis can be easily verified by comparing multiple hypotheses. All in all, PENTA is an ideal approach for its (i) input that is blind to output; (ii) invariant representation for a given functional category; and for the (iii) availability of easy means to test multiple hypotheses through analysis-by-synthesis.

On a side note, although both PENTA and AM aim to explain the surface realisation of speech prosody, note that the former is a theory of articulation whereas the latter is one of intonation phonology. The connection between intended articulatory targets that PENTA aims to

unearth and abstract phonological units is an empirical question, and can only be identified when both are well understood. This dissertation will thus focus on the connection between articulatory targets and surface F0 realisation in several aspects of Japanese prosody.

CHAPTER 2: ACOUSTIC PROPERTIES OF LEXICAL ACCENT

2.1 Background

This Chapter investigates the nature of lexical pitch accent in Japanese. Whereas this topic has been extensively studied from the perspectives of physiology (e.g. Sawashima, Kakita, & Hiki, 1973; Simada & Hirose, 1970; Sugito, 2003), engineering (e.g. Fujisaki, 1977), production (e.g. Kubozono, 1993; Pierrehumbert & Beckman, 1988) and perception (e.g. Sugito, 1982; Sugiyama, 2012; Vance, 1995), my goal is to draw from these findings as well as the new corpus reported in this Chapter, and motivate a revised account of the tonal inventory of this language. Specifically, I seek to argue that there is only one High tone in Japanese; what causes the surface acoustical variation of the High tone in different accent conditions is pre-low raising, a well-known articulatory phenomenon.

Since the body of previous work on Japanese lexical prosody is immense, literature review will be limited to phonetic studies that are related to my proposal. Following a brief introduction to the phonology of Japanese pitch accent, key findings on F0 scaling, F0 alignment, the distinction between final accented and unaccented words, as well as non-F0 cues to pitch accent will be reviewed. Then, the methodology of the production experiment will be described, before a report of the general acoustic analysis of the corpus. The findings in this Chapter will partly form the basis of my proposal in the next Chapter.

2.1.1 The phonology of Japanese pitch accent

In Japanese, a word is either accented or unaccented. An accented word ends in a low tone, whereas an unaccented word ends high. Pitch accent is associated with one of the morae of the accented word. A word can have at most one pitch accent, which is phonetically characterised by an accent peak followed by a sharp fall (known as the accentual fall). If pitch accent is hosted by a heavy syllable (e.g. CVV, CVn), it is associated with the first mora of that syllable.

For example, a word *hana* can be unaccented (LH) 'nose', bear initial accent (HL) '(a girl's name) Hana', or final accent (LH) 'flower'[8]. The difference between accented and unaccented words becomes clearer when appending a nominative case particle *–ga* to the nouns (forming *hana-ga*) — unaccented (LH-H), initial accent (HL-L), and final accent (LH-L).

[8] According to the tone marking scheme by Yamada (1892), cited in Sugiyama (2012).

Although unaccented and final accented words appear to be identical when in isolation, they can be easily distinguished by attaching a following particle to them.

Many describe Japanese as a pitch accent language. This is because tonal specification is only required for one mora, then the tonal pattern of the rest of the word can be worked out by phonological rules. According to Poser (1984), the tonal pattern of a word follows these rules: ① make everything up to and including the accent High, ② make everything following the accent Low, and ③ make the first mora Low if the first syllable is light and the following mora is High. Using these rules, the word in **Figure 5** would have a H tone on the penultimate syllable –*sa*, which bears the pitch accent (per ①). All syllables after the pitch accent bear a L tone (per ②). Then the first mora bears a L tone (per ③). Finally, make all the remaining syllables H (per ①).

o	ya	su	mi	na	**sa**	i
L	H	H	H	H	H	L

***Figure 5**. Tonal pattern of the word oyasumina**sa**i 'good night'.*

AM Theory (Pierrehumbert & Beckman, 1988) offers a more parsimonious representation. In AM, tones are sparsely specified, meaning that syllables can be unspecified for tone. The example in **Figure 1** (Page 14) was a noun phrase consisting of an unaccented adjective *akai* 'red' followed by an initial accented noun *seetaa-wa* 'sweater' (the –wa is a topic marker). Based on Poser's (1984) rules the surface tonal pattern of this phrase would be *akai* (LHH) *seetaa-wa* (HLLL-L). However, as **Figure 1** showed, in AM this phrase only required five tones to represent the intonation contour of the entire domain. The initial L% was delimitative, marking the beginning of the phrase, and provided an anchor for the phrase-initial rise (cf. Poser's rule ③). The final L% was associated with the AP and marked its end. H*L was a bitonal accent, which represented pitch accent in Japanese. The H* and L stood for the accent peak and the ensuing accentual fall. Finally, the H in *akai* was a phrase accent usually associated with the second mora of a word. The surface F0 contour was the linear or sagging interpolation between these tones (Pierrehumbert, 1981).

2.1.2 F0 scaling

Scaling, or the F0 level of a given tone, is central to any model of speech prosody. Many factors are known to affect the height of tones. When developing their F0 synthesiser for Japanese, Beckman and Pierrehumbert (1986b) defined several explicit rules on F0 scaling, based on their earlier experimental findings:

High versus Low	A low tone is lower than a high tone in the same local pitch range setting.
Intrinsic prominence of accents	The H in an accent is higher than the phrasal H tone.
Boundary tone weakening	The L% boundary tone is higher if the first syllable of the upcoming phrase is long or accented.
Boundary strength	The L% boundary tone is lower at an intermediate phrase boundary than at an accentual phrase boundary, and lower yet at an utterance boundary.

***Table 1**. F0 scaling rules in Beckman and Pierrehumbert* (1986b).

The first rule 'High versus Low' should be self-explanatory, whereas the second rule states that H* (accent peak) should be higher than H- (phrasal high tone associated with the second mora of an unaccented word). Although the distinction between H* and H- was not originally introduced based on scaling difference, but on their function, 'Intrinsic prominence of accents' stipulates that there are two distinct F0 levels of High tone in Japanese. The third rule is devised for situations where AP-initial rise is absent, that is, when a word starts with a heavy syllable (e.g. *kenka* 'quarrel' HHH). Since Beckman and Pierrehumbert (ibid.) chose to retain the %L tone, a weak-boundary specification is necessary to account for the absence of initial rise. The final rule defines how much lowering a word should end in depending on its prosodic position. With these rules, together with other rules that concern sentential aspects of intonation, Beckman and Pierrehumbert (1986b) illustrated how F0 scaling could be determined in synthesis.

However, tone identity and phonology are not the only source of variation in scaling; there are also phonetic effects, such as pre-planning, speech rate and microprosodic effects. The pre-planning effect of sentence length on F0 scaling has been reported for English. It was shown that the first F0 peak in a longer sentence is higher, compared to the first peak in a shorter sentence (Cooper & Sorensen, 1981; O'Shaughnessy & Allen, 1983). A comparable phenomenon was also reported for Japanese word prosody. Selkirk and colleagues (2004) showed that AP-initial (or MiP in their terms) rise has a larger excursion size in a longer word than in a shorter word. They compared 3-mora, 5-mora, and 7-mora words, and found that this trend held across speakers and word length contrasts. Similarly, Warner (1997) observed that five-mora noun phrases had a higher phrasal peak (H-) than four-mora noun phrases.

While most studies focus on acoustical landmarks such as F0 peaks or sharp turning points, the transition between these landmarks has received less attention. **Figure 6** shows hypothetical F0 contours predicted by AM Theory (Pierrehumbert & Beckman, 1988). In the right panel F0 gradually drops from the phrasal H to word end, and this is described in AM terms as a linear interpolation between H- and L%. The more interesting issue is in the left panel, where an abrupt initial rise is followed by a gradual rise from the phrasal H to the accent peak (H*). The same pattern is also observed in the data reported in this Chapter (see Appendix 5), and is most obvious when accent occurs late in a word. This transition is easily accounted for in terms of

linear interpolation, between a lower phrasal H and a higher accent peak; but it is less straightforward in Target Approximation terms. While the focus of Chapter 2 will be on accent peak, this interesting issue will be revisited in §6.2, where I will argue that articulatory strength has a role to play in this upward transition.

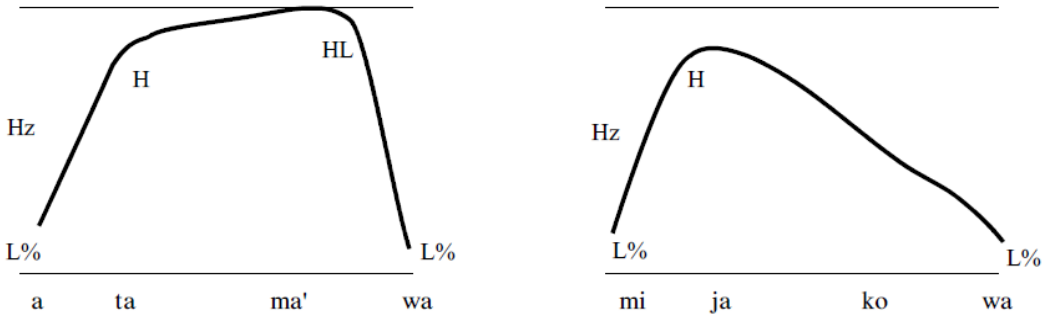


***Figure 6**. F0 curves of ata**ma**-wa vs. miyako-wa* (Warner, 1997, p. 45).

The effect of speech rate on F0 has been widely observed. Cooper and Sorensen (1981) found that overall F0 was lower when one spoke slowly. On a related note, in a perception study higher F0 was found to be perceived as faster speech (Rietveld & Gussenhoven, 1987). Also, fast speech and time pressure are associated with target undershoot (Cheng & Xu, 2014; Lindblom, 1963; Xu & Wang, 2001), because there is a physical limit to how fast F0 can be changed (Xu & Sun, 2002). Other effects of speech rate have been reported, but details varied across studies as well as speakers (e.g. Caspers & van Heuven, 1993; Fougeron & Jun, 1998). As such, these known effects of speech rate need to be taken into consideration when interpreting data on F0 scaling.

Finally, segments can affect the scaling of F0 too. Low vowels are known to bear lower F0 than high vowels (Whalen & Levitt, 1995) across languages. The mechanism behind this phenomenon has been debated, some plausible accounts include laryngeal configuration (Sapir, 1989; Shadle, 1985), subglottal pressure (Steele, 1986), and strategy to enhance vowel contrast (Diehl & Kluender, 1989). Meanwhile, under consonantal perturbation F0 rises immediately after a voiceless stop, and vice versa (Ohde, 1984), due possibly to changes in vocal cord tension. All in all, scaling in Japanese prosody is subject to at least lexical, phonological, and phonetic factors.

2.1.3 F0 alignment

Whereas it is a non-issue in PENTA, the temporal alignment of F0 turning points (e.g. peaks, valleys, elbows) is phonologically specified in models like AM. The timing of F0 landmarks are found to be affected by such factors as phrase position (Arvaniti & Ladd, 1995; Myers, 2003), speech rate (Caspers & van Heuven, 1993) and tonal crowding (i.e. when two or more tones are

associated with the same tone-bearing unit or with adjacent units, see Arvaniti, Ladd, & Mennen, 2006; Arvaniti & Ladd, 2009).

A number of studies have been conducted looking at the temporal alignment of tones in Japanese. Ishihara (2006) conducted a detailed analysis of tonal alignment in Japanese. He investigated the effect of syllable structure and speaking mode on syllable timing, and found that (i) in CVCV, F0 peak was aligned with the beginning of the vowel of the syllable following the accented syllable, (ii) for CVn, with the end of the first mora of the accented syllable; and (iii) for CVV, about 70 percent into the two-mora vowel of the accented syllable for CVR and CVV. This pattern held across speakers and speaking modes. Cho (2011) reported that accent peak occurred earlier in the syllable when it was not phrase-final. Finally, there are also paralinguistic purposes to F0 timing. Hasegawa and Hata (1995) found that native listeners associated a delayed accent peak with a female speaker, and argue that peak delay signaled femininity.

2.1.4 Final accented vs. unaccented words

The contrast between final accented and unaccented words has attracted serious attention from phonologists and phoneticians alike. Recall the example *hana* (unaccented) 'nose' vs. *ha**na*** (final accent) 'flower', in this minimal pair the only difference is the presence or absence of pitch accent, yet when said in isolation they are reported to be identical (Vance, 1995); as McCawley puts it, 'a final-accented phrase… is indistinguishable from an unaccented phrase' (McCawley, 1968, p. 139). In J-ToBI (Venditti, 2005), the two accent conditions are represented as %L H- (unaccented) vs. %L H*(+L) (final accent). However, since the pitch accent falls on an utterance-final syllable, there is no tone-bearing unit for the following +L part to realise, resulting in %L H- vs. %L H*. Their distinction is only obvious when they are followed by at least a mora where the L tone can be realised. Here the question arises as to whether the two accent conditions are actually the same in terms of articulatory targets — if final accented words are perceived as different from unaccented words only if they are followed by a L tone, are they really different, or are they underlyingly the same as unaccented words but became different as a result of the L?

Sugito (1982) investigated whether native Japanese speakers made a distinction between an accent peak and an unaccented counterpart in terms of F0. She recorded 14 subjects producing 2-mora words in isolation, including the commonly adopted minimal pair *hana* 'nose' vs. *ha**na*** 'flower', and only three speakers made an observable distinction between the two classes of words. Moreover, two of these three speakers were news announcers who were used to making accurate articulation. It thus appears that for the average native speaker these two accent conditions are not distinguishable.

Numerous researchers have looked into the issue further subsequently. Vance's (1995) experiment reached a similar conclusion: only one out of four speakers in his experiment made a consistent distinction between the pair *hana* vs. *ha**na***. Warner (1997) explored four-mora and

five-mora words, and found clear acoustical differences between accented and unaccented words, using both normal speech and reiterant speech. However, her stimuli were followed by the copula verb *–da* which allowed accentual fall to realise, thus boosted accent contrast. More recently, Sugiyama (2012) revisited this issue with a larger set of stimuli (20 minimal pairs), looking at both production and perception. She found no significant difference in initial rise or accent peak between the two accent classes when produced in isolation. In her perception test, listeners' performance in the word identification task was at chance level for both male and female voices, showing no preference for either class. She also raised the question whether speakers' inability to distinguish final-accented and unaccented words was due to contrast neutralisation at the end of an utterance (cf. word final consonant devoicing in German). While few would disagree that the particle is a contrast booster, if the contrast is not perceivable without the booster, it can be problematic to assume such a contrast exists.

An alternative to Sugiyama's neutralisation hypothesis would be to argue that accent peak is derived through pre-low raising. This is a local anticipatory tonal process where a High tone immediately before a Low tone rises above its usual pitch target. In other words, the H_1 in the sequence $H_1L_1H_2$ would have a higher phonetic realisation than that in $H_1H_2H_3$, ceteris paribus. We shall return to this issue in §2.4.

2.1.5 Other acoustic cues

There are more cues to prosodic contrasts than F0. For example, intensity and duration ratio are both cues to lexical stress in English (Fry, 1955), whereas in Cantonese creaky voice is an effective cue to Tone 4 (Yu & Lam, 2014); it is thus an interesting question whether Japanese only uses F0 as a cue to pitch accent. For perception, Sugito (1982) reported that intensity did not matter even when the accented mora was devoiced. On the other hand, Neustupný (1966) conducted a production test and analyzed F0 and the intensity of 181 words produced in isolation by four male speakers. From the mixed results he concluded that accent in Japanese was realised by an inconsistent set of interacting features, of which intensity is part.

Neustupný's observation was later disputed in Beckman (1986). She compared the phonetic correlates of accent (both acoustic and perceptual) in English and Japanese, and concluded that for Japanese accent and intensity were not as strongly correlated as Neustupný claimed. She examined the F0, duration, and intensity of six minimal pairs produced by six speakers. The measurements were submitted to t-tests to examine whether certain accent patterns differed from each other in terms of those phonetic correlates. Beckman concluded that the accent patterns did not have significant correlates in peak or average vowel intensity, or in duration.

In her review of the literature, Sugiyama (2012) pointed out that this discrepancy could be due to Beckman's methodology. First, Beckman used t-test rather than ANOVA, resulting in a loss of statistical power; second, the alpha level was set at 0.01, meaning fewer significant

differences would be observed than if α = 0.05. Sugiyama (2012) thus suggested that Beckman (1986) could have come to a different conclusion had her statistical analysis been performed differently. All in all, it seems that although one may not be able to go so far as to claim that there are no other cues to accent than F0, non-F0 cues are at best secondary.

Finally, Sugiyama (2013) investigated if formants are affected by accent condition, based on previous findings that formants cue intended F0 in whisper. Her results showed no correlation between F0 movement and formant frequencies.

2.2 Methodology

A production experiment was conducted to examine the abovementioned issues. The details of the resulting corpus are described in this Section, and will form the basis of the analyses in Sections 2.3, 2.4, 6.2, and Chapter 3.

2.2.1 The stimuli

A total of 33 Japanese words were chosen as stimuli (see **Table 2**). The target words varied in length (1~4 morae), accent condition (unaccented/initial-/medial-/penultimate-/final-accent), and syllable structure (CVCV, CVn, CVV). The accent condition of the target words is based on two website, namely the Online Japanese Accent Dictionary (Hirano et al., 2013; Nakamura et al., 2013) and an online pedagogical dictionary of Japanese pitch accent[9]. The stimuli were checked by a phonetician who is a native speaker of Japanese. The tokens were presented in the unaccented carrier sentence *jiten-ni* ___*-mo nottemasu* 'The word ___ too is found in the dictionary'.

[9] http://accent.u-biq.org/

1-mora	CV		
Unaccented	*ne* ‘price’		
1 (H-L)	***ne*** ‘root’		
2-mora	CV	CVV	CVN
Unaccented	*mane* ‘imitate’	*mai* ‘dance’	
1 (HL-L)	***me****mo* ‘memo’	***mei*** ‘May’	***men*** ‘face’
2 (LH-L)	*mu****ne*** ‘aim’		
3-mora	CV	CVV	CVN
Unaccented	*mimono* ‘ornamental plant’	*mimei* ‘dawn’, *neimo* ‘shoot’	*momen* ‘cotton’
1 (HLL-L)	***me****nami* ‘small wave’	***mei****mu* ‘fog’, ***ni****mei* ‘two people’	***nin****mu* ‘mission’
2 (LHL-L)	*na****na****me* ‘oblique’	*me****mai*** ‘dizzy’	*ni****man*** ‘20,000’
3 (LHH-L)	*mimo****no*** ‘attraction’	*nui****me*** ‘seam’	
4-mora	CVCV	CVV	CVN
Unaccented	*monomane* ‘mimicry’	*meimei* ‘naming’	*nennen* ‘annually’
1 (HLLL-L)		***mu****umin* ‘Moomin’	***nan****nen* ‘what year’
2 (LHLL-L)	*mi****na****mina* ‘everyone’		
3 (LHHL-L)	*nama****na****ma* ‘lively’	*mei****mei*** ‘individual’	*men****men*** ‘everyone’
4 (LHHH-L)	*anoma****ma*** ‘as it is’	*nimai****me*** ‘second piece’	*ninen****me*** ‘second year’

Table 2*. List of stimuli used and corresponding English gloss. For simplicity, tonal representation used in the first column of this Table follows Haraguchi* (2002)*, which comprises only H and L.*

A number of factors were taken into consideration when designing the stimuli. First, past studies of Japanese word prosody often used only bimoraic target words (e.g. Sugiyama, 2012) or words with a range of initial consonants (e.g. Warner, 1997, also recall the effect of segmental perturbation discussed in §2.1.2 above), leaving the possibility that the results could have been different with longer words and words that have similar segments. Stimuli used in the present study cover a wider range of phonological contexts (length, accent condition, and syllable structure). Also, the use of only nasals as initial consonants avoids most of the distortions from segmental perturbation of F0. Second, considering that speech rate is generally fast in Japanese (in the present data mean mora duration is 118 ms for normal speech and 162 ms for slow speech, respectively; see **Table 3**) and can lead to F0 target undershoot, I recorded each target sentence at two speech rates to control for the effect of speed of articulation. Third, though less directly relevant to the present research question, introducing three types of syllable structure (i.e. CV, CVV, CVn) into the stimuli could give us further insights into the shape of F0 contours under different conditions, which may be relevant in future studies on prosody modeling.

Speech rate	Normal					Slow					All
Word length	1	2	3	4	Subtotal	1	2	3	4	Subtotal	
FO	139	121	126	117	123	178	157	153	145	153	138
HS	147	133	131	124	130	193	176	169	157	167	149
KU	124	121	118	112	117	151	150	148	138	145	131
MT	107	98	99	92	97	154	145	144	144	145	121
OY	122	125	126	118	123	197	190	191	176	186	154
RT	106	100	99	94	98	145	142	138	128	135	117
YM	153	146	142	136	141	226	219	209	197	208	174
YT	128	119	117	110	116	169	164	155	147	154	135
Grand Total	128	120	120	113	**118**	177	168	163	154	**162**	**140**

***Table 3**. Average mora duration of each speaker (ms)[10].*

Because only nasal stops were used in syllable-initial positions, some low frequency words had to be included. In light of this, subjects were given time to rehearse and familiarise themselves with the experiment material until they felt comfortable enough to start recording. No F0 patterns peculiar to the less familiar stimuli were observed in subsequent analyses.

2.2.2 Recording procedures

Eight native speakers (four of each sex, mean age 28, s.d. 4.72, see **Table 4**) of Tokyo Japanese from the Greater Tokyo Area (Tokyo, Saitama, Kanagawa, and Chiba) served as subjects. They were living in London at the time of recording. All speakers had been living in the U.K. for less than half a year, except for speaker RT who had been in London for 1.5 years, and speaker OY for five years. No atypical F0 behavior was observed in speaker OY. None of the speakers reported any history of speech, language, or hearing impairment.

Initial	Age	Sex	Born	Grew up	Father from	Mother from	Occupation
YM	25	F	Saitama	Saitama	Kyoto	Saitama	Student
OY	34	M	Tokyo	Tokyo	Tokyo	Kagawa	Student
HS	29	M	Tokyo	Tokyo	Tokyo	Tokushima	Student
KU	31	M	Tokyo	Kanagawa	Tokyo	Kanagawa	Student
RT	31	F	Tokyo	Tokyo	Tokyo	Nagano	Clerk
MT	32	F	Tokyo	Tokyo	Tokyo	Tokyo	Student
YT	21	F	Tokyo	London	Tokyo	Tokyo	Student
FO	21	M	Osaka	Tokyo	Nara	Yamaguchi	Student

***Table 4**. Information of participants in Chapters 2 and 3*

The recording took place in a quiet room in University College London, using a RØDE NT1-A microphone placed approximately 30 cm away from the speakers. Subjects were seated

[10] While not directly related to the present discussion, note the effect of word length on average mora duration. One-way ANOVA shows that this effect is significant $F(3,524) = 8.24$, $p<0.001$. A shorter word (in terms of mora count) has longer mean mora duration than a longer word. Post-hoc Bonferroni tests further reveal that 1-mora vs. 4-mora words, 2-mora vs. 4-mora words, and 3-mora vs. 4-mora words, have significantly different mean mora duration.

in front of a computer screen, on which stimuli were displayed one by one in random order, in standard Japanese orthography (mixed used of *kanji* and *kana* syllabary). They produced each sentence first at normal speed, then immediately followed by a slow production. Though speech rate was not stipulated in actual terms, subjects were instructed to speak obviously slower in the second production. When an undesired emphasis was placed on the particle *–mo* (see discussion on "prominence-lending rise" in Venditti, Maekawa, & Beckman, 2008), in which case *–mo* sees an abrupt F0 rise, the subject was asked to repeat the utterance without any emphasis. From each subject a total of 33 sentences × 5 repetitions × 2 speech rates = 330 tokens were collected. The sampling rate was 44.1 kHz.

Before recording began, participants were briefed about the experiment and granted their written consent to being tested. Speakers were then interviewed about their linguistic background and history of speech and hearing impairment. All speakers were remunerated for their time.

2.2.3 Annotation

Sound files were annotated using ProsodyPro (Xu, 2013), a Praat (Boersma & van Heuven, 2001) script for prosody analysis. Each sound file was labeled (see **Figure 7**), and markings of vocal pulses were manually rectified (to correct apparent errors in the vocal pulse markings generated by the autocorrelation algorithm in Praat that could lead to octave jumps). The identity of the labels per se is meaningless to ProsodyPro, which extracts measurements from any labelled interval. The labels in this corpus were named in such a way to facilitate subsequent automatic generation of functional annotation for F0 synthesis in §6.2. Segmentation was done by the 'mora', such that a light syllable (CV) counts as one mora while a heavy syllable (CVN or CVV) counts as two. In the latter case, two labeled intervals equal in duration were assigned. Apart from the target word itself, the mora before (*-ni*) as well as the one after (*no-,* both part of the carrier sentence) were also labeled during annotation in order to capture any carryover effect extended from or into the target word. Other parts of the carrier sentence were not analyzed in this study. The script then generated all the acoustical measurements from individual files, as well as ensemble files containing data ready for graphical and statistical analyses. Unless otherwise stated, throughout Chapters 2 and 3 F0 values analyzed have been converted into semitones using utterance-initial F0 value as reference[11].

[11] Converting Hz values into semitones based on utterance-initial F0 allows one to adjust for general F0 level difference between normal and slow speech. In the present data, utterances of the same speech rate generally start at a similar F0, whereas slow speech has a lower global F0 level than normal speech.

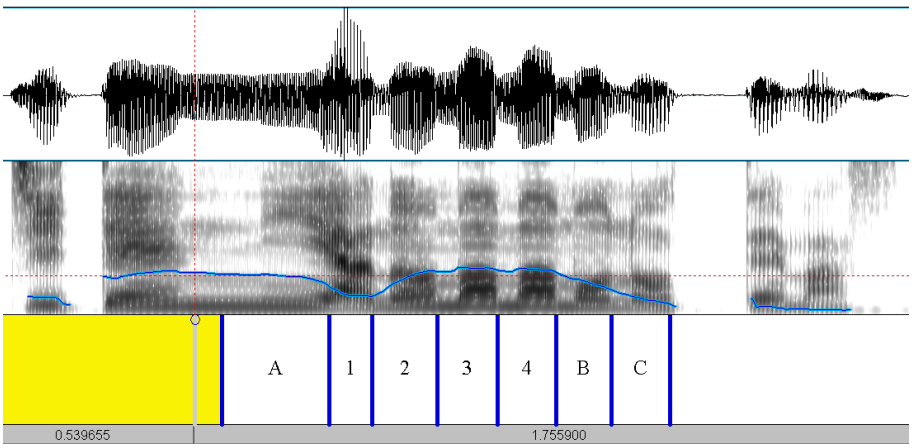


***Figure 7**. A screenshot of annotation using ProsodyPro.*

2.3 General acoustic analysis

In §2.1 we saw that surface F0 is influenced by a wide range of factors. Before I present and test the hypothesis in §2.4, it is necessary to survey in detail the acoustic properties of the corpus. Some of the findings below will be a confirmation of previous studies, while others are new, and the purpose of this Section is to lay the foundation for subsequent analysis in §2.4.

2.3.1 F0 scaling

First I measured accent peak F0, defined as the maximum F0 value in the accent host mora as well as the mora that immediately follows. Here I introduce a new independent variable, peak-to-end distance. It is the distance (number of morae) between pitch accent and word end. For example, for the word *mi**na**mina* in **Table 2** (Page 29), peak-to-end distance is three morae because accent falls on the second mora and the word ends after the fourth mora.

Figure 8 shows time-normalised F0 contours of 3-mora accented words averaged across 40 repetitions from eight speakers. In this diagram, ***me**nami* (initial accent) has a higher and later peak than other words, whereas all accented words have a higher peak than their unaccented counterpart. It also appears that when the pitch accent occurs early in a word, the word ends in a lower F0. Below I will examine if these trends are observed across word length conditions, and the effects of accent condition, word length, and peak-to-end distance, all of which are related to the others.

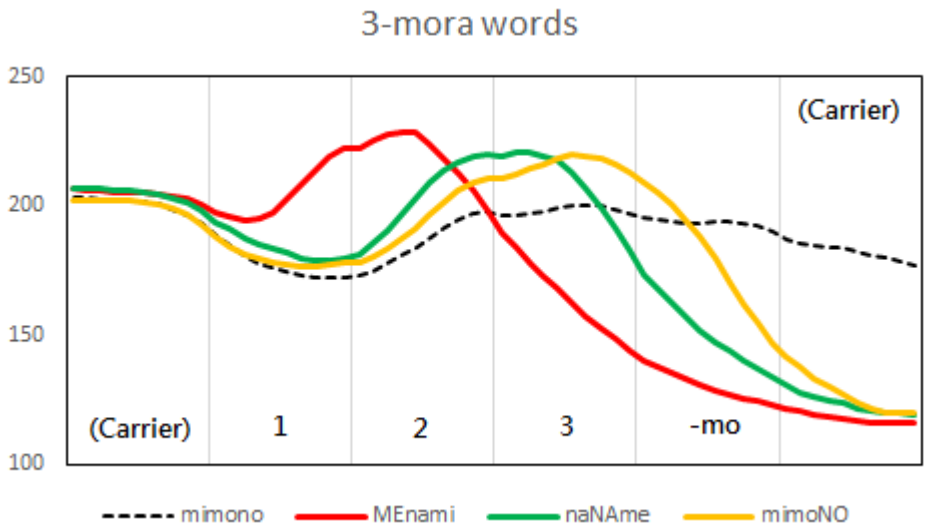


***Figure 8**. Averaged F0 contours of 3-mora CVCV accented words. For this diagram, the capitalised syllable bears pitch accent. X-axis shows normalised time, whereas Y-axis is F0 in Hz.*

Visual inspection of **Figure 9** suggests that accent peak is realised lower when it occurs later in time. The top left panel shows that an early accent (e.g. Condition "1") has a higher accent peak. One-way ANOVA also shows that accent condition (i.e. accent on the first/second/third/fourth mora, $F(3,364) = 38.76$, $p<0.001$) has a significant main effect on accent peak F0. Post-hoc Bonferroni tests further reveal that an early accent has a higher peak F0 than the other accent conditions, with initial accent having a significantly ($p < 0.001$) higher peak F0 than all other accent conditions.

One possible explanation for this is background declination, i.e. F0 gradually decreases over time. The function/source of declination is not well-understood, though some (see Sugahara, 2003b and references therein cited) suggest that it is in part due to decreasing subglottal air pressure. In other studies, the falling of F0 over time is also attributed to communicative functions, for example topic (Wang & Xu, 2011) and focus (Pierrehumbert & Beckman, 1988); but whether the gradual fall in F0 observed here is related to any communicative function requires further investigation. In any case, my data clearly show a lowering accent peak F0 as a function of time. Interestingly, despite the clear pattern in **Figure 9**, in the post-hoc Bonferroni tests reported above, significant difference in peak F0 was observed only in contrasts that involved initial accent, while other accent conditions were not significantly different from one another. To verify that accent condition affects peak F0, I performed an additional one-way ANOVA on a subset of 3-mora words (leaving only accent condition to vary), and the main effect of accent condition on peak F0 remained highly significant, $F(2,128) = 15.180$, $p<0.001$, showing that accent condition robustly affects scaling.

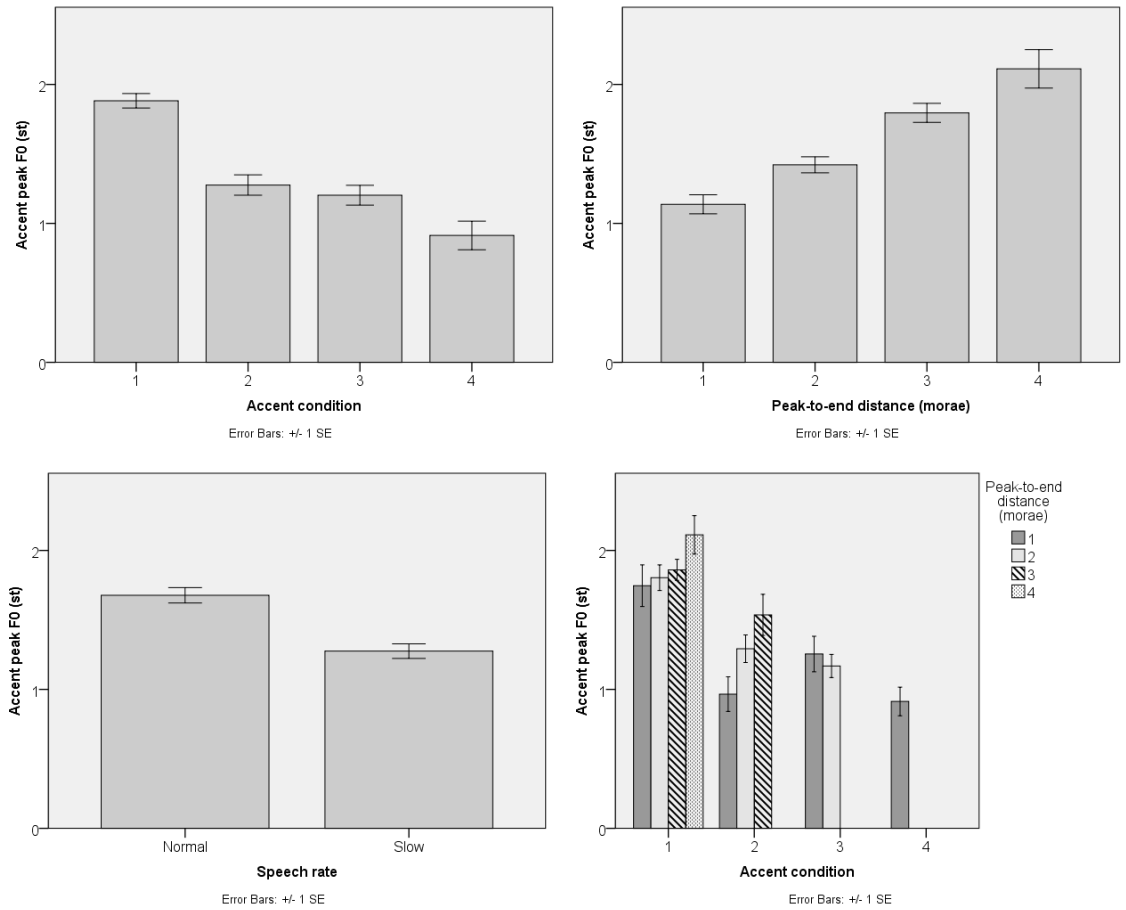


***Figure 9**. Boxplots showing accent peak F0 under various conditions*

There is also an obvious trend that a greater peak-to-end distance coincides with a higher accent peak. One-way ANOVA shows that peak-to-end distance (i.e. one to four morae) has a significant main effect on peak F0, $F(3,364) = 23.79$, $p<0.001$. Post-hoc Bonferroni tests confirm this observation, and show that a greater peak-to-end distance is always associated with a higher peak F0 in the data ($p < 0.05$ in all contrasts). A possible interpretation is that peak-to-end distance is actually reflecting the effect of accent condition; an accent that is far away from word end is also an early accent. To verify this, I looked at the subset of initial accented words (leaving only word length, i.e. peak-to-end distance to vary). **Figure 9** (bottom right panel) shows that a greater peak-to-end distance is still associated with higher peak F0 in this subset (see the first four bars from the left). However, although the general trend holds, there is considerable overlap among categories; indeed ANOVA revealed that the effect of peak-to-end distance in this particular subset was non-significant.

The effect of word length on accent peak F0 is significant too, albeit small, $F(3,364) = 3.265$, $p<0.001$ (one-way ANOVA). As the three factors mentioned above (accent condition, peak-to-end distance, word length) are closely related to one another, the significant effect in one may be merely reflecting that in another. Here a natural question would be: which of these factors is a better predictor of the variations in accent peak F0. To this end, the residual sums of squares from the above one-way ANOVAS were compared: accent condition = 158.96, peak-to-

end distance = 175.35, word length = 204.243. With the most variation in the data left unexplained, word length thus appears to be the worst predictor of the three. This issue will be further discussed in §2.4 and §6.2.

Speech rate (**Table 3**) affects accent peak as well. **Figure 9** (bottom left panel) shows that accent peak is lower in slow speech. One-way ANOVA shows a significant main effect of speech rate on accent peak F0, $F(1,366) = 27.86$, $p<0.001$. Post-hoc Bonferroni test shows that accent peak in slow speech is on average 0.40 semitones lower than in normal speech. Here one may question whether normalisation may have distorted the results; that is, the utterance-initial F0 value being used for normalisation is lower in slow speech. However, holding other factors constant, a smaller reference value in normalisation should result in a greater accent peak, which is contrary to the present results. Thus it is safe to conclude that accent peak is lower in slow speech than in normal speech. (See also Appendix 5 for averaged F0 contours of normal vs. slow utterances)

Next I consider the scaling of the L tone in accented words. Specifically I am interested in the excursion size of accentual fall under various conditions. **Figure 10** (top left panel) suggests a clear trend; when the accent peak is further away from word end, giving a speaker more time to decrease F0, the L tone is realised at a lower F0. In turn, this also means that a greater peak-to-end distance is associated with a larger accentual fall (top right panel). One-way ANOVA reveals that peak-to-end distance has a significant main effect ($F(3,364) = 17.10$, $p<0.001$) on the minimum F0 of an accented word (in semitones). Results from post-hoc Bonferroni tests show that final-accented words (where peak-to-end distance is 1 mora, including monomoraic accented words) end in a significantly higher minimum F0 ($p<0.001$ in all contrasts) than medial and initial accented words. This trend holds among initial accented and medial accented words (i.e. among words of which peak-to-end distance is not 1 mora) too, although the differences are not significant. On the other hand, it is not clear (bottom panel) whether peak-to-end distance has any effect on the maximum F0 velocity (i.e. rate of F0 change, measured as semitones per second) during accentual fall.

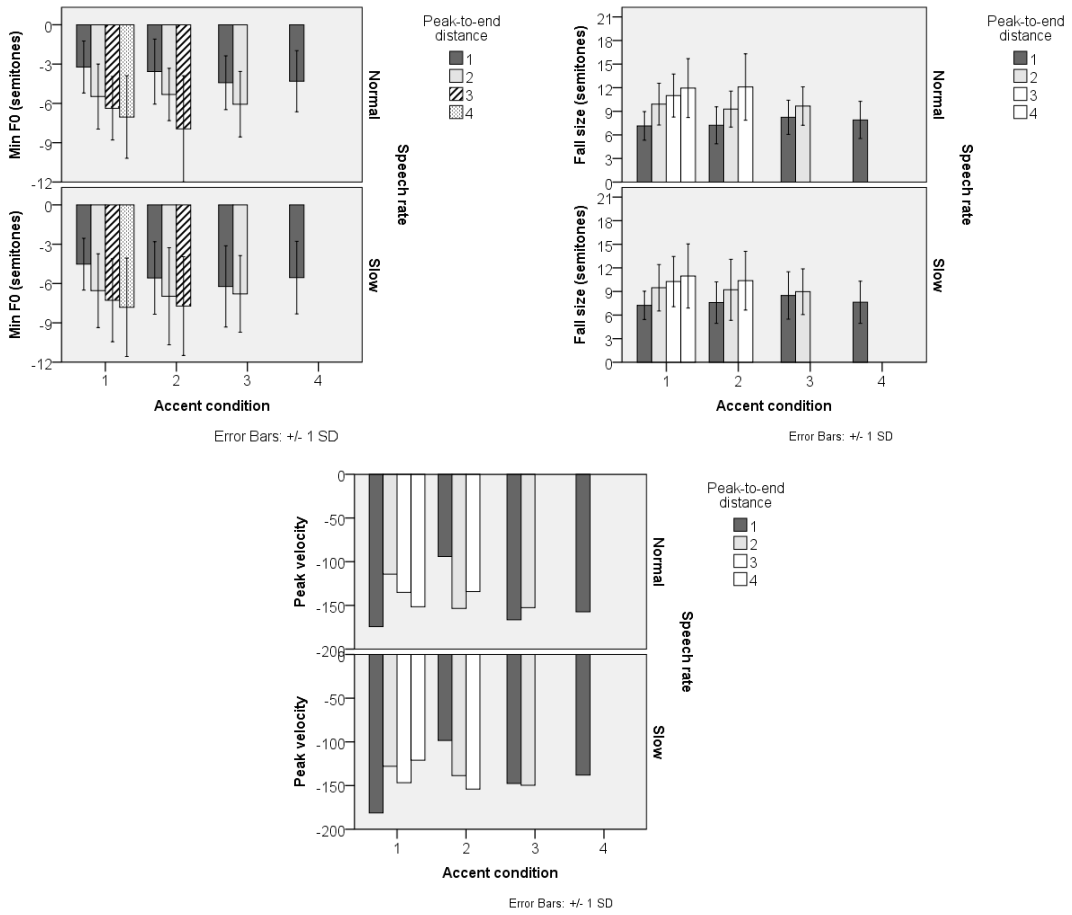


***Figure 10**. Effects of accent condition, speech rate, and peak-to-end distance on the F0 of (i) minimum F0 of prosodic word, (ii) excursion size of accentual fall, and (iii) minimum F0 velocity in accentual fall.*

Above I have demonstrated that accent peak, i.e. what is known as H* in J-ToBI, has a variable height which is largely influenced by where it occurs. While it is unknown whether declination is playing any functional role in the stimuli, my experiment design that controls speech rate, accent condition, word length, and syllable structure has enabled us to see exactly how time and peak height interact. The relation between time and accent peak will become more relevant to the research question in the next Section where peak alignment is discussed.

2.3.2 Peak alignment

In PENTA, the timing at which acoustical landmarks occur is seen as a result of target approximation; hence the concept of segmental anchoring or alignment is not given a great deal of importance. That said, F0 peak timing can be useful as it indirectly informs us about the underlying articulatory target, as well as factors like articulatory strength. Below I consider the timing of accent peak in the data, using peak delay ratio as measurement:

$$\textbf{Peak delay ratio} = \frac{\textbf{Actual time of accent peak – Actual time of host mora onset}}{\textbf{Duration of host mora}}$$

The resulting quotient, if >1, will mean that accent peak occurs after the offset of the host mora, whereas if it is <1 it occurs within the host mora. I measured Pearson's *r* (N = 1840[12]) by comparing peak delay ratio and several variables, defined below in **Table 5**, and a number of strong correlations emerged in **Table 6**:

F0Max	Accent peak. Maximum F0 value in the host mora (excluding the first 2 time points)[13] of pitch accent and the mora that follows
F0Mean	Mean F0 of the accent host mora
F0A	F0 at the onset of accent host mora
F0B	F0 at the offset of the mora that immediately follows the accent host
F0C	F0 at the offset of the prosodic word, i.e. the final particle *-mo* in the carrier sentence
F0Rise	F0Max less F0A
F0Fall1	F0B less F0Max
F0Fall2	F0C less F0Max
V1	Final F0 velocity of the accent host mora.

Table 5**. Variables used in **Table 6

In particular, a late peak is related to a high F0B (*r* = 0.740), high V1 (*r* = 0.794), and a small F0Fall1 (*r* = 0.690, note that *r* is positive because F0Fall1 = F0B - F0Max). F0B is defined as the F0 at the offset of the first post-accent mora, thus a late peak would mean there is not enough time for F0 to drop sufficiently from the peak, resulting in an undershot (less low) F0B. A high V1 (final F0 velocity of accent host mora) is related to a late peak due to physiological constraints on sudden changes in pitch direction (i.e. 'inertia'). In other words, if F0 is still sharply rising at the offset of accent host mora, it is impossible for it to start falling sharply at the turn of a new mora. These two observations stem from time constraints on articulation (Cheng, 2012; Xu, 2001). Likewise, F0Rise and F0Max are positively related to peak delay ratio because it takes time for F0 to rise to a high peak, leading to peak delay (see also **Figure 8** in Page 33 where an early accented-word is associated with both a later and higher peak). These observations can be easily accounted for with articulatory explanations, without assuming that peak delay is phonologically specified.

	Correlations (N = 1840)									
PDR		F0Max	F0Mean	F0B	F0C	V1	F0A	F0Rise	F0Fall1	F0Fall2
	Pearson *r*	0.414	-0.472	0.74	-0.215	0.794	-0.39	0.597	0.69	-0.328
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001

***Table 6**. Pearson's correlations of peak delay ratio (PDR) vs. numerous variables.*

[12] That is, with data unaccented words excluded.
[13] This is to avoid the syllable-initial high F0 values due to consonantal perturbation.

Note that the negative correlations with F0C (r = -0.215) and F0Fall2 (r = -0.328) require a different explanation. Target approximation would predict that a late peak leaves less time for reaching a target, leading to an undershot (higher) F0C, i.e. r>0. Nevertheless, here r<0 means that when F0 needs to reach F0C within a shorter time (due to late peak), a lower F0C tends to be reached. This is because F0C measured at word end, which in some cases is several morae away from the accent peak. Since time pressure is not at issue here, this correlation should be taken to mean that a greater peak delay leads to a larger accentual fall. I verified these results by performing the same test on averaged data (across 5 repetitions of each speaker, N = 368), and obtained the same conclusion, peak delay ratio~F0C r = -0.279, and F0Fall2 r = -0.400. A negative correlation was consistently observed across all individual speakers between peak delay ratio~F0C (ranging from r = -0.174 to r = -0.428). The same consistency is found across syllable structures that for peak delay ratio~F0Fall2, r = -0.241 for CVCV, r = -0.302 for CVn, r = -0.381 for CVV. Recall that unaccented data have been removed for this analysis, hence any negative correlation is not a result of bimodal distribution from the two accent classes. This issue will be taken further in Chapter 3.

Next I investigate how speech rate, syllable structure, accent condition, and peak-to-end distance may affect peak timing. Here, to avoid excessively low p-value I use data averaged across speakers (N = 368). Word length is not considered here given the observation in §2.3.1 that it is not a good predictor of F0 scaling. Several one-way ANOVAs were conducted to examine their effects on peak delay ratio. The effect of speech rate on peak delay ratio was non-significant, perhaps because of longer mora duration, i.e. large denominator of the ratio. I further ran an one-way ANOVA on actual peak delay (ms), and found a significant main effect of speech rate as well ($F(1,366) = 26.89$, $p<0.001$); in general, accent peaks in slow speech is 30 ms later than normal speech ($p < 0.001$). Regarding peak-to-end distance ($F(3,364) = 72.25$, $p<0.001$), **Figure 11** shows a clear trend — where word length is held constant, the earlier an accent (phonologically) the more peak delay. For example, for an initial-accented 4-mora word (rightmost solid grey bar in **Figure 11**), the accent is 4 morae away from word end, and the accent peak occurs the latest; in contrast, if pitch accent falls on the final mora, which is right before word end, accent peak occurs relatively early (cf. first four bars from the left in **Figure 11**). Post-hoc Bonferroni tests confirmed that all the contrasts were statistically significant ($p<0.001$), with the exception of peak-to-end distance conditions 3 vs. 4 (p = 0.30). A comparable phenomenon has been reported in Greek wh-questions (Arvaniti & Ladd, 2009), where the variability in peak timing is influenced by proximity to other tones (e.g. boundary tones); peak timing is thus considered a phonological factor in their account. From an articulatory viewpoint, it could be argued that because a Japanese word usually starts with an initial rise across the first two morae, an early accent which coincides with the initial rise will have a larger F0Rise than a late accented word, the latter of which F0Max would instead be preceded by a high plateau (see **Figure 8**, Page 33). In other words, reaching the accent peak from an already high plateau means a less steep rise, and in turn leads to less peak delay.

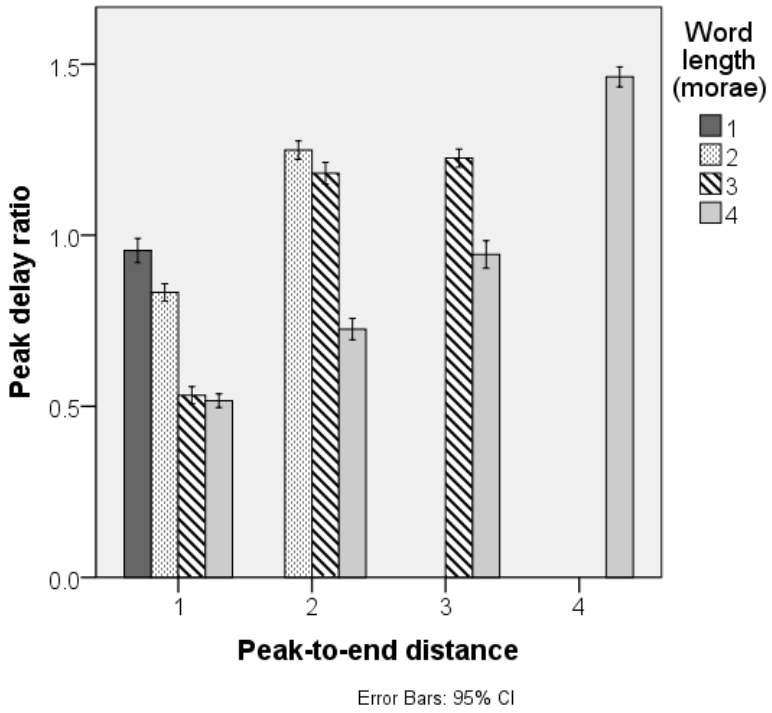


***Figure 11**. Barplot showing mean peak delay ratio under different peak-to-end distance and word length conditions.*

So far the results of accent peak F0 and peak delay ratio tie in well. An early accent has a higher accent peak, as does normal speech. In other words, the later in time a pitch accent occurs the lower the accent peak is. Here declination is not the only possible explanation, the fact that Japanese has initial rise forces an early accent to undergo a sharp rise, leading to a higher velocity (V1) and in turn higher F0Max (due to inertia). At the same time, because it is difficult to turn from a sharp rise to a steep fall, an early accent is associated with a greater peak delay ratio. On the other hand, a late accent is preceded by a high plateau, from which it is easier to turn to a fall, hence a smaller peak delay (see **Figure 8**).

2.3.3 Formant analysis

Since the effect of accent on formant frequency is understudied, I analysed the formant trajectories of the following pairs of words in the data: *ne* vs. ***ne***, *mimei* vs. ***ni****mei*, *mimono* vs. *mimo****no****, meimei vs. mei****me****i,* and *nennen vs. **na**nnen.* The contrast in the first, the third, and the fourth pairs is minimal, differing only in accent condition; whereas the second and the fifth pairs are near-minimal. The annotation of these 10 words (10 words * 5 repetitions * 8 speakers * 2

speech rates = 800 utterances) was fed into FormantPro[14] (also described in Cheng & Xu, 2013; Xu, 2007), to extract formant values for graphical and statistical analyses. The script uses the Burg algorithm implemented in Praat to extract continuous formants, and applied a trimming algorithm to remove excessive and sudden bumps in the formant trajectories (Cheng & Xu, 2013).

Visual inspection of **Figure 12** shows that for the chosen word pairs, accent condition has no effect on formant frequencies; the respective formant trajectories of accented and unaccented words overlap with each other. Subsequent t-tests comparing mean formant frequency of relevant morae did not reach statistical significance either. These results agree with Sugiyama and Moriyama (2013) who found no correlation between F0 and formant frequency. Although Higashikawa and colleagues (1996) reported significant effect of intended pitch on formant frequency in whisper, such effect does not apply to normal Japanese speech.

[14] Xu, Y. (2007-2015). FormantPro.praat. Available from: http://www.phon.ucl.ac.uk/home/yi/FormantPro/

Normal speaking rate Fast speaking rate

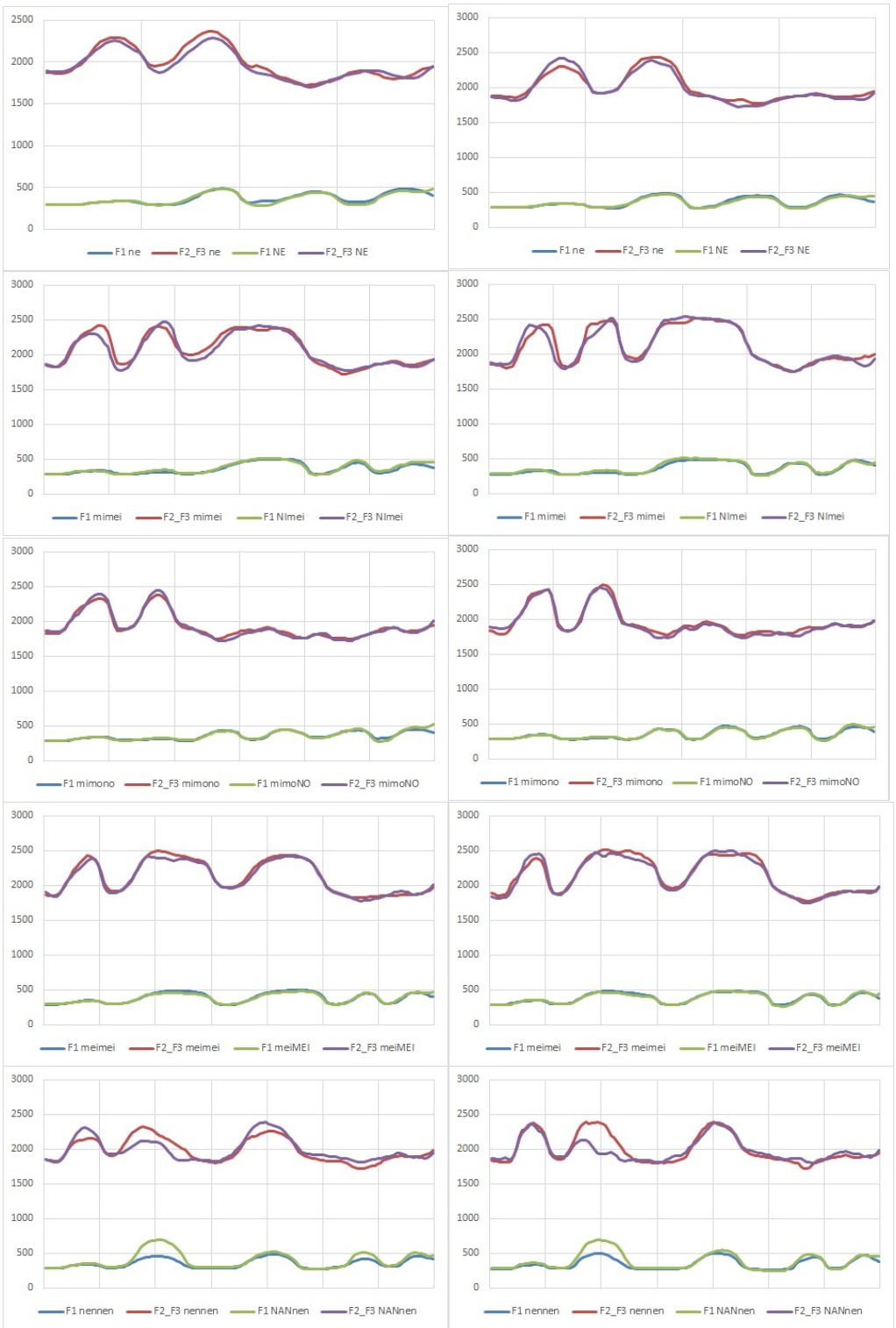


Figure 12*. Formant contours averaged across 40 repetitions from eight speakers. X-axis shows normalised time, with vertical lines representing mora boundaries; Y-axis shows formant frequency in Hz.*

2.3.4 Discussion

In the present Section, we have seen that the scaling of accent peaks is influenced by a complex interplay between accent condition and peak-to-end distance (or word length). Both an early accent and a greater peak-to-end distance are associated with a higher accent peak F0. Identifying the source of variation in accent peak F0 is important, as it the paves way for the discussion in §2.4, when the difference between accented and unaccented words is considered.

The effect of accent condition has been well known, and is reflected in Beckman and Pierrehumbert's synthesiser (1986b) where they stipulated a gradually declining topline; the later an accent peak occurs, the lower a topline it will be bound by. The effect of word length is expected too, given Selkirk and colleagues' (2004) findings. But the two effects could also be interpreted differently, as the distance between accent peak and word end, which also showed a significant effect on accent peak F0.

As for the mechanism behind these effects, a possible account is related to final lowering (Cooper & Sorensen, 1977). Cooper and Sorenson argue that the final lowering which marks the end of a phrase could be due to 'a generalised relaxation response of the speech-processing machinery as it nears completion of the processing of a constituent' (1977, p. 691, i.e. end of phrase). When such relaxation slows down vocal fold vibration, F0 is lowered. The following rise that marks the beginning of a new phrase, then, is the reverse, i.e. a new round of excitation. Following this logic, the effect of word length on scaling would be a result of planning based on word length. Anticipating a longer word, one needs a larger operating pitch range for subsequent falling movement, thus a higher peak. This is also related to the fact that a smaller peak-to-end distance sees lower peak — there is less time for the fall, thus requiring a smaller pitch range and lower peak.

Another way of looking at the effect of accent condition is that accent peak is subject to declination (cf. topline in Beckman and Pierrehumbert's synthesiser), which is argued by some to be due to gradual drop in subglottal air pressure ('t Hart, Collier, & Cohen, 1990). Ohala (1978), however, argues that declination (or downdrift) is not an automatic mechanism because the rate of drop in subglottal air pressure is never as large as the decline in F0, and that it is often not found in question intonation. He instead suggests that F0 declination is the result of active laryngeally caused changes in vocal fold tension. For now, we will move on with the conclusion that accent peak height is primarily determined by accent condition, complemented by peak-to-end distance, as shown in §2.3.1.

The observation that a late accent (e.g. final accent) has earlier F0 peak (relative to the onset of accent host mora) in part echoes with accounts of tonal crowding, as briefly mentioned in §2.3.2. In many languages the timing of a given tone tends to be (pushed) earlier phrase-finally, when it is adjacent to a boundary tone; the early timing of late accent in the data is thus reminiscent of this phenomenon. However, the gradient nature of peak timing revealed by the present corpus, illustrated in **Figure 11** and in the strong correlation between peak delay ratio and F0Max, requires an alternative explanation. Through controlling for word length and accent

condition, there is a gradient pattern of peak timing in relation to where an accent falls in a word — a late peak in an early accent, and vice versa. This is thus more than a binary context of whether there is crowding or otherwise. Putting this into context, it appears that the observed peak timing pattern is a result of accent peak height. First, the strong correlation between F0 and peak timing suggests these two are closely related. Second, as discussed above the scaling of accent peaks is articulatorily determined, one way or another. It is thus not ideal for any framework to have to specify peak timing for each word length × accent condition combination to account for the gradience, as opposed to leaving timing a free parameter to be derived from other parameters like F0 and articulatory effort. The results in §2.3.2 thus tie in well with target approximation where F0 alignment is not specified in the input.

There is still one issue mentioned in §2.1.2 that is left unresolved — the interpolation between H- and H*. As the present analysis has only covered the accent peak, this issue will be discussed in §6.2 when F0 modelling results are considered.

2.4 Pitch accent as pre-low raising[15]

2.4.1 Background

This Section argues that there is only one underlying articulatory target for the High tone in Japanese. As reviewed in 2.1, the distinction between final-accented and unaccented words can only be produced and perceived when there is a following syllable for the L tone to realise; in isolation their difference does not usually stand out. Even in studies where this distinction was found in the surface F0, it was only produced by some of the speakers. Although it is known that the following L tone is salient in the perception of accent condition, it is unclear why in production the surface F0 distinction is observed only in the presence of L. Here I attempt to resolve this puzzle by testing a new hypothesis: there is only one H tone in Japanese and the acoustic difference between the two versions is due to a well attested mechanism, namely, pre-low raising of surface F0, or PLR in short. In other words, this H tone underlies both accented and unaccented words with PLR and other factors giving rise to surface F0 variations.

Also known as F0 polarisation (Hyman & Schuh, 1974), anticipatory dissimilation (Gandour, Potisuk, & Dechongkit, 1994; Xu, 1997), regressive H-raising or anticipatory raising (Connell & Ladd, 1990; Laniran, 1992; Xu, 1999), PLR is a local anticipatory tonal variation where the F0 of a High tone becomes higher when preceding a Low tone. For example, the F0 of first H in the sequence HLH would be higher than in HHH, ceteris paribus. PLR has been reported for other languages, including Bimoba (Snider, 1998), Cantonese (Gu & Lee, 2007), Gurma (Rialland, 1981), and Igbo (Laniran & Gerfen, 1997). And it has also been observed in singing, which is referred to as 'preparation' (Saitou, Unoki, & Akagi, 2005). Though widely

[15] A version of this Section was reported in Lee et al (2013).

observed, the underlying mechanism of PLR is still unclear despite some speculations (Gandour et al., 1994; Sugiyama, 2012; Warner, 1997; Xu, 1997). Moreover, the precise condition that triggers PLR is reported to vary from one language to another. For example, in Yoruba PLR is observed when a high tone is followed by a low tone (Laniran & Clements, 2003), in Cantonese it appears to occur mainly in rising tones (Gu & Lee, 2007), while in Mandarin the rising tone, the low tone and the falling tone can all trigger PLR in a preceding tone (Xu, 1997). What is common to all these cases is that the trigger contains a low pitch point, and the preceding tone has a high pitch point. The Japanese case seems to satisfy this condition, as can be seen in **Figure 13**. However, whereas in previous studies PLR is established by showing the surface F0 realisation of the High tone under different contexts, this approach is not applicable to Japanese. This is because the high F0 points in the two curves come from two different accent conditions, attributing the higher F0 of the solid line to PLR could be partially circular; showing the difference of two (arguably) different entities does not prove the existence of PLR for Japanese. One way to reduce the level of circularity is to show that the effect of PLR is gradient rather than all-or-none, because the gradience would be incompatible with the two-H-tone hypothesis, but would be more compatible with a biomechanical account.

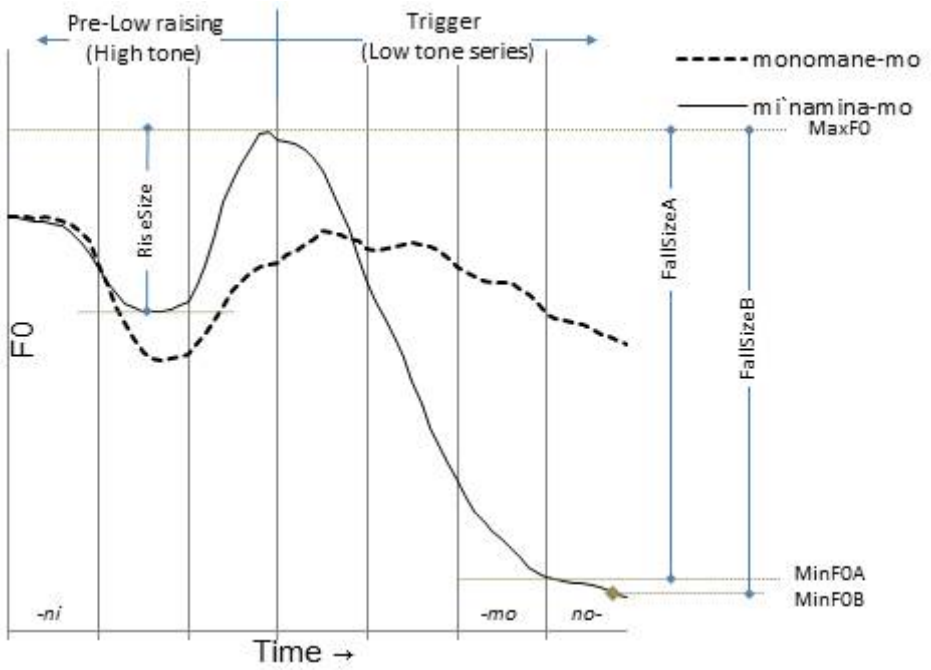


***Figure 13**. An accented word and an unaccented word. F0 contours of monomane (unaccented) 'mimicking' and mi**na**mina (accented) 'everyone' averaged across 40 repetitions from eight speakers.*

One such account, based mainly on the physics of motion, is that PLR is an anticipatory action to increase the peak velocity of F0 movement in order to reach a low F0, given that the production of a low F0 is harder than that of a high F0 since it involves external laryngeal muscles (Atkinson, 1978; Erickson, 1976). Other things being equal, reaching a target quickly requires a high velocity, but peak velocity is positively related to movement magnitude, whether

the movement is articulatorily (Ostry & Munhall, 1985) or acoustically measured (Cheng & Xu, 2013; Xu & Sun, 2002). The present data seem to fit this account. In Xu and Sun (2002), the average speaker could lower F0 as fast as 61~67 semitones per second; here, mean mora duration is 118 ms for normal rate, while mean excursion size of accentual fall is 9.35 semitones (**Figure 10**, top right panel). Thus for cases where peak-to-end distance is only one mora (e.g. final accented words), the speaker would be intending to drop F0 by 9.35 semitones in some 118 ms, or 79 semitones per second, which exceeds Xu and Sun's (2002) reported maximum speed of F0 change. By implication, when producing an accentual fall Japanese speakers may well be at their maximum speed of F0 change. As such, it would be helpful to pre-raise F0 to increase the total movement distance in order to achieve the peak velocity needed to reach a low F0. Such a pre-movement can be also seen in the act of throwing an object or striking a tennis ball with a racket: the harder the throwing or striking force is required, the farther the arm needs to first pull back in the opposite direction.

Another account is based on more specific physiological mechanisms of F0 production. F0 is determined by the tension of the vocal folds, which is controlled by a combination of the intrinsic laryngeal muscles, mainly the cricothyroids (CT), and the extrinsic laryngeal muscles — mainly the sternohyoids (SH) and thyrohyoids (TH) (Atkinson, 1978; Erickson, 1976). When F0 changes from either a high or low level across the mid range, there is a 'switch-over point' in the mid-F0 range around which there is a slight overlap of CT and SH/TH activities. When F0 changes from a low to a non-low range, a post-low bouncing effect is observed in Mandarin (Chen & Xu, 2006) and Cantonese (Gu & Lee, 2007), and possibly in English (Pierrehumbert, 1980, as interpreted by Chen & Xu, 2006). This bouncing effect was argued to be due to a temporal loss of balance between the CT and SH/TH control over the vocal folds during the 'switch over' (Prom-on, Liu, & Xu, 2012). Thus it is possible that the driving force of PLR is a preemptive CT activity to pre-balance the SH/TH activity in anticipation of a very low F0.

Although these two accounts are quite different, positing that the raised H is to either facilitate the downward F0 movement, or counteract the contraction of the external laryngeal muscles, both would predict a gradient negative correlation between the F0 values of H and L, as opposed to a categorical dual-height division. Furthermore, assuming that articulatory planning cannot be fully precise, it is also predicted that the negative correlation is weaker than the highly linear negative correlation in the case of post-low bouncing, a related articulatory phenomenon where F0 is boosted to a high level after a Low target (Prom-on et al., 2012). The result of testing these two predictions will therefore either support or reject the biomechanical accounts, thus providing evidence either for or against the single-H hypothesis for Japanese pitch accents. If supported, the single-H hypothesis for Japanese would provide yet another piece of evidence for PLR as a general articulatory mechanism applicable to languages in general, whenever a low tonal target occurs.

2.4.2 Methodology

The corpus described in §2.2 is used for the present analysis. My goal is to examine whether accent peak F0 is gradient and is determined by the following low target. Linear regression will be performed to examine the relationship between these two tone targets. For each utterance in the corpus, the following measurements were taken:

MaxF0	Maximum F0 (in semitones, same for variables ii-vi) in the host mora of a pitch accent and the following mora, wherever it occurs.
MinF0A	Minimum F0 of the final mora of the target prosodic word, i.e. the final particle –mo in the carrier sentence.
MinF0B	Minimum F0 value of the mora immediately after the target word, i.e. no- in the carrier sentence, with the last 30 ms of the mora excluded (to avoid the effect of segmental perturbation on F0 from the following geminate consonant /tt/).
RiseSize	The difference between MaxF0 and minimum F0 of all morae that precede the accent peak.
FallSizeA	The difference between MaxF0 and MinF0A.
FallSizeB	MaxF0 less MinF0B.
VMaxRise	Maximum velocity of initial rise.
VMaxFall	Maximum velocity of accentual fall
PeakDelay	The difference between accent host onset time and accent peak time.

***Table 7**. Variables used in **Table 8**.*

2.4.3 Results

The main hypothesis is that the height of H is a function of the following L, as opposed to previous report of H as a function of word/phrase length (Selkirk et al., 2004). An examination of the properties of unaccented words in the data is thus necessary before proceeding to examine the behavior of H, to check if word length is a confounding factor on F0. **Figure 14** shows that contrary to the observation of Selkirk et al (2004), it is only for some, but not all, speakers that a longer unaccented word has a higher maximum F0, echoing with the results in §2.3.1 that the word length was the worst predictor of accent peak F0. A one-way ANOVA was performed to examined the effect of word length (one~four morae) on the maximum F0 of unaccented words. Although there was a significant main effect of word length $F(3,156) = 4.486$, $p = 0.005$, post-hoc Bonferroni tests revealed that most contrasts did not reach statistical significance ($p > 0.05$). The one contrast that was significant was one-mora vs. four-mora words. This is also in line with our observation in Lee et al (2014) and in §6.1 below that adding word length as an extra predictor did not yield noticeable improvement in modeling accuracy — if word length systematically influences F0 in the data, having it as an additional predictor should have allowed the model to capture more variations, which is not the case. Having established the inconsistency of word length effect on the height of unaccented F0 peak across speakers in the data, next we proceed to investigate if H is gradient.

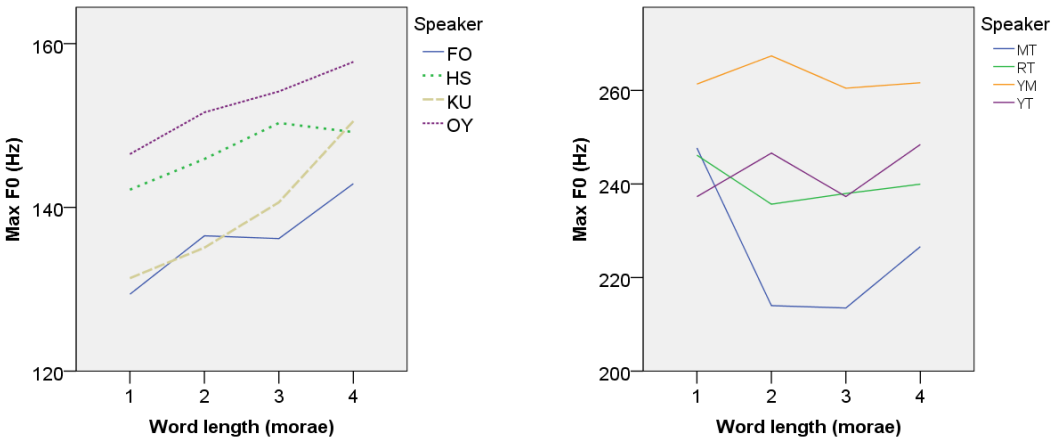


***Figure 14**. Mean Max F0 of unaccented words by word length (morae).*

Figure 15 displays time-normalised F0 contours averaged across five repetitions by all subjects. One can see that peak-to-end distance is positively related to accent peak height, but inversely related to the F0 of the right edge of target word. That is, other things being equal, the earlier the pitch accent in a word, the higher its peak F0 and the lower the F0 at word end. **Figure 16** shows how peak-to-end distance affects MaxF0. The four bars on the left represent words that bear initial accent, and differ from one another in terms of peak-to-end distance. For example, the leftmost bar represents initial accented-words of which accent peak is one mora away from word end, i.e. it is a one-mora word. MaxF0 is higher when peak-to-end distance is greater (e.g. the fourth bar from the left). This echoes with the one-way ANOVA result in §2.3.1, where peak-to-end distance was found to have a significant effect on accent peak F0, $F(3,364) = 23.79$, $p<0.001$.

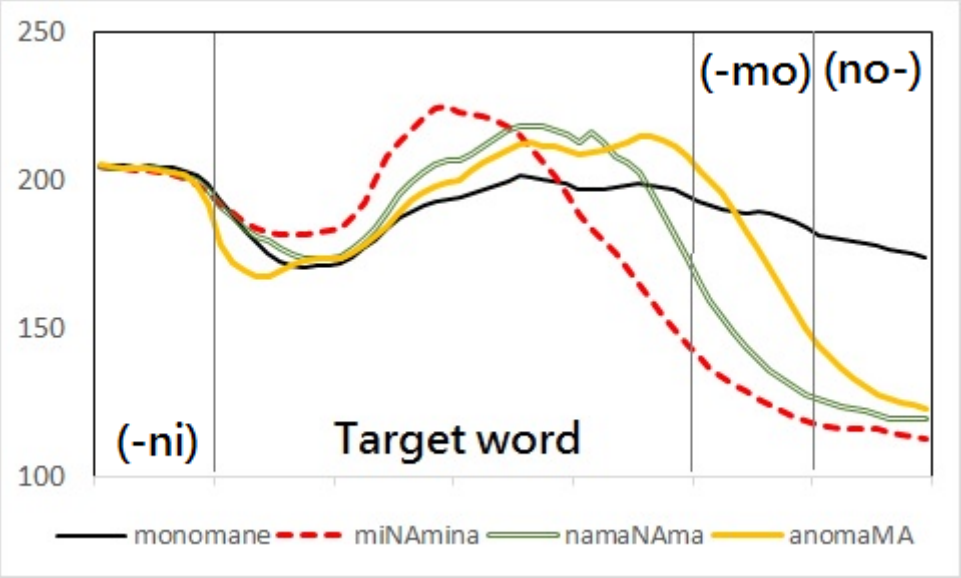


***Figure 15**. Time-normalised average F0 contour of four 4-mora words. X-axis shows normalised time, while Y-axis shows F0 in Hz. The four intervals are parts of the carrier sentence, the target word, and the particle –mo.*

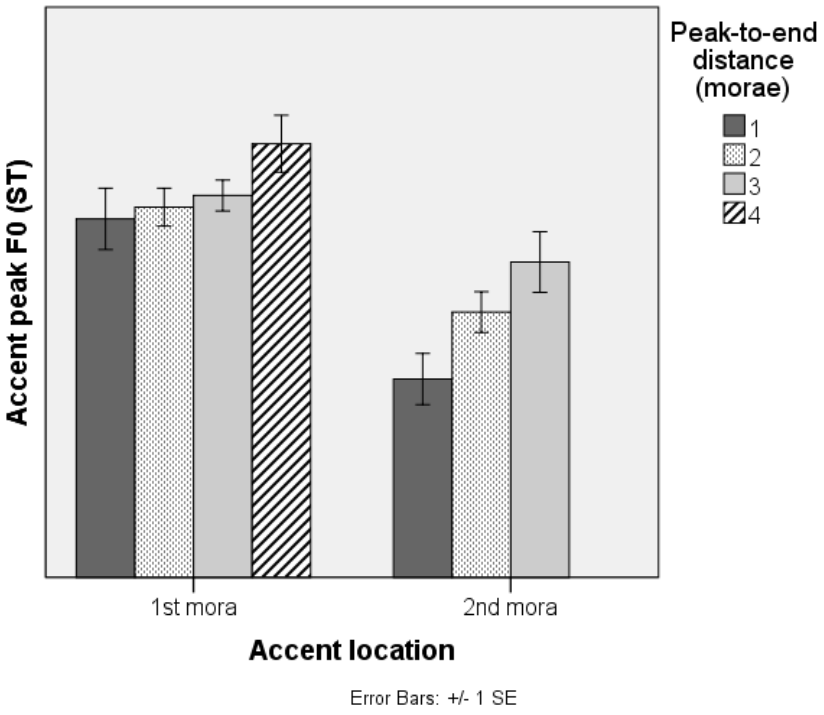


***Figure 16**. The effect of peak-to-end distance (number of morae between the accented mora and word end) on MaxF0 in semitones (y-axis). Bar colors represent the length of peak-to-end distance.*

Next recall the behavior of the L target (word end minimum F0) reported in §2.3.1. A significant main effect of peak-to-end distance on minimum F0 was observed, $F(3,364) = 17.10$, $p<0.001$. Final-accented words (where peak-to-end distance is 1 mora) were found to end in a significantly higher F0 than medial and initial accented words. The effect of peak-to-end

distance on accent peak F0 and minimum F0 thus appears to support the hypothesis that the realisation of H and L is an inverse relationship.

I then compared the measurements given above for possible correlations (N = 2640, including unaccented words). The F0 data were first converted to semitones using the utterance-initial F0 of each utterance as reference. The reason for doing so, rather than using another reference like speaker mean F0, was to avoid distortion from the lower F0 register in slow speech rate observed across all speakers (also Footnote 11). MaxF0 and MinF0A were inversely correlated, r = -0.354 (two-tailed, $p<0.001$), suggesting that a lower word-end F0 is associated with a higher accent peak. However, part of the negative correlation comes from the bimodal distribution of accented and unaccented words, where accented words naturally have a higher MaxF0 and lower MinF0A. Therefore, to assure that raising, if there is any, is gradient within accented words, I repeated the same correlation analysis with unaccented words excluded, and the results are shown in **Table 8** and **Figure 17**. It can be seen that MaxF0 is positively correlated with PeakDelay (r = 0.415). That is, in an accented word, when peak occurs later in or after the accented mora, it tends also to be higher. Meanwhile, PeakDelay is also inversely related to MinF0A, r = -0.201 (or r = -0.250 when normalizing data with word-initial F0 value instead). That is, the lower the word-end F0, the later the F0 peak, which in turn is related to peak height. Similarly, a lower value of MinF0A is also associated with a larger initial rise: for RiseSize~MinF0B, r = -0.198 (see also **Figure 17**).

Correlations (N = 1840)

	MinF0B	VMaxRise	VMaxFall	RiseSize	FallSizeA	FallSizeB	PeakDelay
MaxF0				0.694	0.318	0.267	0.415
MinF0A	0.767	-0.064	0.057	**-0.198**	-0.955	-0.745	**-0.201**
MinF0B				-0.214	-0.732	-0.967	-0.103
VMaxRise			-0.125	0.145	0.067		
VMaxFall					-0.060		
RiseSize					0.394	0.382	0.229
FallSizeA						0.786	0.314
FallSizeB							0.205

***Table 8**. Pearson's correlations of normalised data (converted into semitones using utterance-initial F0 value). Non-significant (p>0.05) correlations are not displayed. Data of unaccented words have been removed.*

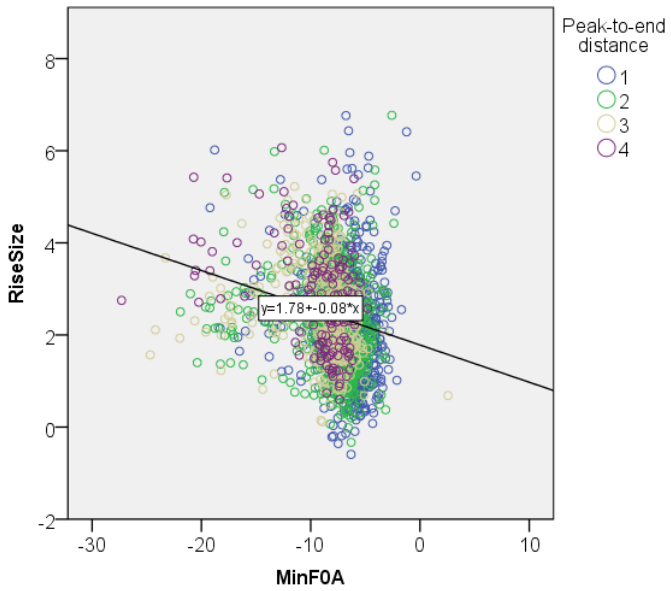


***Figure 17**. Scatterplot of MaxF0~MinF0A (N = 2640, r = -0.198, p < 0.001)*

Given the design of the stimuli, which contrast accent condition and word length at the same time, it is possible that part of the effect could be confounded. Hence, a logical extension of the analysis would be to further divide the data into subsets according to four peak-to-end distance and word length conditions (see **Table 9**). Here a gradient pattern emerges — RiseSize~MinF0A was *r* = -0.317 when accent was 4 morae away from word end, *r* = -0.205 when 3 morae away, *r* = -0.190 when 2 morae away, and *r* = -0.142 when 1 mora away. Note that this is not to be mistaken for the word length effect of F0, because when data was grouped by word length the same gradient pattern became much weaker, with 1-mora words losing negative correlation between RiseSize~MinF0A altogether. The gradient pattern across peak-to-end distance conditions agrees with the comparison of residual sums of squares in §2.3.1 that peak-to-end distance captures variation in accent peak F0 better than word length does. Meanwhile, for unaccented words, MaxF0~MinF0A *r* = 0.639, suggesting a completely opposite behavior to accented words. Finally, grouping data by speech rates reveals that there is less raising effect in slow speech, for PeakDelay~MinF0A, *r* = -0.229 in normal speech and *r* = -0.194 in slow speech.

Peak-to-end distance	Pearson's r RiseSize~MinF0A		Word length	Pearson's r RiseSize~ MinF0A
4 morae	-0.317		4 morae	-0.225
3 morae	-0.205		3 morae	-0.210
2 morae	-0.190		2 morae	-0.202
1 mora	-0.142		1 mora	0.306
Combined	-0.198		Combined	-0.198

***Table 9**. Pearson's r of RiseSize~MinF0A (converted into semitones using utterance-initial F0 value). Data were subsetted into four groups according to peak-to-end distance and word length.*

2.4.4 Discussion

The above analysis has yielded support for PLR as the mechanism of raising F0 of the accented H in Japanese. First, there is evidence of PLR in the PeakDelay~MaxF0 and PeakDelay~MinF0A correlations. On the one hand, other things being equal, a higher MaxF0 should take longer to achieve, hence a greater peak delay. This is confirmed by the positive PeakDelay~MaxF0 correlation. On the other hand, a greater PeakDelay relative to the accent host mora would leave less time for the movement toward the low F0 at word end, resulting in an undershoot of the low F0. But this is contradicted by the negative PeakDelay~MinF0A correlation, which shows that a greater peak delay is associated with a lower F0. Thus a lower MinF0A has led to higher MaxF0, which in turn led to greater peak delay. This is consistent with previous findings about PLR in other languages, except that no measurement of peak delay was taken in the earlier studies.

The second piece of evidence is that these correlations become stronger as the accent is further away from word end. Given that Japanese is generally spoken very fast (mean mora duration 117 ms for normal speech and 161 ms for slow speech in the data), time pressure may have masked part of the PLR effect in the correlations. As peak-to-end distance is reduced, carryover assimilation to the preceding H- due to inertia is more likely to obscure any raising effect, and this also explains why when the subsets in **Table 9** are collapsed the general correlations were quite weak. The fact that the negative correlation in RiseSize~MinF0A is the strongest when accent is four morae away from word end indicates that a lower F0 indeed gives rise to greater initial rise. Meanwhile, *r* is the smallest in RiseSize~MinF0A when accent is adjacent to word end, in which case PeakDelay is smallest because there is a lack of time to reach the low tone, which in turn has led to relatively low MaxF0, thus also a weaker negative correlation. Note that the effect of peak-to-end distance is not to be confused with gradient measurements like PeakDelay and speech rate. Recall that for PeakDelay~MinF0A, $r = -0.208$ at normal speech rate, but $r = -0.146$ in slow speech. This is because, when given more time, the low tone is better reached and so less variable, leading to a weaker correlation. Note also that although PeakDelay per se is not an indicator of PLR, it serves to confirm that accent peak height is the result of articulatory processes, which in turn supports the view that raised peak results from the adjustment of muscle activities, as outlined in 2.4.1.

These two pieces of evidence are in support of the hypothesis that variation of F0 height associated with an accent is a function of the height of the following low F0; the lower the following low, the higher the preceding high. But there is also a previously unreported interaction between categorical and gradient effects. At the word level, there seems to be an effect of gross pre-planning, based on the number of morae available to the speaker for achieving the upcoming low tone, which the speaker can deduce from lexical knowledge. Within a mora, the exact amount of PLR is dependent on how well the low tone is actually achieved,

better at the slow speech rate but worse when accent is adjacent to the targeted low. The height of acoustical landmarks in Japanese prosody thus appears to be shaped by both mora-sized planning and mechanistic articulatory inertia, working in opposite directions.

The absence of consistent effect of word length on the height of unaccented words is interesting. Post-hoc Bonferroni tests showed that although words that are several morae different in length (i.e. 1-mora vs. 4-mora) have significantly different peak F0, most other contrasts do not reach statistical significance, e.g. 3-morae words are not significantly different from words of any other word length. In Selkirk et al (2004), word length ranged from 3 to 7 syllables/morae, whereas in the present study words are 1 to 4 morae long. Another difference lies in how height is measured — maximum F0 of the entire word in the present study, whereas in Selkirk et al (2004) it was 'the F0 at the start of the vocalic nucleus of the initial syllable… and the F0 at the peak of the initial rise (if there was one) or the F0 at the beginning of the second syllable (if there was not)'. The differences in choice of word length and measurement could both be the sources of discrepancy between my results and theirs.

Truckenbrodt (2004) also reported a PLR-like phenomenon in German. He found that a H tone before a downstep was higher than a H tone not followed by downstep. His account of the phenomenon, however, attributes the raised H to the following downstepped accent, rather than specifically to the intervening L in the HLH sequence. Xu & Wang (2001) argued that downstep in a HLH sequence actually consists of two independent mechanisms triggered by L: anticipatory raising (equivalent to PLR) that raises the first H, and carryover lowering that lowers the second H. The present finding is a further demonstration of the anticipatory raising effect, as has been found for different languages (Connell & Ladd, 1990; Gandour et al., 1994; Gu & Lee, 2007; Laniran & Clements, 2003; Laniran & Gerfen, 1997; Rialland, 1981; Xu, 1997, 1999). Truckenbrodt (2004) also attributed the phenomenon of final lowering to lack of upcoming downstep. Although the present study is not designed to examine final lowering, my confirmation of L as the trigger of PLR is at least at odds with his proposal, because the L% boundary tone at the end of an utterance should have been a sufficient trigger for the raising of the final H.

The finding of gradient nature of PLR is consistent with the biomechanical accounts mentioned in the Introduction. Unlike post-low bouncing (Prom-on et al., 2012), however, PLR is a result of local pre-planning in anticipation of an imminent low tone. The current data indicate that the amount of F0 increase depends on the predicted amount of forthcoming lowering based on the number of post-accent morae. Interestingly, the present results also suggest that pre-planning can be done only at the level of the smallest unit of individual movement, which is likely the mora in the case of Japanese. That is, speakers seem to anticipate that a low tone will be better reached as the amount of time available is increased based on the count of number of morae, as shown by the positive relation between MaxF0 and peak-to-end distance. Meanwhile, as seen above, the within-mora effect (mechanical in nature and more gradient) interacts with pre-planning at the word level (mora-by-mora and more discrete).

Logically, the existence of PLR only enhances the surface difference between the two versions of observed H in Japanese, and does not entail that they are underlyingly the same. Further evidence of their identity comes from their perceptual and acoustic indistinguishability in contexts where PLR does not apply, i.e. utterance-finally or when in isolation (Sugito, 1968; Vance, 1995). Combined together, there seems to be sufficient evidence for us to conclude that there is only a single underlying H in Japanese.

The present finding has implications for models of tone and intonation, especially those that simulate articulatory dynamics of F0 production, such as the Fujisaki model (Fujisaki, Wang, Ohno, & Gu, 2005) and the PENTA model (Prom-on, Xu, & Thipakorn, 2009; Xu, 2005). Both models assume that surface F0 contours result from laryngeal responses to muscle activities, which can be simulated by a spring-mass system. PENTA, in particular, assumes that the most basic laryngeal movements are unidirectional toward underlying pitch targets that are specified by communicative functions such as lexical (encoded through tone, pitch accent and word stress), focal and sentential contrasts. Although it has been demonstrated that computationally simulating these unidirectional movements can generate F0 contours that are close to those of natural speech in English, Mandarin and Thai (Liu, Xu, Prom-on, & Yu, 2013; Prom-on et al., 2009; Xu & Prom-on, 2014), there is residual variability that is not yet fully modeled. Prom-on et al. (2012) showed that part of the residual variability comes from post-low bouncing, which can be modeled by adding an F0-raising force at the end of a low-approaching movement. The present results show another source of residual variability not yet accounted for by PENTA or any other computational model. The present version of PENTA would have to capture this variability as a context-specific target change, akin to Tone 3 sandhi in Mandarin (Yip, 2002), but such an account would bear no relevance to the underlying mechanism being proposed, and achieves nothing more than introducing an ad hoc predictor in the model. Given its planning nature, and relatively weak correlation with low F0, the modeling simulation of PLR would be rather different from that of post-low bouncing, which needs to be explored in future studies.

Furthermore, the present findings also have implications for Japanese tonal phonology as well as tonal phonology in general. Some of the implications are relatively direct. For example, the gradient nature of pre-low raising means that this type of surface difference in F0 is unlikely due to a phonemic or tonomic contrast. This would significantly simplify the tonal phonology of Japanese. Other implications are more indirect. For example, the pervasiveness of PLR as seen in both Japanese and many other languages suggests that, downstep, of which PLR is part (proposed by Xu, 1997 and supported by the current data), is also unlikely a contrastive phonological process. Also the exclusion of downstep or its lack thereof, from part of the explanation of final lowering (contra Truckenbrodt, 2004), means that the latter is likely due to an independent mechanism, e.g., to contrast statements from questions (Liu & Xu, 2005). Most importantly, the identification of PLR as an independent process as well as its possible articulatory mechanism, together with the identification of other articulatory mechanisms, such as post-low bouncing, inertia-triggered extensive F0 transitions (Gandour et al., 1994; Xu, 1997, 1999), undershoot (Xu & Wang, 2009) and peak delay (Xu, 2001) suggests that surface F0

patterns cannot be used as direct basis for positing underlying phonological tones or tonal processes such as scaling, downstep, and spreading. Instead, for each observed F0 pattern, it is imperative to determine not only the potential contrastive function behind it, but also the likely articulatory mechanisms involved.

Finally, the generalizability of the present finding needs further empirical support from other languages. One of my next steps will be to investigate whether the same correlations can be observed in languages where PLR is well established, e.g. Thai, Cantonese and Mandarin. Moreover, although I have found evidence of mora-level categorical pre-planning, it is also possible that the syllable is the real tone bearing unit of the language, as is the case for languages like Mandarin and English. This issue will also be examined in Chapter 3.

2.5 Chapter conclusion

In this Section, I have used a quantitative approach to show that in Japanese the F0 peak associated with a pitch accent varies with its following low tone. I have found evidence that the variable F0 peak height is the result of pre-low raising. Pearson's *r* reveals an inverse relation between accent peak and the following low tone, and that such a relation becomes more pronounced when the peak is further away from the low tone. That the effect of PLR is masked when the accent peak is close to word end may explain the absence of similar findings in the literature. These results suggest that in Japanese a low tone raises its preceding high tone, which is consistent with our current understanding of the physiology of vocal fold tension control in F0 production.

CHAPTER 3: SYLLABLE AS THE TONE-BEARING UNIT FOR JAPANESE[16]

3.1 Introduction

Whether pitch accent is carried by the mora or the syllable in Japanese has been a long-standing debate. Many researchers have investigated the reality of the mora in terms of isochrony (see Warner & Arai, 2001 for a review), or illustrate the role of the mora as a unit of rhythm and of prosodic measurement in phonology (e.g. Labrune, 2012); while others argue the temporal domain of Japanese is the syllable (McCawley, 1968), or neither of the two (Yoshida, 1990). Although the behavior of the mora in phonology is now relatively well understood, it does not necessarily follow that it is also the domain of tone articulation. In this Chapter, I show evidence that while the mora may well play a role in phonology, it is the syllable that tonal articulation is based on, with reference to the notion of Target Approximation (TA).

As introduced in Chapter 1, TA is a theory of articulatory mechanism that delineates how surface fundamental frequency (F0) results from underlying tone targets. It assumes that each tone-bearing unit (TBU), typically the syllable, is assigned a pitch target that is either static (slope ≈ 0, i.e. flat) or dynamic (slope ≠ 0, i.e. rising or falling). The surface F0 is the result of asymptotic approximation of the underlying target in full synchrony with the TBU. At the boundary between two adjacent targets, the final articulatory state of the first one is transferred to the second one. Such transfer often results in a delay of the apparent alignment of an F0 landmark (e.g. peak, valley). The temporal alignment of these turning points is not specified in the model, as it is the result of syllable-synchronised realisation of underlying pitch targets and is thus predictable by TA. Equally speaking, it is possible to induce the characteristics of an underlying target from surface F0 alignment, as we shall see below.

The size of TBU matters in tonal articulation. Where there is insufficient time (e.g. a short TBU) to reach a tonal target, TA would predict a target undershoot (cf. also Lindblom, 1963 on the articulation of segments). Hence whether tonal targets are hosted in the mora or the syllable, or even larger units in Japanese, would lead to very different proposals of its tonal inventory as well as different predictions of its surface realisation of intonation. The issue of TBU in Japanese has never been touched upon from this articulatory point of view, and the present Chapter seeks to fill this gap.

Here the hypothesis that the syllable is the domain that carries underlying tonal targets in Japanese is tested, using accentual fall as a test case. The accentual fall is particularly suitable for this purpose, because it involves a falling F0 contour that spans at least two morae, and can occur both in heavy syllables (e.g. ***me****n*) and in light syllable sequences (e.g. ***me****mo*).

[16] This Chapter was presented at the Linguistic Society of Hong Kong Annual Research Forum 2014.

I will test the two TBUs with different tonal inventories. Following conventional analyses (Pierrehumbert & Beckman, 1988), I adhere to a parsimonious [High] and [Low] inventory for the mora hypothesis. It is possible to propose additional tones just in case, but such an expanded tonal inventory would be theoretically ungrounded and introduce unnecessary degrees of freedom. On the other hand, for the syllable hypothesis I propose a tonal inventory of [High], [Low], and [Falling]. The [Falling] target[17] is necessary in this case, because if syllable was indeed the TBU, a [High] target on the accented heavy syllable (cf. Figure 18 right panel) would not yield (under TA) a peak F0 in the middle of the vowel which is then followed a sharp fall, as observed in §2.3.2 and Ishihara (2006). The introduction of the [Falling] target for CVV accent hosts is not entirely arbitrary; similar context-dependent target assignment is also observed in English (Liu et al., 2013; Xu & Xu, 2005), where the slope of the underlying target of the statement-final position varies according to focus condition, thus arguably an allomorphic alternation. The proposed [Falling] target is an underlying target at the stage of tonal articulatory planning, and is considered an 'allophonic' implementation of the pitch accent in the context of a heavy syllable (whereas the 'allophone' in a CVCV context is the [High][Low] target sequence).

The two hypotheses are to be distinguished from phonological theories that view both types of pitch accent as an H+L sequence. For example in the J-ToBI (Venditti, 2005) annotation convention, although pitch accent is represented as H*+L, whether it is a mora or a syllable that hosts the tone is a non-issue, because more than one tone can be attached to a TBU. The goal of this Chapter is to show that there are both High and Falling targets in Japanese and that tonal articulation is based on the syllable rather than the mora.

Figure 18. *Conceptual representation of the underlying targets of* memo-mo *'memo also' (High-Low-Low) in the left panel, and* men-mo *'face also' (Falling-Low) in the right panel. The dashed lines represent underlying targets.*

This hypothesis will make the following predictions: First, as illustrated in Figure 18, bearing a different number and nature of underlying targets, light and heavy syllables should see differential timing of F0 peak. Second, the respective articulatory parameters (e.g. slope

[17] The original idea came from a pilot study, in which a native speaker produced ***me****n* and ***me****i* at a very slow speed with a falling pitch pattern like the Mandarin Falling tone, whereas her slow production of ***me****mo* was clearly a [High][Low] sequence. This led me to hypothesise that CVV/CVn and CVCV had different underlying targets.

and height) of the two hypothetical tone targets should be considerably different and form a bimodal distribution. Third, using PENTAtrainer1, a prosody synthesiser, modelling an accented long syllable as a Falling target should yield better accuracy than as consecutive High-Low morae, other factors held constant.

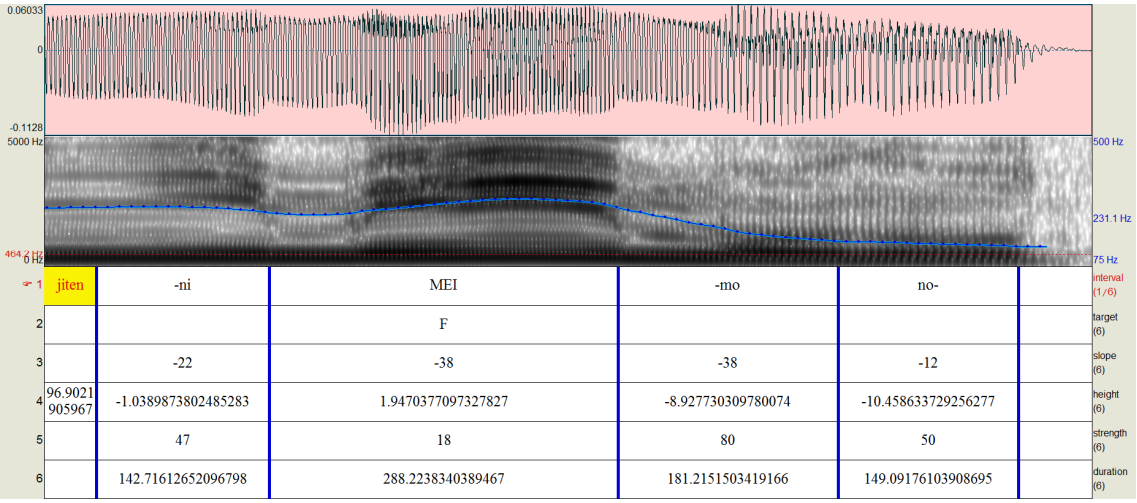
3.2 Methodology

The corpus used in the present study is the accented subset of that reported in Chapter 2. The 23 accented Japanese target words varied in length (1~4 morae), accent condition (initial/medial/penultimate/final accent), and syllable structure (CVCV, CVn, CVV). From eight speakers, altogether 1,840 utterances (23 target words × 8 speakers × 5 repetitions × 2 speech rates) are analysed. The target words are framed in the carrier sentence *jiten-ni X-mo nottemasu* 'The word X too is found in the dictionary'.

PENTAtrainer1 (Xu & Prom-On, 2010-2015) and PENTAtrainer2 (Prom-on & Xu, 2012) are semi-automatic software packages for analysis and synthesis of speech melody based on communicative functions and TA (Xu & Wang, 2001; Xu, 2005), written in the form of Praat (Boersma & van Heuven, 2001) scripts. The basic idea of the PENTAtrainers is to extract the articulatory parameters of F0 targets [height (*b*), slope (*m*), and strength (*λ*)] by means of analysis-by-synthesis based on the quantitative Target Approximation (qTA)(Prom-on et al., 2009). For further details about the PENTAtrainers see Prom-on et al (2009) and Xu & Prom-on (2014) and see Lee et al (2014) and §6.2 for an example of how these tools are applied on the current data set.

PENTAtrainer1 extracts target parameters locally unit by unit through exhaustive search. For each target interval (typically the syllable), PENTAtrainer1 compares all possible combinations of *b, m,* and *λ* within the search ranges and finds the parameter combination that generates F0 contours with the least difference from the original. For the present study, I used a feature of this tool which allows imposing specific target slope, so as to test whether it is a Falling (negative slope) or High-Low (zero slope) target that yields better accuracy. A further subset (N = 800, accented CVCV words excluded) of the dataset that consists of words with a long accented syllable (e.g. ***me****n*, contra ****me****mo*) will be analyzed using PENTAtrainer1. Two annotation schemes will be compared for their overall learning accuracy. The two schemes differ only at the accented syllable, where in the Mora scheme the accented syllable is modelled as having consecutive [High][Low] targets, while in the Syllable scheme it is modeled as having a single long [Falling] target. **Figure 19** is the user interface of PENTAtrainer1, in which tiers three to six are automatically extracted local articulatory parameters. Users need to annotate (with any label) on Tier 1, then the script will extract articulatory parameters for all labelled intervals. Tier 2 allows restriction of target for hypothesis testing. By specifying a target for a given interval, the search range of that particular syllable will be restricted. Annotation on this tier is optional, but is useful when there are specific hypotheses to be tested. For example, in

Figure 19, for the syllable that has a F target specified, PENTAtrainer1 will extract a negative target slope, and optimise height and strength values to fit around the imposed negative slope. The target imposing feature will allow us to direct test High vs. Falling targets, the result of which will be presented in §3.3.3.



1	jiten	-ni	MEI	-mo	no-	
2			F			
3		-22	-38	-38	-12	
4	96.9021 905967	-1.0389873802485283	1.9470377097327827	-8.927730309780074	-10.458633729256277	
5		47	18	80	50	
6		142.71612652096798	288.2238340389467	181.2151503419166	149.09176103908695	

***Figure 19**. Snapshot of annotation with PENTAtrainer1.*

PENTAtrainer2 extracts qTA targets globally from an entire corpus by means of analysis-by-synthesis based on simulated annealing, a machine learning algorithm (Xu & Prom-on, 2014). Two annotation schemes were tested: 'Mora' and 'Syllable' (as in A. Lee et al., 2014), which respectively represents the scenarios of mora and syllable being the TBU. The extracted articulatory parameters from these two schemes will shed light on the nature of the underlying targets in question. See Lee et al (2014) and §6.2.2.2 for more details of annotation.

3.3 Results and discussion

3.3.1 Evidence 1: F0 peak timing

The first set of results comes from peak delay, defined as the time lapse between the onset of the accent host mora/syllable and the F0 peak of the pitch accent. Mean peak delay in words with a short accented syllable (N = 208, hypothesised as a [High] target) and those with a long accented syllable (N = 160, hypothesised as the [Falling] target) are 118 ms and 158 ms, respectively.

In §2.3.2 the effects of speech rate and peak-to-end distance on peak delay ratio were reported, here I am interested in the effect of syllable structure on actual peak delay (ms). **Figure 20** shows the general pattern of actual peak delay in the data. Long syllables have a slightly later F0 peak than short syllables. A one-way ANOVA was conducted to examine the effect of syllable structure (CVn/CVV/CV) on actual peak delay, and revealed a significant main effect of

syllable structure F(2,365) = 4.424, p = 0.013. A Helmert contrast was specified to compare the peak delay in CV vs. CVV+CVn, which were significantly different (p = 0.004). Meanwhile, post-hoc Bonferroni test showed that the difference between CVV and CVn was not significant.

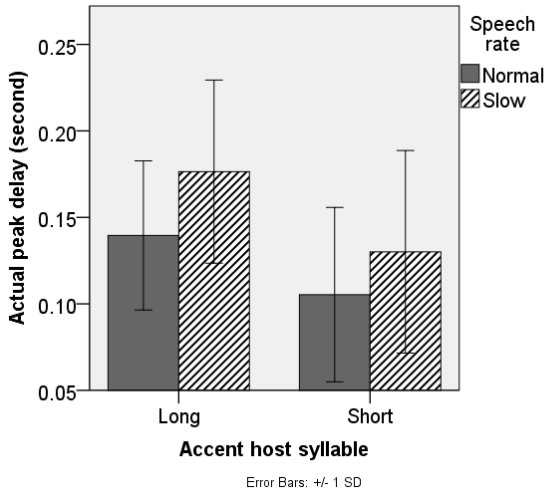


Figure 20. *Mean actual peak delay (ms) of hypothesised Falling and High targets.*

In the present dataset, in a word like ***me****mo* (High-Low, see **Figure 21**) F0 peak occurs near the middle of the word, after which F0 needs to change course for the Low target. That peak delay is greater in an accented long syllable is surprising, and needs to be explained in the context of the syllable being the TBU. Sato (1993) reported that in Japanese the duration of CVn is shorter than CVCV (the ratio of CVn:CV being 1.50~1.71:1), meaning that the mora boundary in CVn should occur earlier than its CVCV counterpart. Then, if a long syllable carried tones as two morae like in CVCV, we should expect a smaller peak delay, because at the TBU boundary target approximation would be reset. The fact that the opposite is observed thus implies that pitch accent in CVn/CVV and that in CVCV have different underlying targets. By extension, that pitch accent in CVCV ([High][Low]) reaches its peak earlier is due to the need to change course from [High] to [Low] at the syllable boundary; whereas CVn/CVV, though shorter than their CVCV counterpart, has an entire long syllable to realise the falling contour.

Evidence from Mandarin (Xu, 2001) shows that F0 peak of a [High] target occurs near the end of the syllable, that of a [Falling] target occurs considerably earlier, whereas the peak of a [Rising] target is the latest. Judging from **Figure 20**, if CVn/CVV were two TBUs like CVCV, accent peak would fall on the post-accent mora, meaning in the accent host per se there is a rising F0 contour that extends way into the following mora — comparable to the Mandarin rising tone. However, to hypothesise that a long accented syllable constitutes a [Rising][Low/Falling] target sequence would be theoretically ungrounded and explains the data in circularity. Hence, the remaining possibility is that a long accented syllable is a single TBU with a [Falling] underlying target, which will be verified with further evidence in the following subsections.

Potentially, one could argue that CVV/CVn consists of two TBUs (still a [High][Low] target sequence), with the mora boundary very late into the long syllable (based on the

observation that the accent peak is close to the TBU boundary). Albeit more parsimonious than the present proposal, this alternative was not considered because actual peak timing (or peak delay ratio) is highly variable depending on linguistic factors such as word length and accent condition (see discussion in Chapter 2 and **Figure 11** in Page 39). For example, for a CVn word the F0 peak of a pitch accent that is four morae away from word end occurs 76% into the syllable at normal speed, but 27% into the syllable when the peak-to-end distance is one mora. To propose that there is a gradient TBU boundary in CVn/CVV words is possible, but it too would also be explaining the data in circularity. Then why is a pitch accent articulated as a [Falling] target in a heavy syllable, when it is known to have an earlier peak than [High] and [Rising]? This is because the TBU is longer (syllable) in CVn/CVV than in a CV; Xu (2001) reports that when a syllable is lengthened (due to slower speech rate in his study) F0 peak shifts rightwards accordingly, i.e. aligned to the end of the TBU. Compared to a light syllable, CVn/CVV is longer in duration, hence a significantly later accent peak.

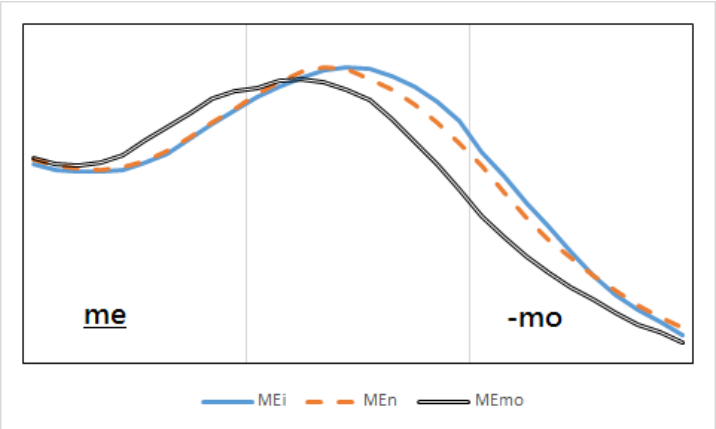


***Figure 21**. F0 contours of the minimal triplet **me**n 'face' vs. **me**i 'May' vs. **me**mo 'memo' spoken at normal speed, averaged across 40 repetitions. X-axis shows normalised time, and y-axis F0 in semitons. The first two intervals represent the target words, and the third interval is the following particle –mo. Vertical lines are mora boundaries.*

3.3.2 Evidence 2: Global articulatory targets from PENTAtrainer2

Having established that long and short accented syllables bear different underlying targets, I now look into their behavior further using PENTAtrainer2. The articulatory parameters of the two hypothetical underlying targets were learned for each speaker in the corpus. Pending verification, the two targets are referred to as [Falling] and [High] for long and short accented syllables, respectively.

Results in Figure 22 are in line with my expectations. Firstly, [Falling] targets have greater negative slope than their [High] counterparts, which have a positive or near flat (flat≈0) slope. Second, the two target types are distributed as two separate clusters, confirming that they should not be viewed as the same. Note that the mildly positive slopes in (word-medial)

High targets learned by PENTAtrainer2 reflect the upward interpolation between phrasal High (H-) on the second mora and the accent peak (H*), the latter of which is usually higher in F0 (Beckman & Pierrehumbert, 1986b). They, however, are not to be taken to mean that Japanese pitch accent is underlyingly a rising target. Like in **Figure 18** (left panel), F0 is rising in the first syllable to approach the [High] target, but the target itself is not rising. Meanwhile, the positive slopes in word-initial [High] targets reflect word-initial rise that marks AP boundaries in Japanese (Pierrehumbert & Beckman, 1988). With [Falling] targets, of which the domains are long syllables, the slopes are unambiguously negative, showing that a [Falling] target in a long syllable is compatible with Japanese lexical prosody. As will be shown in §6.2, F0 resynthesis based on these learned targets is highly accurate, suggesting that these targets truly reflect the articulatory targets intended by the speakers.

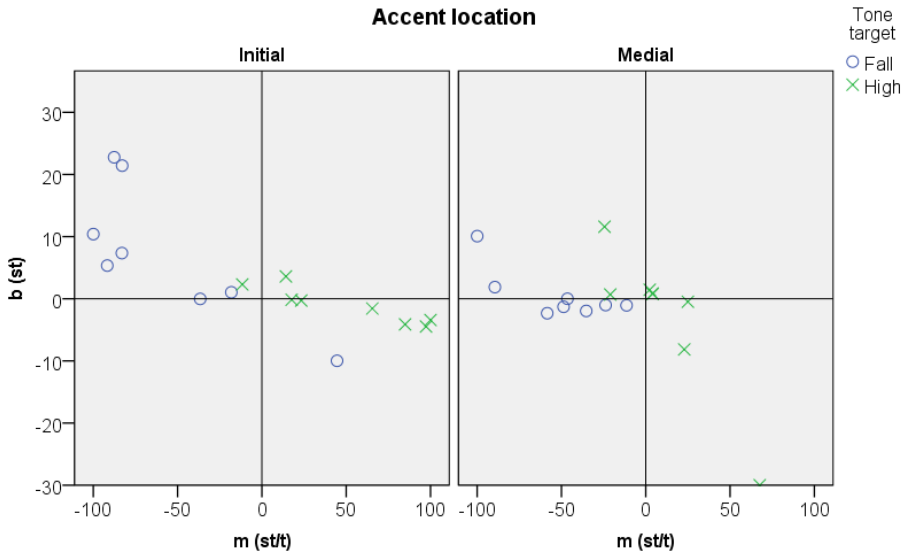


Figure 22. *Distribution of articulatory targets (m = slope, b = height) in word-initial accents (left panel) and word-medial accents (right panel). Each point represents one speaker.*

3.3.3 Evidence 3: Learning accuracy with restricted target search

To ensure that the evidence so far is not but acoustical artifacts of comparing one mora vs. two morae, additional confirmation was sought using PENTAtrainer1. In Lee et al (2014), PENTAtrainer1 was trained using the present dataset without underlying targets specified (see also §6.2 for more details of PENTAtrainer1). Here the procedures are replicated but with accent targets imposed. This function of PENTAtrainer1 allows us to directly test whether a [Falling] underlying target yields better fit than a [High][Low] sequence. To impose a specific target, the user simply needs to input the tone target (e.g. H, L, F, R) in the appropriate interval label of the Praat textgrid file. Only intervals that are being tested require input, others can be left blank. A labelled interval will have a restricted search range — for example when the

syllable is labelled as H, the slope value is restricted to 0, while the other two parameters are optimised to achieve the best fit.

Target	RMSE[18]	Pearson's *r*
Falling	0.1534	0.9969
High-Low	0.1780	0.9941
Unspecified[19]	0.1545	0.9969

Table 10. Overall learning accuracy in RMSE and r using PENTAtrainer1 under two imposed targets.

The data subset (N = 800) tested here contains only words with a long accented syllable. I compared two annotation schemes, one with the long syllable as two intervals bearing H and L, another as a single interval bearing F. Where pitch accent is a Falling target the learning accuracy is better, with higher *r* and lower RMSE (*r* = 0.9969, RMSE = 0.1534) than when accent is split into two morae (i.e. [High][Low] sequence) (*r* = 0.9941, RMSE = 0.1780). Although the improvement from [High][Low] targets is small, a paired samples t-test confirms that the difference in learning accuracy between the two types of imposed target is statistically significant (t(7) = 4.017, *p* = *0*.005 for *r,* t(7) = -7.388 *p<0*.001 for RMSE). Therefore, the improvement achieved as a result of the imposed Falling target, albeit small, is meaningful. Recall that the two annotation schemes differed only at the accent host syllables and are otherwise identical, hence any improvement is solely attributable to the Falling target and to the syllable being the target approximation domain.

The slight improvement in RMSE from the unspecified scheme is counter-intuitive, as PENTAtrainer1 is supposed to obtain the best possible local fit when target is not specified. Paired samples t-test shows that their difference (Falling vs. Unspecified) is significant t(159) = -2.985 *p* = 0.003. In any case, the similarity/identity of learning accuracy in Falling vs. Unspecified shows that a [Falling] target yields the best possible local fit for F0 contours in heavy accent syllables.

Finally, other factors held constant, two labelled intervals should have led to better accuracy than having one interval, as a larger degree of freedom should capture variability in the data better; the higher learning accuracy with the [Falling] (i.e. one interval) observed here thus makes a strong case that it is the real underlying target in accented heavy syllables instead of [High][Low].

[18] RMSE (root-mean-square error) and Pearson's *r* are two correlation coefficients used for evaluating F0 contours. RMSE measures the difference between natural and synthesized F0 contours while correlation coefficient indicates the linear relationship between them. See Prom-on et al (2009) for a more detailed description.

[19] This is different from Lee et al (2014) and §6.2.3.1, where the accuracy reported is of a larger dataset (N = 2640). In **Table 10** the accuracy is reported for the accented heavy syllable subset (N = 800).

3.4 Discussion and Chapter conclusion

This Chapter set out to answer the mora/syllable dichotomy in Japanese from a Target Approximation perspective, and has found three pieces of evidence in support of the syllable as the domain of tonal target approximation — namely F0 peak timing, bimodal distribution of target slope and height, and PENTAtrainer1 learning accuracy. These results suggest that Japanese is like languages such as Mandarin and English, where the syllable bears tones during tonal articulation.

There is one important question that the present dataset was unable to answer — whether a morpheme boundary within a heavy syllable would be articulated as one TBU or two. For example, in ***ki****i* キー 'key' vs. ***ki****.i* 奇異 'strange', the latter word consists of two morphemes (*ki* 奇 and *i* 異). If a morpheme boundary makes a heavy syllable pronounced as two TBUs (CV.V), one should expect an earlier accent peak than in a CVV case (see §3.3.1 above). Future research should look into this issue using appropriate minimal pairs to examine whether morphology affects articulatory planning in this situation.

Finally, it must be emphasised again that acoustical evidence presented in this Chapter serves to argue for the syllable as the unit of articulatory planning, and should not be interpreted as evidence against the existence or functional necessity of the mora for other purposes.

CHAPTER 4: DECLARATIVE FOCUS PROSODY

4.1 Introduction

4.1.1 Basic facts about Japanese focus prosody

In a focused statement in Japanese, F0 is raised on the focused item whereas the excursion size of initial rise is compressed after focus (Sugahara, 2002, 2003a). In cases where on-focus F0 range expansion can be difficult, like in an unaccented word, there is the optional focus-marking strategy known as 'prominence-lending rise' (Venditti et al., 2008). Prominence-lending rise is the 'additional F0 rise on the phonological word-final/penult mora' (S. Ishihara, 2015, p. 600, where it is referred to as a "late high rise"), often a particle or a case marker but not necessarily so. There are also non-F0 cues to focus, such as duration and formant frequency (Maekawa, 1997), though F0 is by far the best understood cue to date. There are conflicting reports over the existence of certain focus markers, such as pre-focus modification of duration (A. Lee & Xu, 2012; Maekawa, 1997) and pre-focus F0 reduction (Hwang, 2011; A. Lee & Xu, 2012); but that post-focus compression (PFC) marks focus in Japanese is hardly disputed.

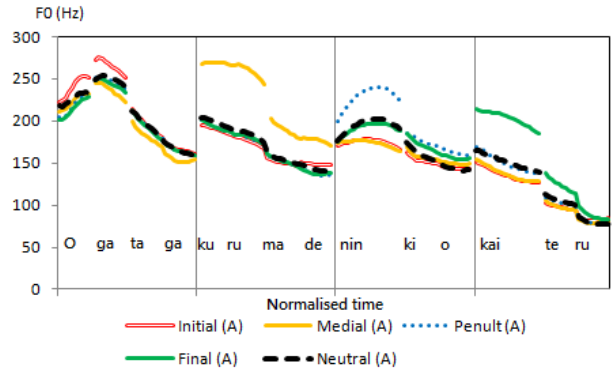


***Figure 23**. F0 contours of the sentence* o**ga**ta-ga **ku**ru-made **ni**nki-o **ka**iteru *'Ogata wrote "ninki" until he came' spoken in four focus conditions averaged across 7 speakers.* (from A. Lee & Xu, 2012)

Figure 23 illustrates how F0 changes on-focus and post-focus in an accented utterance in Japanese. Consider the yellow contour (narrow focus on *kuru-made*) vs. the dashed black contour (neutral focus). There is substantial on-focus raising of F0 (higher yellow contour in the second interval from the left) as well as post-focus compression of F0 range (smaller F0 range on the yellow contour in the third interval from the left). However, there is no strong evidence for any pre-focus modification (cf. Hwang, 2011).

A related notion is the new/given contrast, which shows similar prosodic markers. Sugahara (2003a) reports that givenness lowers F0 in Japanese in addition to PFC, comparable to the contrastive focus/neutral distinction. However, newness is not the same as focus, which

'indicates the presence of alternatives that are relevant for the interpretation' (Krifka, 2008, p. 247); the notion of newness per se does not entail the presence of an alternative. Focus and newness are encoded differently/differentially in prosody too — Sugahara (2003a) shows differential patterns in PFC depending on whether the post-focus material is given or new.

Finally, there is a strategy known to effectively elicit focus prosody in the absence of contrastive focus — WH-words (Deguchi & Kitagawa, 2002; S. Ishihara, 2002; Maekawa, 1997). A WH-word has a higher F0 peak and is followed by a compressed F0 range, just like a contrastive focus. Kubozono (2007) argues that WH-words can elicit more natural-sounding focus prosody by not involving contrast in the discourse, which is deemed to lead speakers to exaggerate the focus markers. In the present study Kubozono's strategy is not adopted, for reasons to be elaborated in §4.2.

4.1.2 Autosegmental-Metrical representation

Japanese focus is generally represented in Autosegmental-Metrical Theory as Intonational Phrase boundary insertion immediately before the focused item, which results in 'pitch reset', and deletion of AP boundaries after focus (dephrasing), which causes the non-realisation of the initial rise (Pierrehumbert & Beckman, 1988). This representation has a number of theoretical implications that have attracted serious debates. One important issue is whether F0 is reset at the focused item as a result of Intonation Phrase boundary insertion. If this is the case, the resetting of F0 would undo any downstep effect from previous accented words; indeed Pierrehumbert and Beckman (1988) found no significant effect of the previous accent condition on the F0 peak of the focused item, and contended that F0 is reset under focus.

Figure 24 illustrates the workings of prominence marking in J-ToBI. The solid black contour shows an utterance under final focus on the verb *o**yo**ideru* 'swimming'. Here, the L% after *–wa* is both the inserted IP-initial boundary to mark focus as well as the AP-final boundary that marks the prosodic word. In contrast, the dashed grey line shows the neutral focus counterpart of the same utterance, where the same medial L% stands for AP-final boundary only. The result of the difference in identity between the two L%'s is that downstep is 'blocked' in the final focus utterance (solid black curve), leading to a stark contrast between the two versions of the same accent peak. Further, the inserted IP boundary would also lead to such boundary effects as phrase-final lengthening on the *–wa* (Venditti et al., 2008).

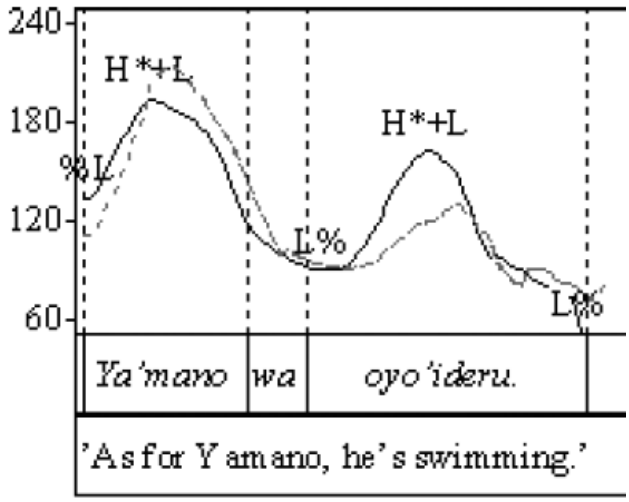


Figure 24*. F0 contours of two versions of* **ya**mano-wa o**yo**ideru*. The gray dashed line represents neutral focus, whereas the solid black line shows narrow focus on* o**yo**ideru (from Venditti et al., 2008, p. 465)*.*

Findings in subsequent work have enabled a better understanding of the implications of Pierrehumbert and Beckman's (1988) original proposal. For example, in reexamining the notion of on-focus pitch 'reset', Kubozono (2007) found significant difference in the on-focus F0 peak height between a downstep-prone condition and otherwise, which led him to posit a local F0 boost notion for Japanese focus marking instead. His proposal captures the fact that the downstep effect from preceding words is not undone at the focused item. He attributed the discrepancy between his conclusion and Pierrehumbert and Beckman's to the materials they used in their production tasks. Specifically, he pointed out that Pierrehumbert and Beckman (1988) used leading questions to elicit contrastive focus, which informants would tend to exaggerate. To elicit a more natural focus, he instead used WH-words which lexically attract focus.

In terms of post-focus dephrasing, Sugahara's (2003) study revealed that Intonational Phrase (or 'Major Phrase' in her original wording) boundary was also subject to deletion when the post-focus word was given; when the post-focus word was new, the boundary was intact. Her findings suggest that the level at which post-focus dephrasing occurs is determined by the interaction between focus and the new/given distinction.

4.1.3 Post-focus compression (PFC)

At this point it is important to define what PFC is. To the best of my knowledge, the term was first explicitly used in Sugahara (2003a) where she measured the excursion size of the Japanese AP-initial rise; elsewhere, Xu (1999) used the term 'suppressed' to describe lowered and compressed pitch range after focus in Mandarin. Subsequently PFC was increasingly studied as a phenomenon and explored from a typological perspective (Xu, Chen, & Wang, 2012). It refers to the compression of pitch range after prosodic focus widely observed across

languages, but exactly how it is measured varied among researchers. In American English, for example, it is realised as lowered and compressed F0 range for statements, and as raised and compressed F0 range for questions (Eady & Cooper, 1986; Liu & Xu, 2007). On the other hand, in utterances without much F0 fluctuation, like a Mandarin statement that consists of only High tone words (e.g. Xu, 2005, p. 236 Fig. 8a), 'PFC' can manifest itself as mean F0 lowering instead. Thus in examining whether a language makes use of PFC, or more specifically pitch range, lexical prosody must be taken into account or the wrong conclusion about focus marking would be reached.

In the case of Japanese, PFC is robustly observed in AP-initial rise, after both accented and unaccented words. Looking across an entire AP, PFC could also be taken to mean a compressed pitch range between an accent peak and the minimum F0 of that phrase; but measured this way PFC has been reported to be absent after unaccented words. In both Ishihara (2011a) and Lee & Xu (2012), F0 range was not found to be compressed after an unaccented focus if PFC was measured as the difference between F0 maxima and minima within that word. As Ishihara (2015, p. 601) describes it, where the focus is unaccetned 'the pitch contour exhibits a high plateau following the focal F0 rise' which, as we shall see in §4.3, resembles post-focus raising. In this plateau there is observable but compressed AP-initial rise, which is taken by Sugahara (2003a) as evidence of PFC.

Consider **Figure *25***, where the focused item (N2) is unaccented, the F0 range of the first word after focus (first three points of the dashed curves in the red region) is larger than its neutral focus (solid curves) counterpart; there is only 'compression' if one confines his measurement to the first AP-initial rise in the post-focus contours (excursion between the first two points of the dashed curves in the red region) — the definition of PFC adopted by Sugahara (2003a). Comparing the findings in Sugahara (2003a), Ishihara (2011a) and Lee & Xu (2012), it is thus clear that whether there is PFC in Japanese largely depends on how it is defined.

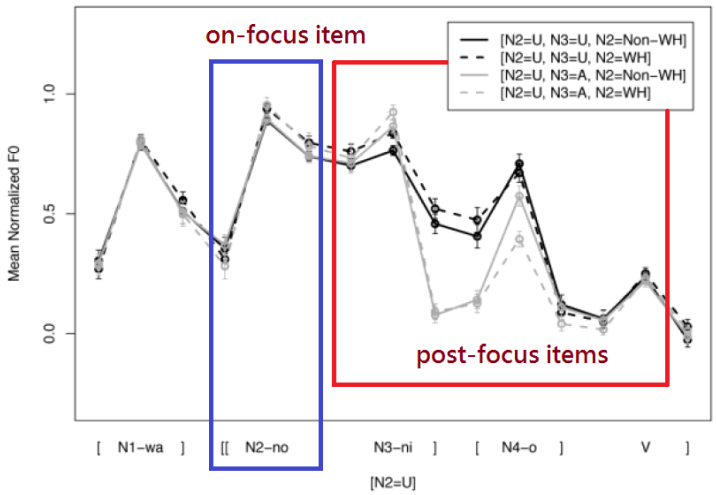


Figure 25. *F0 contours from Ishihara* (2011a) *illustrating the effect of unaccented focus on PFC (my annotation).*

Here I argue for an alternative account: PFC, as the difference between phrase-internal maxima and minima, is realised only after a pitch accent, elsewhere F0 is a result of carryover effect from the preceding word. In turn, I claim that although typologically PFC is a focus marker in Japanese, it does not follow that PFC is applied across the board. I illustrate my argument through a newly collected corpus that controls for accent condition of all words in the sentence, focus condition, and word length. While introducing within-language variation in terms of where a phonetic feature (i.e. PFC) can be realised, my account has the advantages of rendering Japanese comparable to other languages with regards to how PFC is measured, and conforming to the notion of focus trizone (Xu, Xu, & Sun, 2004). According to Xu and colleagues (2004), narrow focus controls the pitch range of the on-focus, pre-focus and post-focus regions, each with a different strategy. Because the boundary tone (%L) that forms part of the initial rise (%L H-) is 'shared' between two words, restricting PFC to initial rise (which serves to mark the beginning of a prosodic word) would mean this phonetic feature is no longer purely post-focus but is looking across both the on-focus and the post-focus regions — confusing for those to whom the focus trizone notion matters and confounding a post-focus feature with the effect of on-focus raising. In addition, attributing F0 movements in unaccented utterances to carryover effects means resorting more to phonetic and mechanistic notions and less to abstract ones, which is desirable in my experimental approach. Whereas the basic facts about Japanese focus prosody have been well established in previous research, the goal of this Chapter is to offer a new angle to PFC realisation and typology through a comprehensive experiment design that exhausts logical possibilities.

4.1.4 Research questions

The overarching goal of this Chapter is to offer a detailed phonetic description of PFC across different accent conditions and focus conditions. If PFC is only consistently observed in AP-initial rise (cf. N3 in **Figure *25***), what could we say about the rest of the post-focus region? Other interesting questions that I seek to address include the existence of pre-focus modification and the scope of PFC[20]. First, researchers have yet to agree on whether there are pre-focus F0 modifications. Hwang (2011) reported lowering of F0 before focus in a production study with four speakers, whereas in Lee & Xu (2012) pre-focus lowering was observed only for some speakers. Second, Ishihara (2011a) notes that in some cases PFC does not extend beyond one word in the post-focus region. I intend to examine this further and find out what influences the scope of PFC.

[20] In this Chapter we are only interested in the effect of the accent-focus interaction on the scope of PFC. For good discussions on how syntax affects the scope of focus please see Deguchi & Kitagawa (2002), Ishihara (2002, 2003, 2007).

4.2 Methodology

A production experiment was conducted with 13 Japanese speakers from the Greater Tokyo area (Tokyo, Saitama, Kanagawa, Chiba). Participants were recruited on the website http://mixb.net. One speaker turned out to be a English-Japanese bilingual, one withdrew from the experiment after deciding the target sentences were difficult, another completed the task but did not produce most of the contrasts I intended to elicit (e.g. focus). Data from these three speakers were discarded. Hence here data from ten speakers are reported, five from each sex, aged between 24 and 36 (mean 30.3, S.D. 4.15, see **Table 11**). All participants were remunerated a small sum for their time, and granted their written consent to being tested.

Initial	Age	Sex	Born	Grew up	Father from	Mother from	Occupation
KK	25	M	Saitama	Saitama	Tokyo	Akita	Student
KM	24	M	Tokyo	Tokyo	Tokyo	Yamagata	Unemployed
TM	30	F	Saitama	Saitama	Nagano	Tokyo	Dog trainer
TK	25	M	Chiba	Tokyo	Tokyo	Tokyo	Student
OY	36	M	Tokyo	Saitama	Tokyo	Kagawa	Student
YK	30	F	Kanagawa	Kanagawa	Kanagawa	Kanagawa	Student
IK	34	F	Saitama	Saitama	Saitama	Saitama	Unemployed
KE	33	M	Tokyo	Tokyo	Tokyo	Hiroshima	Student
MY	25	F	Tokyo	Kanagawa	Kanagawa	Tokyo	Student
KT	33	F	Tokyo	Kanagawa	Kanagawa	Tokyo	Actress
AM	33	M	Tokyo	Tokyo	Kanagawa	Tokyo	Student

***Table 11**. Information of participants in Chapters 4 and 5*

4.2.1 Stimuli

As the main focus markers have already been identified in the literature reviewed above (using WH-words like in Kubozono, 2007; or leading questions with natural contexts like in Sugahara, 2003a), this study will focus on examining the interplay between pitch accent and focus, as well as the post-focus F0 patterns in cases where PFC is absent. To this end, unnatural stimuli will be used to elicit strictly minimal contrasts and to control for possible confounds from microprosodic variations. Such a design is unusual and generally not preferred, but given a wealth of insights from previous literature on this topic, it is now possible to take a more stringent approach at the expense of naturalness. As we shall see, the focus markers used by speakers in this study are not different from those previously reported.

Altogether 128 target sentences (2 sentence lengths × 8 accented conditions × 2 sentence types × 4 focus conditions) were elicited, as shown in **Table 12**. The accent condition of the target words is based on two website, namely the Online Japanese Accent Dictionary (Hirano et al., 2013; Nakamura et al., 2013) and an online pedagogical dictionary of Japanese

pitch accent[21]. The stimuli were checked by a phonetician who is also a native speaker of Japanese[22]. There are four possible focus conditions for each target sentence, namely initial, medial, final, and neutral. Each target sentence was repeated five times. The speaker was to produce the leading question and the statement in pair. From the 6400 utterances collected, a further 149 had to be discarded due to mis-production of accent condition. A total of 6251 were retained for statistical analysis. For a more focused discussion, in this Chapter only the subset with statements is reported.

		Word I		Word II		Word III		
Short	Accented	**mei**-ga May が May-NOM	×	**mo**mo 腿 thigh	×	-o **mi**ta を見た -ACC saw	×	?
	Unaccented	mei-ga 姪が Niece-NOM		momo 桃 peach		-ni nita に似た -DAT resembled		。
Long	Accented	**muu**min-ga ムーミンが Moomin-NOM	×	**bu**dou 武道 martial arts	×	-o **mi**ta を見た -ACC saw	×	?
	Unaccented	noumin-ga 農民が Farmer-NOM		budou 葡萄 grapes		-ni nita に似た -DAT resembled		。

Table 12*. Corpus used in Chapters 4 and 5.*

Several methodological issues were taken into consideration when designing these stimuli. First, initial nasal consonants were used as far as possible, to facilitate subsequent segmentation of data and that a continuous F0 trajectory was tracked. Second, minimal pairs that contrast only in accent condition were used to avoid confounds from consonantal perturbation (e.g. Hombert, Ohala, & Ewan, 1979) and vowel intrinsic F0 (e.g. Sapir, 1989; Shadle, 1985). As a result of these two strategies, a part of the stimuli were meaningless (e.g. 'Moomin watched martial arts') or ungrammatical (***-ni nita* → *-ni niteita* 'resembled'). Another setback of these stimuli is, as will become clearer in Chapter 5, that yes/no questions in Japanese involve a raised F0 on the verb (Kori, 2013; Maekawa, 1991), which could be confusable with a narrow focus. While acknowledging these shortcomings, these strict minimal pairs serve to illuminate the role that the accent-focus interaction plays in the realisation of PFC. Needless to say, when interpreting the results from these target sentences, one needs to check against the findings from previous studies to assure that any observations are not peculiar to the present stimuli.

Including more accent (i.e. non-initial accents) and sentence length (i.e. four word sentences or more) conditions would have benefited this study but this was logistically unfeasible. The 10 speakers reported here took approximately 50 minutes to complete the production task; adding more parameters would make the experiment exponentially longer and

[21] http://accent.u-biq.org/

[22] The consultant confirmed that the stimuli could elicit the intended accent conditions, but warned that some of the stimuli were either ungrammatical or meaningless.

exhausting. Again, these factors have been controlled for in previous studies (e.g. A. Lee & Xu, 2012 for four word sentences and; Sugahara, 2003a for non-initial accented words), whose findings will be taken into consideration when interpreting the results of this study.

Moreover, although it has been suggested (Kubozono, 2007) that lab speech, especially one that elicits focus with information contrast, is more exaggerated and yields extra boosted foci, I decided to adhere to the present procedure. Eliciting focus this way ensures that true minimal contrasts between questions vs. statements are obtained, which will be the focus of Chapter 5. Though Kubozono's (2007) using WH-words may yield more natural utterances, such an approach would not give us truly minimal contrasts of sentence types and focus conditions.

4.2.2 Recording procedure

Recording took place in a sound-proofed room in University College London, using a RØDE NT1-A microphone placed approximately 30 cm away from the speakers. Participants were seated in front of a computer, which displays one question-statement pair at a time. The stimuli were presented in random order, and the repetitions were collected over five random occasions. Stimuli were presented in standard Japanese orthography (mixed use of hiragana, katanana, and kanji), with the focused item underlined and boldfaced. Participants were given oral instructions about the task, and time to practice before recording began. All participants were briefed about the experiment and granted their written consent to being tested. Speakers were then interviewed about their linguistic background and history of speech and hearing impairment. All speakers were remunerated for their time.

4.2.3 Processing of data

The raw sound data were first chunked into individual utterances, and subsequently segmented by the mora on Praat. Heavy syllable were segmented into two morae equal in duration. Vocal pulse markings were manually checked and rectified. The segmented data were then fed into ProsodyPro (Xu, 2013) to extract acoustical measurements for further analyses.

4.3 Results

In this Section, results of several statistical analyses are presented for each of the three (initial, medial, final) focus conditions. Five dependent variables were analysed, namely **MaxF0** (maximum F0 of word), **MeanF0** (mean F0 of word), **MinF0** (minimum F0 of word), duration

(word duration in ms), and intensity (mean intensity of word in dB). A repeated measures ANOVA was conducted on each of these variables separately to compare to effects of focus condition (narrow focus vs. neutral focus) and accent condition of Word 1, Word 2 and Word 3. Measurements of these variables were taken in each of the focus trizones (on-focus, pre-focus, post-focus). For, example, to examine the effect of (initial) focus on post-focus **MeanF0**, the mean F0 of Word 2 is compared between initial focus and neutral focus. Likewise, the effect of (final) focus on pre-focus duration is examined by comparing the duration of Word 2 in final focus and neutral focus.

4.3.1 Initial focus

As found in previous studies, there is on-focus raising of F0; both **MaxF0** and **MeanF0** of the focused word were higher than their neutral focus counterpart. The peak of the green curves in **Figure 26** are higher than that of the black curves. When focus is accented (see green curve in panals 111S and 121S), the F0 peak of Word 2, and in most cases of Word 3 as well, are lower than that of the neutral focus counterpart (black curve). This is a typical case of post-focus compression of F0 range in the general sense of F0 maxima less minima. On the other hand, when focus is unaccented (panals 211S and 221S), Word 2 always starts higher than neutral focus, but the two focus conditions will converge eventually. The time required for this convergence appears to be a matter of actual time rather than of prosodic structure — we will return to this issue with more details in the discussion in §4.4.1.

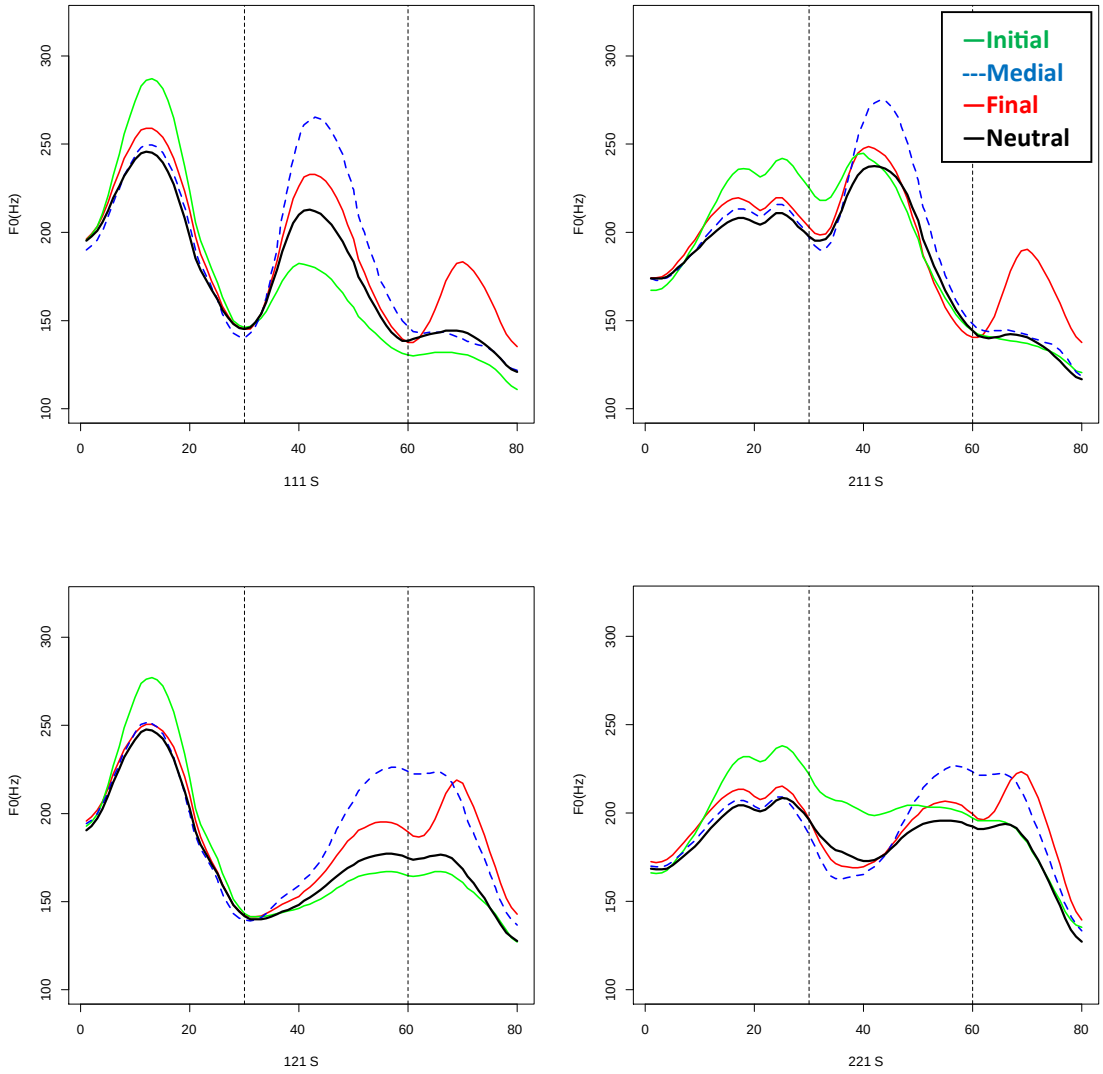


***Figure 26**. Averaged F0 contours (across 50 repetitions each) of four sentences: **mei**-ga **mo**mo-o **mi**ta 'May looked at the thigh' (111S), **mei**-ga momo-o **mi**ta 'May looked at the peach' (121S), mei-ga **mo**mo-o **mi**ta 'the niece looked at the thigh '(211S), and mei-ga momo-o **mi**ta 'the niece looked at the peach' (221S). X-axis shows normalised time, whereas vertical lines represent word boundaries. Line colour represents different focus conditions.*

Repeated measures ANOVA shows significant main effects of focus on **MaxF0** ($F(1,9) = 61.90$, $p<0.001$) and **MeanF0** ($F(1,9) = 91.36$, $p<0.001$). Post-hoc pairwise comparisons also confirm that on-focus MaxF0 and MeanF0 are higher than neutral. Focus also has a significant effect on on-focus duration ($F(1,9) = 56.51$, $p<0.001$), while its effect on on-focus intensity is but marginal ($F(1,9) = 4.37$, $p = 0.066$), i.e. an accented focus is only louder than neutral by 0.996 dB (see **Figure 27**), contra our observation in Lee & Xu (2012). On-focus duration is longer than neutral by 13.3 ms (post-hoc Bonferroni test, p <0.001), contra Maekawa's (1997) observation of largely unchanged duration of the focused item.

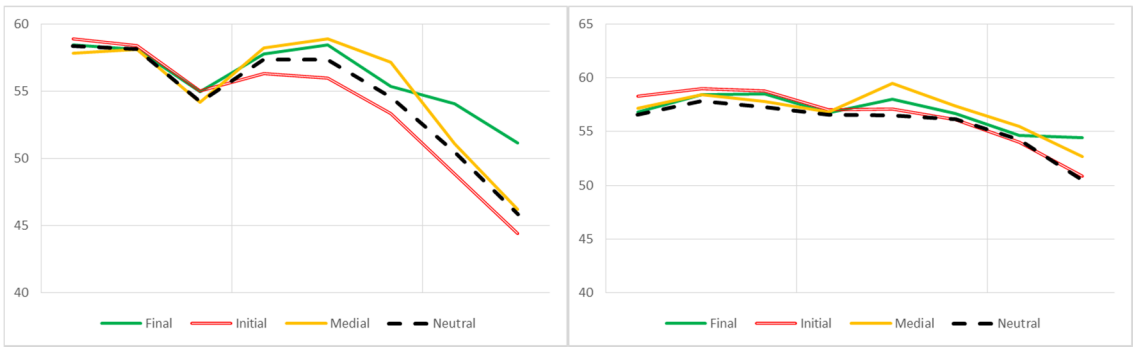


***Figure 27**. Averaged intensity profile of **me**i-ga **mo**mo-o **mi**ta 'May looked at the thigh' (left) and mei-ga momo-ni nita 'the niece resembled a peach'(right). X-axis is normalised time, while Y-axis show intensity (dB).*

Somewhat surprisingly, there is no significant main effect of focus on any of the post-focus measurements, but a significant interaction of focus and the accent condition of Word 1 on post-focus **MaxF0** ($F(1,9) = 34.01$, $p<0.001$), **MeanF0** ($F(1,9) = 32.93$, $p<0.001$) and **MinF0** ($F(1,9) = 27.81$, $p = 0.001$). Further examination of **Figure 26** reveals that post-focus F0 goes in opposite directions depending on the accent condition of the focused item; where (initial) focus is unaccented, post-focus **MaxF0** and **MeanF0** are higher than neutral counterparts, vice versa, hence masking the main effect. Note that all measurements in this study are taken within one word (i.e. three to four morae), such that the post-focus measurements of initial focus are taken from Word 2, rather than the whole of the post-focus region.

4.3.2 Medial focus

This subsection reports results of the acoustic analysis of medial focus. Because the stimuli in this study contains three words (SVO), here medial focus is equivalent to penultimate focus.

Like in initial focus, significant on-focus raising of F0 peak (blue dashed curve) is observed. Post-focus F0 movement is accent-dependent — when focus is accented (panels 112S and 332S in **Figure 28**), Word 3 appears to be indistinguishable from neutral (i.e. dashed blue vs. black curves), whereas when focus is unaccented (panels 341S and 342S in **Figure 28**) Word 3 is higher (i.e. dashed blue curve higher than black curve).

ANOVA shows significant main effect of focus on on-focus **MaxF0** ($F(1,9) = 104.04$, $p<0.001$) and on-focus **MeanF0** ($F(1,9) = 87.33$, $p<0.001$). The same effect is also significant on on-focus **intensity** ($F(1,9) = 34.72$, $p<0.001$, see also **Figure 29**) and **duration** ($F(1,9) = 313.99$, $p<0.001$).

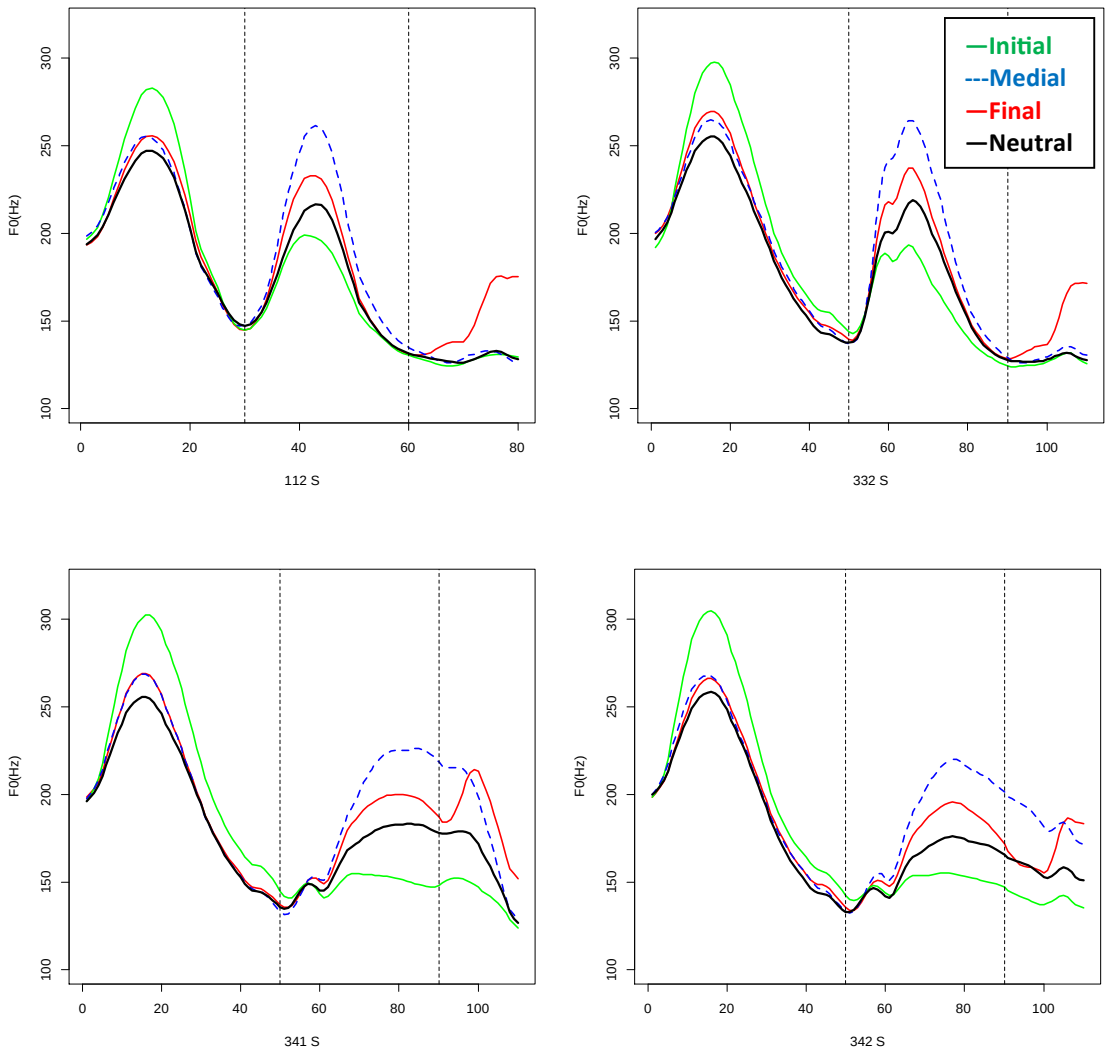


***Figure 28**. Averaged F0 contours of **me**i-ga **mo**mo-ni nita 'May resembled the thigh' (112S), **mu**umin-ga **bu**dou-ni nita 'Moomin watched martial arts' (332S), **mu**umin-ga budou-o **mi**ta 'Moomin looked at the grapes' (341S) and **mu**umin-ga budou-ni nita 'Moomin resembled the grapes' (342S) in four focus conditions.*

It is also observed that **MaxF0** ($F(1,9) = 38.90$, $p<0.001$), **MeanF0** ($F(1,9) = 38.54$, $p<0.001$) and **MinF0** ($F(1,9) = 7.17$, $p = 0.025$) are higher after focus in general. Post-hoc test shows that **MaxF0** and **MeanF0** are respectively 1.602 semitones and 1.267 semitones higher after focus than neutral, which is unexpected given the known effect of PFC. There is also post-focus **intensity** raising ($F(1,9) = 10.01$, $p = 0.011$). Judging from the results in initial focus, I then examined the interaction between focus and accent condition of Word II (the focused item), which turned out to be highly significant, post-focus **MaxF0** ($F(1,9) = 47.72$, $p<0.001$), **MeanF0** ($F(1,9) = 21.86$, $p = 0.001$), and **MinF0** ($F(1,9) = 7.22$, $p<0.025$). Post-focus **MaxF0** is only 0.103 semitones higher than neutral when focus is accented, but 3.103 semitones higher when focus is unaccented. The same pattern is also found in **MeanF0** and **MinF0** after focus. This

confirms that F0 is indistinguishable after an accented focus[23], but higher after an unaccented focus.

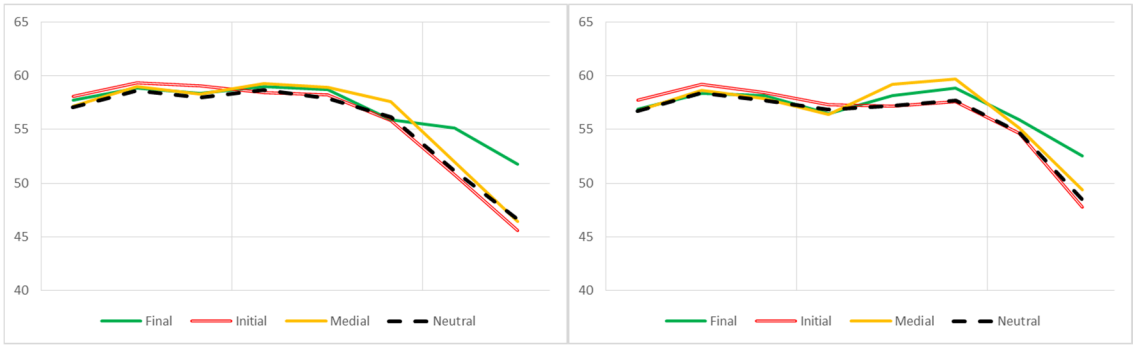


***Figure 29**. Averaged intensity profile of mei-ga **mo**mo-o **mi**ta 'the niece looked at the thigh' (left), mei-ga momo-o **mi**ta 'the niece looked at the peace' (right).*

Before focus, lengthened duration ($F(1,9) = 21.70$ $p = 0.001$) is observed. Other than that, focus has no significant main effect on any of the pre-focus measurements.

4.3.3 Final focus

When focus is sentence-final, there is on-focus raising of MaxF0 ($F(1,9) = 26.17$, $p = 0.001$), MeanF0 ($F(1,9) = 27.54$, $p = 0.001$) and MinF0 ($F(1,9) = 15.49$, $p = 0.003$), comparable to on-focus effects in other focus conditions. Also, the focused item is longer in duration ($F(1,9) = 56.27$, $p<0.001$) and higher in intensity ($F(1,9) = 73.425$, $p<0.001$, see also **Figure 31**), compared to its neutral counterpart.

[23] This is believed to be a property of penultimate focus in SOV languages. See discussion in §4.4.

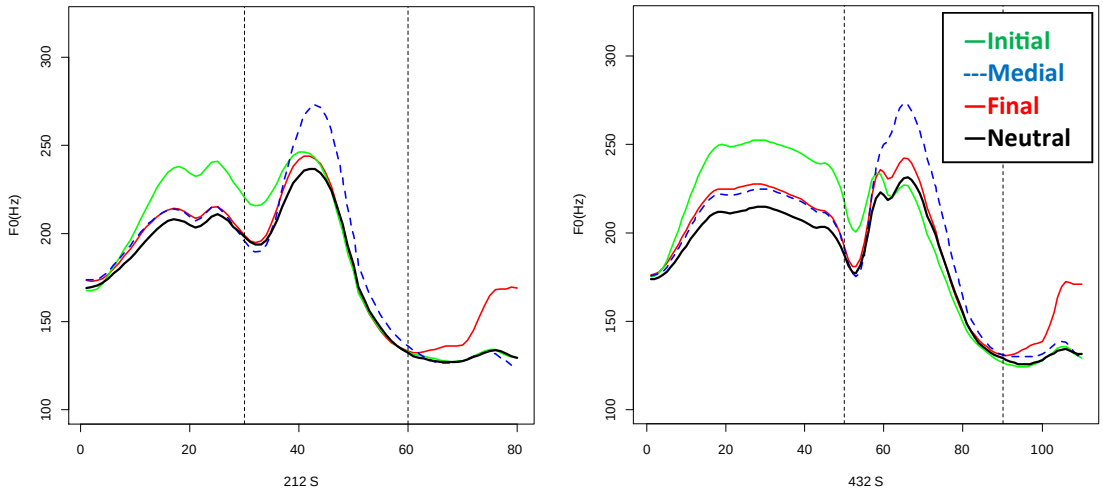


***Figure 30**. Averaged F0 contours of mei-ga **mo**mo-ni nita 'the niece resembled the thigh' (212S), and noumin-ga **bu**dou-ni nita 'the farmer resembled martial arts' (432S) in four focus conditions.*

There are also considerable pre-focus enhancement effects. Both **MaxF0** (F(1,9) = 16.50, *p* = *0*.003) and **MeanF0** (F(1,9) = 11.06, *p* = *0*.009) are higher before focus, forming multiple peaks across the utterance (see any solid red curve in **Figure 30**, compared with corresponding black curves). Besides, the pre-focus item is louder (F(1,9) = 10.087, *p* = *0*.011) and longer in **duration** (F(1,9) = 61.80, *p*<*0*.001) than the corresponding position under neutral focus.

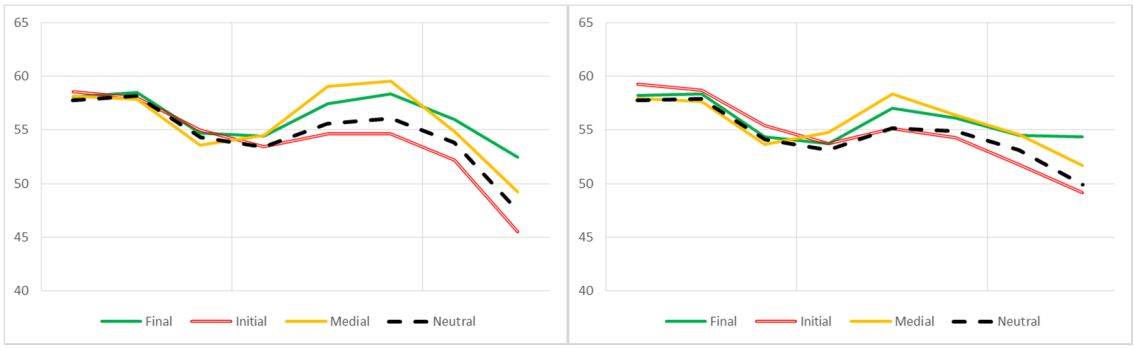


***Figure 31**. Averaged intensity profile of **me**i-ga momo-o **mi**ta 'May looked at the peach' (left) and **me**i-ga momo-ni nita 'May resembled a peach' (right).*

An interesting interaction to note, though not directly relevant to the research question, is the interaction between focus and the accent condition of Word I. This interaction effect is statistically significant on on-focus MaxF0 (F(1,9) = 5.56, *p* = *0*.043) and MeanF0 (F(1,9) = 16.49, *p* = *0*.003). This means that when Wd I is accented, Wd III (the focused item) is lower than when Wd I is unaccented, a typical piece of evidence of downstep. We shall return to this issue in the Discussion below.

4.4 Discussion

4.4.1 Conditional realisation of PFC

This Chapter set out to examine the realisation of PFC under various accent and focus conditions. The findings generally agree with previous studies, confirming the effectiveness of the elicitation method. Where PFC is taken to mean maxima less minima within a given post-focus word, it is observed after an accented focus; when the focused word is unaccented, F0 is higher after focus to a level much higher than neutral (cf. Ito, 2002b). But it is also observed that the F0 curves of initial focus and neutral focus converge, at the end of the short statement (e.g. panel 111S in **Figure 26** in Page 73) and in the middle of the long statement (e.g. panel 331S in **Figure 34** in Page 81). This leads us to believe that (i) PFC is conditionally realised after a pitch accent, and (ii) elsewhere F0 has a common underlying target with neutral focus but is subject to strong carryover effect and weak articulatory strength.

The first claim of this Chapter is a quasi-reiteration of Ishihara (2011a), that the realisation of PFC requires at least one pitch accent in the preceding domain. However, it is unclear to us if lowering could be due to a preceding pitch accent that is two words away or more (e.g. green curve of panel 341S in **Figure 28**). That is, while the lowering of the green curve in Word 2 is a clear case of PFC, which emerges owing to the pitch accent in Word 1, it is not possible to prove that the lowering in Word 3 is also due to the same pitch accent; it could possibly be a mere continuation (carryover) from the final F0 value of Word 2, which is related to the next claim.

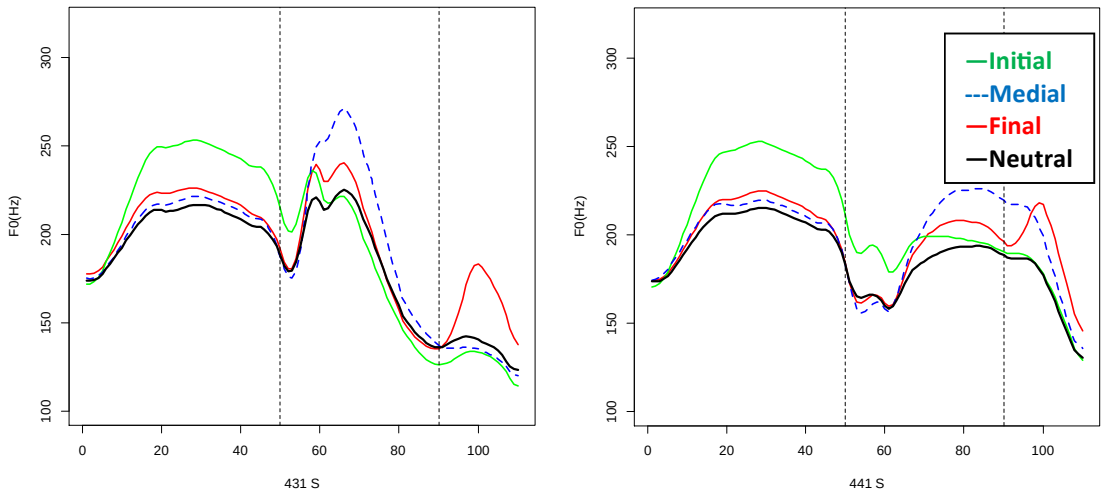


***Figure 32**. Averaged F0 contours of noumin-ga **bu**dou-o **mi**ta 'the farmer watched martial arts' (431S), and noumin-ga budou-o **mi**ta 'the farmer looked at the grapes' (441S) in four focus conditions.*

The second claim is that in cases where PFC is not realised, pitch targets in the post-focus region are the same as neutral focus, although they bear weaker articulatory strength. When there is no preceding pitch accent, PFC is absent. The apparent 'post-focus raising' (cf. Ito, 2002b) is a carryover effect from the previous word that ends at a high F0, due to on-focus raising (e.g. panel 441S in **Figure 32**). But note that in all of these cases, when the post-focus domain is sufficiently long (i.e. for initial accent), the 'raised' post-focus contour of initial focus eventually converges with the neutral contour.

Here one may argue that the convergence at the end of cases like 222S in **Figure 33** is due to the boundary tone L% (in J-ToBI terms) that ends an utterance, but 442S suggests otherwise. In 442S (which differs from 222S only in terms of word length and in turn total duration), initial focus and neutral focus converge in the middle of Word 2 in 442S. On the other hand, in 222S, the two focus conditions do not converge until the end of the utterance. In other words, in both 222S and 442S, initial focus and neutral focus are approximating the same target in the post-focus region, but they meet 'earlier' (i.e. in the middle of the utterance) in the longer sentence (442S) and later (i.e. at the end of the utterance) in the shorter sentence, suggesting that the convergence requires time to take place rather than occurring at a particular location in the prosodic structure (e.g. end of utterance). This is reminiscent of the phonetic realisation of neutral tone in Mandarin (Chen & Xu, 2006) which takes several syllables before F0 contours of different tonal conditions finally converge. Like the Mandarin neutral tone, I argue that the sluggish approximation to target after an unaccented focus is due to weak articulatory effort. By extension, this property of unaccented words could also be interpreted as evidence that unaccented words are phonologically weak and thus do not allow certain focus markers to realise, like in a recent study on French prosodic focus (Turco, Dimroth, & Braun, 2012) .

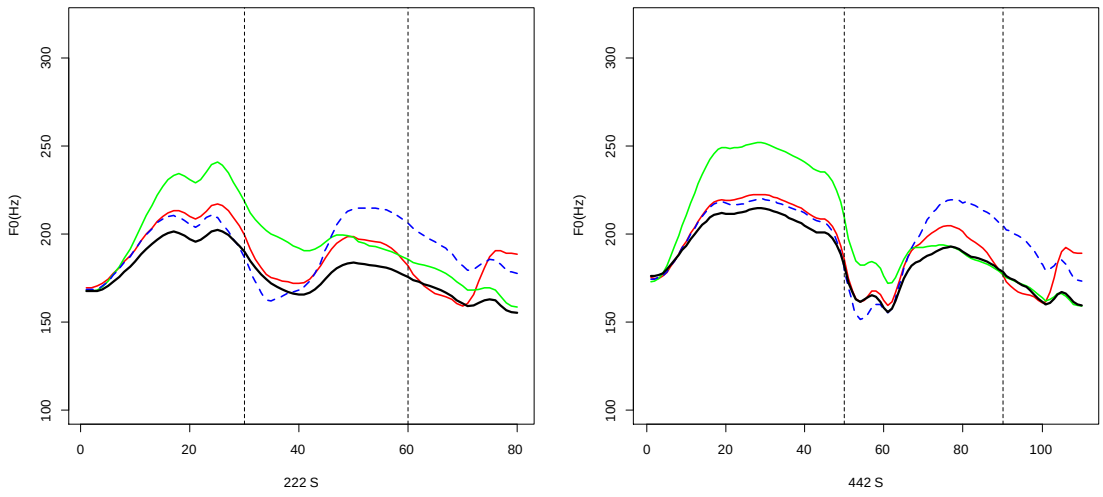


***Figure 33**. Averaged F0 contours of mei-ga momo-ni nita 'the niece resembled a peach' (222S) and noumin-ga budou-ni nita 'the farmer resembled grapes' (442S) in four focus conditions.*

The main reason that the carryover account for unaccented utterances is preferred comes from the fact that 'PFC' in this case is only realised within initial rise. As mentioned earlier, initial rise consists of the boundary tone %L and the phrasal tone H-, the former of which is also part of the preceding word. Thus the under-realisation of the initial rise (i.e. 'PFC') would likely be the artefact of the transition from the high final F0 of the preceding unaccented word. If PFC is indeed at work, one should be able to observe clear reduction effects like compression or lowering on the pitch targets per se, realised across an entire post-focus word rather than only at the beginning; nevertheless it is absent. The lack of PFC in an unaccented word is not due to its lack of sharp turning points either, because in the high flat tone (Tone 1) in Mandarin, PFC is clearly realised as mean F0 lowering; that F0 is higher in Japanese instead could only mean PFC is not at work. In other words, the absence of PFC in this case is language-specific rather than due to the F0 profile of the accent condition in question. Finally, this carryover account has an additional advantage of explaining the post-focus dephrasing effect (Sugahara, 2002) in unaccented words without resorting to formal grammatical notions. Where both the focused item and the post-focus item are unaccented, the initial rise in the latter is reported to be absent or weakened, hence the name 'dephrasing' because initial rise marks the beginning of a phrase – this account account offers an articulatory explanation to this well-established phenomenon.

On the other hand, PFC is not a carryover effect from a low final F0 of an accented word; in panels 341S and 342S in **Figure 28**, for example, where the green (initial focus) and black (neutral focus) curves in Word 2 (post-focus domain) start from the same point and only differ in terms of peak height. Thus PFC is a functional strategy by choice, to lower post-focus pitch targets, accented and unaccented alike. Then where is pitch accent in this articulatory strength account, apart from giving rise to PFC? It disrupts the articulatory strength profile, i.e. it draws the initial and neutral focus curves together sooner than otherwise. In panel 431S in **Figure 32**, although after an unaccented focus F0 initially rises, soon after the accent peak (circa time point 65) the green and black curves start to overlap. Note that in this account there is no PFC after an unaccented focus, so Word II of the green and black curves in panel 431S are deemed to bear the same target. As is clear from **Figure 32**, the two curves converge after two morae, whereas in 441S it takes more than four. I thus contend that an accented word has greater articulatory strength than an unaccented word in the post-focus domain, which allows targets to be reached sooner. More evidence in support of this claim will be presented in §6.3.

At this point, one may question the logic of my claim: if one accepts that there is no PFC in an exclusively unaccented statement, cases of apparent PFC would coincide with downstep, which also has the effect of lowering subsequent accent peaks. If the realisation of PFC is not tone-bound in Mandarin, there is no reason why it is in Japanese. Then, should one as well claim that typologically Japanese does not have PFC, but downstep that yields all the post-focus lowering of F0 peaks? The answer is negative: there is clearly downstep effect in a statement under neutral focus (e.g. panel 111S in **Figure 26**, Page 73), then to claim that there is extra downstep after a narrow focus would be logically no different from naming it PFC,

unless there is physio-articulatory reasons that predict so. This also agrees with previous research (S. Ishihara, 2007; Sugahara, 2003a) in which PFC and downstep were shown to be separate phenomena. Hence, I maintain that Japanese does have PFC, only that it is conditionally realised, and this Chapter has demonstrated that pitch accent is one such condition.

4.4.2 Indistinguishable cases?

In §4.3.2 it was found that PFC was absent even after an accented focus, like panel 331S in **Figure 34** where the blue curve (medial focus) and the black curve (neutral focus) overlap in Word III. That an accented medial focus sees no PFC or raising echoes with our finding in Lee & Xu (2012) and in Ito (2002a), and could be attributed to the fact that Japanese is a SOV language. In Turkish, also an SOV language, F0 is not lowered after a medial (i.e. penultimate) focus (Ipek, 2011). Even in Korean (Y.-C. Lee & Xu, 2010), where PFC is observed, the effect size of focus is much smaller in medial (i.e. penultimate) focus than in initial focus. Ipek (2011) suggests that the object-verb sequence forms a naturally falling contour that resembles post-focus lowering, hence the absence of PFC effect for penultimate focus. Similarly, the same prediction by Focus Projection (e.g. Chomsky, 1972; Gussenhoven, 1999), whereby prosodic prominence on one word could either mark focus on the same word or a larger phrase containing that word, has also been discussed by Venditti and colleagues (2008) with regards to Japanese.

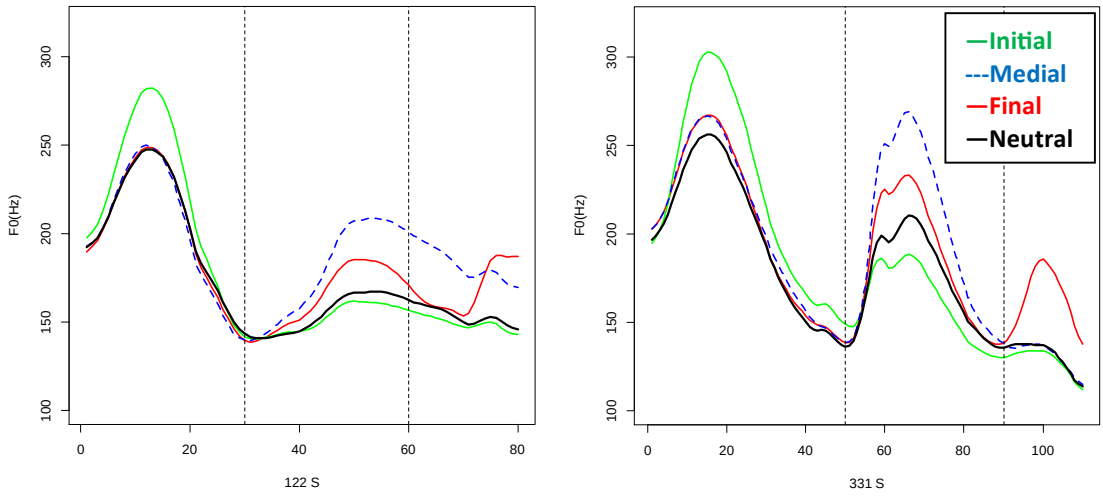


***Figure 34**. Averaged F0 contour of **mei**-ga momo-ni nita 'May resembled a peach' (122S), and **mu**umin-ga **bu**dou-o **mi**ta 'Moomin watched martial arts' (331S) in four focus conditions.*

I mentioned that the Intonational Phrase boundary insertion hypothesis would predict that initial focus is indistinguishable from neutral focus[24] (Venditti et al., 2008), but this is not supported by the data (e.g. panel green curve vs. black curve in 122S, **Figure 34**). Initial focus has a significantly higher F0 peak than neutral in all cases. Perhaps the laboratory nature of this production task led speakers to exaggerate their initial focus peak, and it remains an open question whether the extra high peak makes initial focus more easily identifiable in perception.

4.4.3 No pre-focus reduction

The raising of F0 and intensity as well as the lengthening of duration before focus are surprising, and disagree with previous findings (e.g. Hwang, 2011; A. Lee & Xu, 2012; Maekawa, 1997). The discrepancies can be attributed to the stimuli used here, and possibly the fact that I am using mostly nonsense sentences (the only speaker who shows consistent pre-focus lowering, KT, is a trained voice-over artist). Then, looking at the big picture, pre-focus enhancement may be seen as an optional strategy for unfamiliar utterances — faced with an unfamiliar sentence, speakers can only assure that the salient focus cues are correctly produced, i.e. on-focus and post-focus, and inadvertently emphasise the rest, namely pre-focus. An alternative account would be that Intonational Phrase boundary insertion before the focused item triggers final lengthening in the pre-focus item. To test the validity of the latter hypothesis, a more systematic study using natural stimuli is needed to investigate pre-focus durational modification. Whilst pre-focus modifications have been found to be an optional strategy for marking focus (A. Lee & Xu, 2012), it is unclear whether the intended focus condition of such multiple-peak utterances will be accurately perceived.

The interaction between focus and the accent condition of Word 1 shows that on-focus raising of F0 peak is not a 'reset' (Pierrehumbert & Beckman, 1988) but a local boost (Kubozono, 2007), which is subject to downstep from preceding words. This is noteworthy because my agreement with Kubozono is in spite of the fact that I am using lab speech, which is potentially 'exaggerated'.

4.4.4 Non-F0 focus markers

Regarding non-F0 focus markers, my data showed mixed results. Among all the measurements, only on-focus lengthening (duration) reached statistical significance across all focus conditions, echoing early work on English focus (Eady & Cooper, 1986) where durational correlates were reported to be generally localised (i.e. only on-focus). Meanwhile, intensity is not always raised

[24] See Ishihara (2011b) for a more detailed discussion on this issue and his Optimality Theoretic account.

on-focus — in §4.3.1 initial focus did not have a significant main effect on on-focus intensity, despite significant on-focus raising of MaxF0. One possible reason is that Word I is quite loud already even under neutral focus, hence the speakers have little room to further raise intensity.

There was no shortening before or after focus, contra previous work on English (Weismer & Ingrisano, 1979) and Japanese (Maekawa, 1998). Moreover, there was even pre-focus lengthening for final focus, contradicting Maekawa's (1998) observation of pre-focus shortening. But such a discrepancy is not totally surprising — even in English, findings on non-F0 modifications before and after focus have been inconclusive (e.g. Eady & Cooper, 1986; vs. Weismer & Ingrisano, 1979), suggesting that non-local non-F0 cues are not primary focus-marking strategies employed by all native speakers of Japanese.

This study cannot solve all the puzzles in PFC. Some issues remain and cannot be concluded with the present production data. For example, whether certain focus conditions are easily confusable with neutral and whether there is pre-focus F0 modificaiton cannot be hastily concluded with only production data. As a next step, a perception study is under way to examine the respective ease of focus identification in each of the communicative conditions studied in this work, and will hopefully draw a clearer picture of the complex interplay between pitch accent and focus in Japanese.

4.5 Chapter conclusion

This Chapter reports a production study of Japanese focus prosody that comprehensively controls for accent condition, focus condition and word length. The findings generally agree with previous research, that there is on-focus enhancement and post-focus reduction in various forms. However, based on the data, I propose that PFC is only realised after a pitch accent, and does not include cases of compressed initial rise in exclusively unaccented utterances. Those cases are, I argue, a result of carryover effect from the preceding on-focus item, thus the apparent 'post-focus raising'. There is also the higher contour after an unaccented focus gradually converges with the neutral focus contour, and posit that these two focus conditions have the same underlying pitch target, although the post-focus contour approximates this target at a low articulatory strength.

CHAPTER 5: INTERROGATIVE FOCUS PROSODY

5.1 Introduction

Sentence type and focus are two major communicative functions conveyed by intonation. The two functions are closely connected in Japanese in that a lot of studies on Japanese focus have actually used wh-questions to elicit focus (e.g. S. Ishihara, 2011a; Kubozono, 2007). Now that we have a better understanding of focus, and of question intonation in general from previous literature, our next destination would be how these two functions interact. Through this new corpus, in this Chapter I investigate the acoustic properties of interrogative focus prosody to complete the picture of the basic aspects of Japanese sentential prosody.

An interrogative utterance in Japanese is generally marked by a final rise, as opposed to final lowering that marks a declarative sentence (Beckman & Pierrehumbert, 1986a). It is represented as an utterance-final H% in the J-ToBI annotation convention (Venditti, 2005). Pierrehumbert and Beckman (1988) describe the H% **as an additional tone** that follows L% (which marks the end of a declarative utterance), although the dip from the resulting L%H% sequence is not always obvious. Gussenhoven (2004, p. 202), on the other hand, describes the question final rise as a H_v **in replacement of** the statement marking L_v tone.

Recently Kori (2013) conducted a comprehensive phonetic study on question intonation in Japanese. He reports that a yes-no question has 'a straight or concave-up pitch rise that begins in the final mora', lengthening of the final mora (40-100 ms depending on the verb's lexical accent), delayed accentual fall where the verb is accented, pre-final shortening (10% reduction in duration) as well as expanded pitch range. The striking similarity between focus and interrogation markers begs the question of how questions and statements respectively under narrow and neutral focus differ.

5.2 Methods

The speech data reported in this Chapter is a subset of the corpus described in §4.2.1. Here I provide an acoustic analysis of 3136 interrogative utterances that vary in accent condition, focus condition, and utterance length. Of the 3200 utterances originally recorded, 64 were discarded due to misproduction of accent condition. See §4.2.1 for details of the production data.

As mentioned in§4.2, unnatural stimuli were used in this study for several reasons. First, to make sure that questions and statements are equal in length (in terms of mora count), no question particles (e.g. *no* の) were used after questions. Second, focus was elicited by underlining and boldfacing the word of interest, without providing context to the speaker. This is because by collecting 640 utterances from each speaker, the experiment on average lasted 50

minutes (excluding ethics procedures); providing context before each elicitation would mean an even longer test that is physically too tiring. This elicitation method could cause unnatural production, but as will be demonstrated§5.3 all focus conditions were clearly distinguished by the speakers. Also, in an experiment on Hong Kong English (Fung & Mok, 2014), prosodic focus was effectively elicited by boldfacing the focused word, like in this Chapter. Finally, yes/no questions in Japanese involves a raised F0 on the verb (Kori, 2013; Maekawa, 1991), which could be confusable with a narrow focus. However, in the following Section this study shows that interrogative final focus and interrogative neutral focus are acoustically distinct in our data.

5.3. Results

5.3.1 Initial focus

Comparing questions under initial and neutral foci, repeated measures ANOVA shows that there is significant on-focus raising of **MaxF0** $F(1,9) = 29.50$ $p = <0.001$ and **MeanF0** $F(1,9) = 20.22$ $p = 0.001$. For example, in **Figure 35**, the green curves (initial focus) are always higher than the black curves (neutral focus) in Word I, i.e. on-focus raising. However, the effect of focus is non-significant on post-focus **MaxF0**, **MeanF0** and **MinF0**. Likewise, focus has no significant interaction with any of the post-focus F0-related dependent variables. Apart from F0, focus also leads to significant on-focus **duration** lengthening ($F(1,9) = 39.58$ $p<0.001$), while its effect on post-focus **duration** is non-significant. Effect of focus on **intensity** was not observed, like in declarative utterances in §4.3.1.

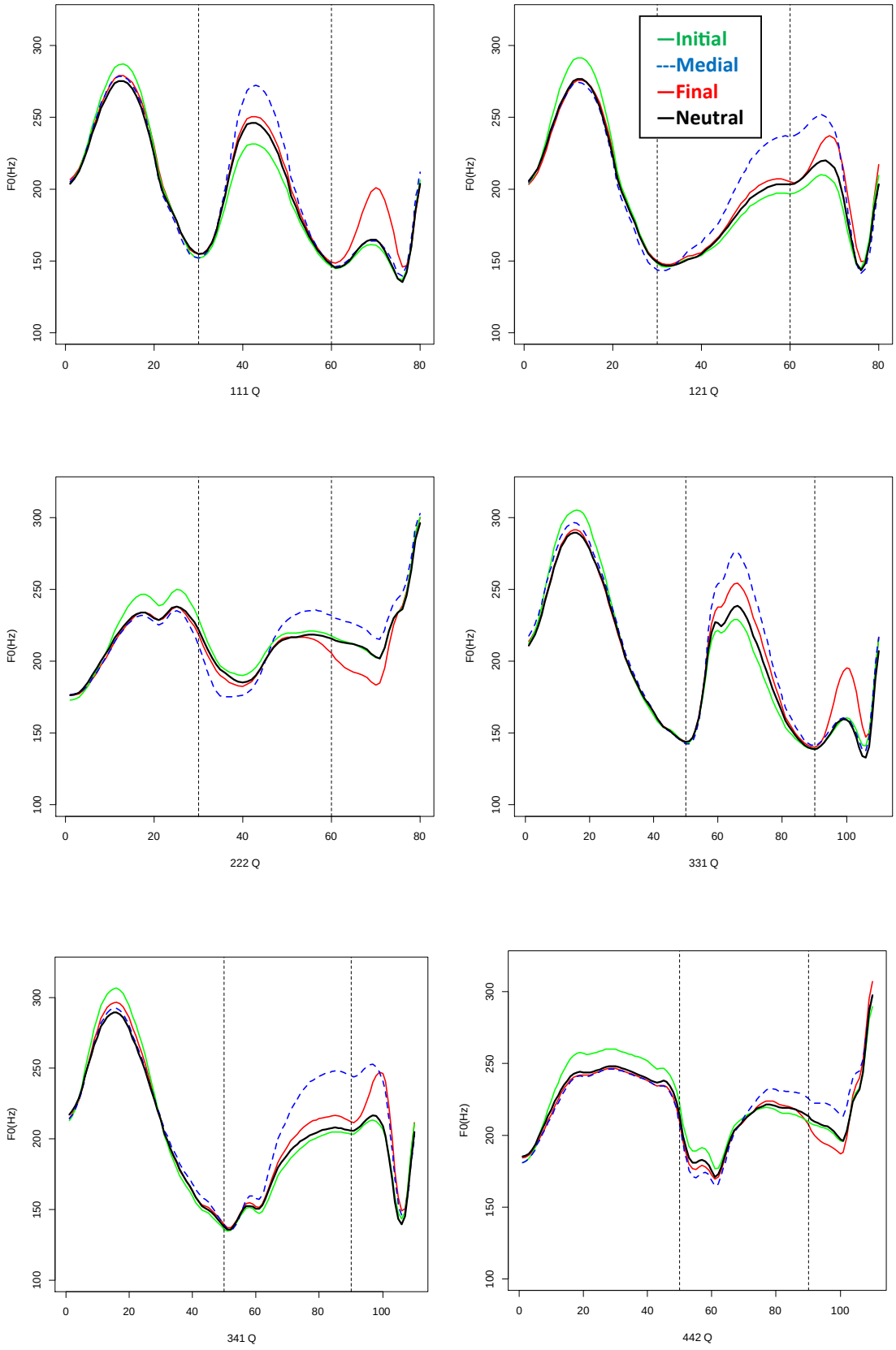


***Figure 35**. Averaged F0 contours of **me**i-ga **mo**mo-o **mi**ta? 'did May look at the thigh?' (111Q), **me**i-ga momo-o **mi**ta? 'did May look at the peach?' (121Q), mei-ga momo-ni nita? 'did the niece resemble a peach?' (222Q), **mu**umin-ga **bu**dou-o **mi**ta? 'Did Moomin watch martial arts?' (331Q), and **mu**umin-ga budou-o **mi**ta? 'did Moomin look at the grapes' (341Q), and noumin-ga budou-ni nita? 'did the farmer resembled grapes?' (442Q) in four focus conditions.*

Figure 35 appears to contradict the statistical analysis above in terms of post-focus F0 realisation. Initial focus (green curve) sentences have a lower F0 than neutral focus sentences (black curve) in the second interval in most of the panels. Where the focused word is accented, it appears that post-focus **MaxF0** is consistently lowered across all conditions, whereas for unaccented focus this is not always true. However, closer inspection of individual data reveals that not all speakers manifest post-focus lowering even after an accented focus. This individual variability thus explains the absence of statistical significance in post-focus **MaxF0**, **MeanF0** and **MinF0**.

5.3.2 Medial focus

For medial focus, on-focus F0 peak is consistently higher than neutral, but there is no visually discernable PFC (cf. Turkish, Ipek, 2011) in **Figure 36**. That is, the blue curve (medial focus) is higher than the black curve (neutral focus) in the second interval, but not lower than it in the third interval from the left. Note that the apparent post-focus F0 height difference in cases of an unaccented focus (e.g. panels 122Q and 342Q in **Figure 36**) is but a carryover effect just like in statements (cf. discussion in §4.4.1)–by halfway into the interval F0 contours of the two focus conditions will converge and overlap.

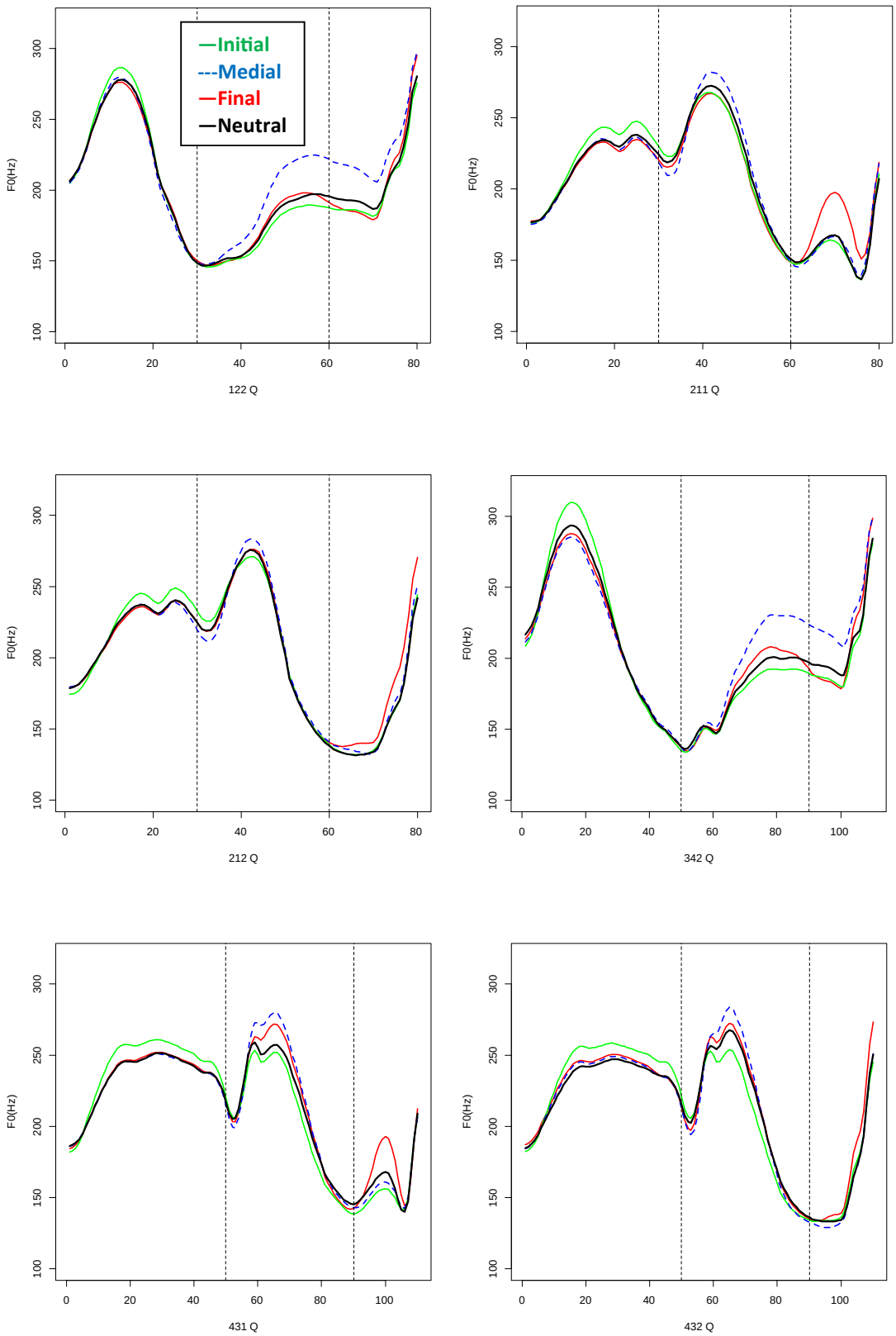


***Figure 36**. Averaged F0 contours of **mei**-ga momo-ni nita? 'did May resemble a peach?' (122Q), mei-ga **momo**-o **mi**ta? 'did the niece look at the thigh? '(211Q), mei-ga **momo**-ni nita? 'did the niece resemble the thigh?' (212Q), **muu**min-ga budou-ni nita? 'did Moomin resemble the grapes' (342Q), noumin-ga **budou**-o **mi**ta? 'did the farmer watch martial arts?' (431Q), and noumin-ga **budou**-ni nita? 'did the farmer resemble martial arts' (432Q) in four focus conditions.*

These observations are confirmed by ANOVA. There is significant main effect of focus on on-focus **MaxF0** ($F(1,9) = 27.55$, $p = 0.001$), **MeanF0** ($F(1,9) = 27.83$, $p = 0.001$), and **MinF0** ($F(1,9) = 16.37$, $p = 0.003$). Focus also has a significant effect on post-focus **MaxF0** ($F(1,9) = 30.76$, $p<0.001$) and **MeanF0** ($F(1,9) = 15.59$, $p = 0.003$). That said, sub-group means reveal that these are not evidence of PFC–where focus is accented, mean post-focus **MeanF0** is 0.013 semitones higher than neutral, but 1.157 semitones higher than neutral when focus is unaccented. This is consistent with **Figure 36**, which shows that in medial focus there is no PFC, but carryover effect that gives rise to an apparent 'post-focus raising'–the significant main effect in ANOVA is merely reflecting this 'raising' (also cf. absence of PFC under medial focus in statements).

Apart from F0, there are also raised on-focus **intensity** ($F(1,9) = 30.48$ $p<0.001$) and on-focus lengthening of **duration** ($F(1,9) = 89.95$ $p<0.001$). Seemingly there is pre-focus lengthening ($F(1,9) = 30.12$, $p<0.001$) in **duration**, though post-hoc Bonferroni comparison ($p<0.001$) shows that it is only by 3.72 ms.

Panel 431Q is the only case in the data that appears to see PFC. In Word III, the blue curve (medial focus) appears to have a lower peak (in turn smaller F0 range) than its neutral focus counterpart (black curve). However, after looking at the averaged F0 contours of individual speakers it was found that only five out of 10 manifest overall post-focus lowering; the others either showed post-focus raising or did not distinguish the two focus conditions in the post-focus domain. I am thus confident to maintain that there is no PFC in questions even after an accented focus.

5.3.3 Final focus

Final focus behaves similarly to medial focus. **Figure 37** suggests that the on-focus item has a higher F0 than the neutural focus counterpart. In all the panel, final focus (red curve) is higher in F0 than neutral focus (black curve) in Word 3. Where focus is accented (e.g. panels 221Q and 441Q), the focused item has a higher accent peak; whereas if it is unaccented (e.g. panels 112Q and 332Q) there is a larger excursion size in the question-final rise. Repeated measures ANOVA reveals that focus has a significant main effect on on-focus **MaxF0** ($F(1,9) = 22.59$, $p = 0.001$) and **MeanF0** ($F(1,9) = 14.35$ $p = 0.004$), as well as **intensity** ($F(1,9) = 50.93$, $p<0.001$) and **duration** ($F(1,9) = 15.74$, $p = 0.003$). Before focus (i.e. Word II), there are also significant raising of **MaxF0** ($F(1,9) = 5.89$, $p = 0.038$) and lengthening ($F(1,9) = 17.79$, $p = 0.002$). That said, pre-focus raised **MaxF0** is not observed when data is not normalised based on speaker mean F0, and judging from weak F value (5.89) it is not as robust as other focus cues. Meanwhile, like with medial focus, pre-focus lengthening is significant but small (3.19 ms, post-hoc Bonferroni comparison $p = 0.002$).

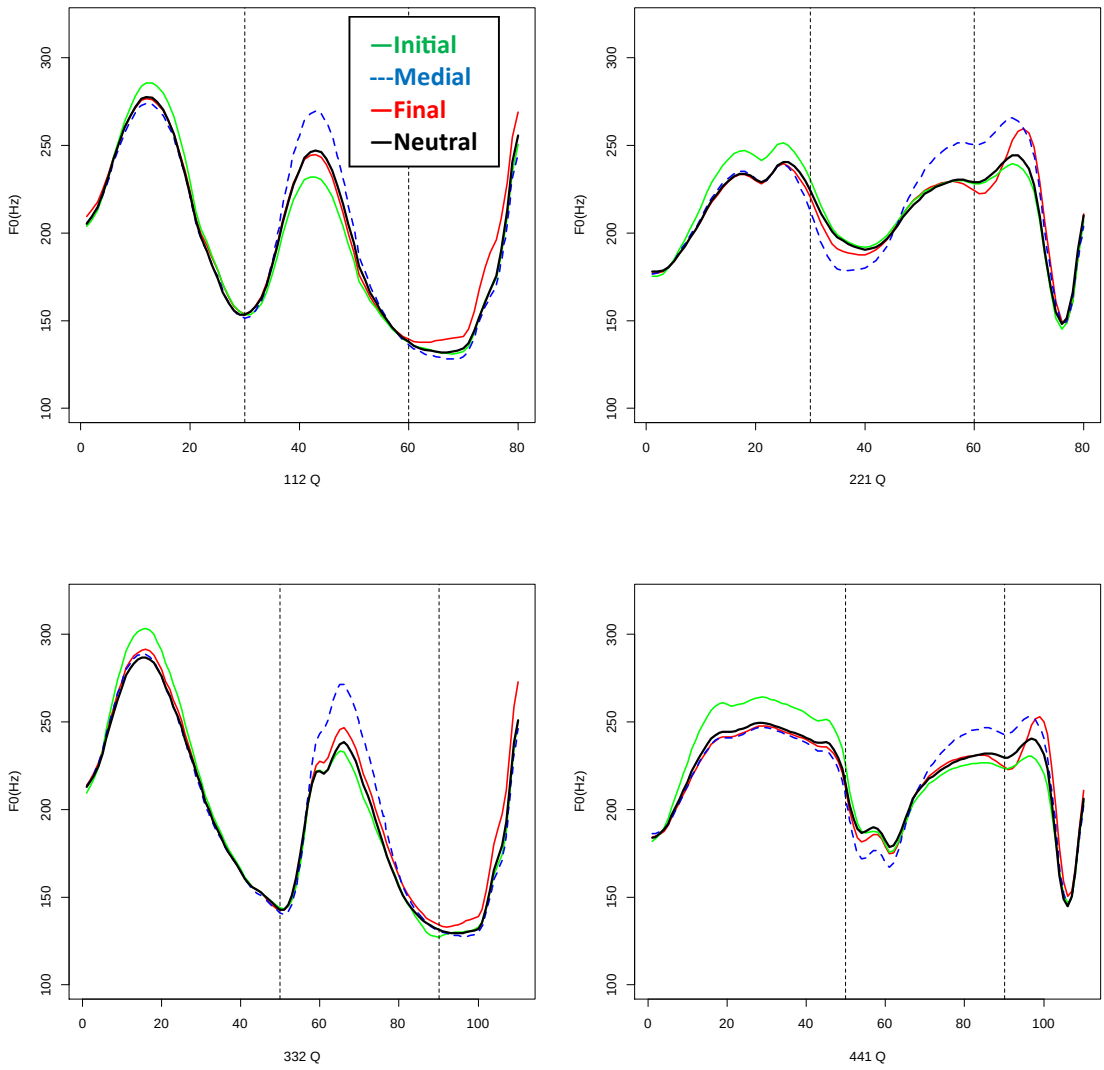


***Figure 37**. Average F0 contours of **mei**-ga **mo**mo-ni nita? 'did May resemble the thigh?' (112Q), mei-ga momo-o **mi**ta? 'did the niece look at the peach?' (221Q), **mu**umin-ga **bu**dou-ni nita? 'did Moomin watch martial arts?' (332Q), and noumin-ga budou-o **mi**ta? 'did the farmer look at the grapes?' (441Q) in four focus conditions.*

5.3.4 Effect of question

Apart from the effects of focus and pitch accent on interrogative prosody reported above, there is also the effect of sentence type, that is, the difference between questions and statements, which lies beyond the sentence-final rise. I combined the present data (questions) with that in Chapter 4 for visual comparison. As shown in **Figure 38** and Appendix 1, questions (solid curves) always have a higher overall F0 than statements (dashed), regardless of focus and accent condition.

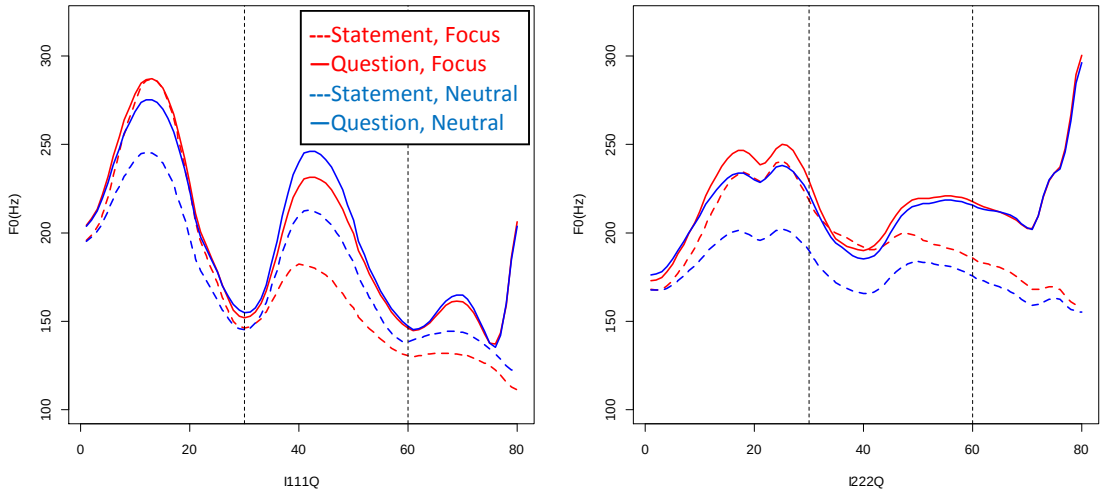


***Figure 38**. Averaged F0 contours of **me**i-ga **mo**mo-o **mi**ta 'May looked at the thigh' (IIIQ) and mei-ga momo-ni nita 'the niece resembled the peach' (222Q) spoken in question vs. statement and initial vs. neutral focus.*

Several repeated measures ANOVAs were further performed. First, across all accent, focus, and word length conditions, sentence type has a significant main effect on sentence mean F0 ($F(1,9) = 55.077$, $p<0.001$), which is 22.71 Hz higher in questions than in statements (post-hoc Bonferroni test, $p<0.001$). I then performed focus condition-specific analyses like in 5.3.1-5.3.3. Results show that in the data, 'question' enhances all measures, across all focus trizones. That is, **MaxF0, MeanF0, MinF0, duration**, and **intensity** are all greater in questions than in statements, other factors held constant. Two of the comparisons (namely on-focus duration for medial focus, and pre-focus duration for final focus) do not reach statistical significance ($F(1,9)$, alpha = 0.05), but even in these cases the general trend remains.

Given the effect of questions on the F0 level, one would expect on-focus MaxF0 of narrow focus statement, narrow focus question, and neutral focus question to be very similar, if not indistinguishable. Indeed, although ANOVA shows significant main effect of sentence type on all of the measurements, in **Figure 39** the **MaxF0** of these three focus*sentence type conditions are very close to one another. One possible explanation is that at this F0 level speakers are already close to the 'ceiling' of their F0 range, hence there is limited room left to them to distinguish the different conditions.

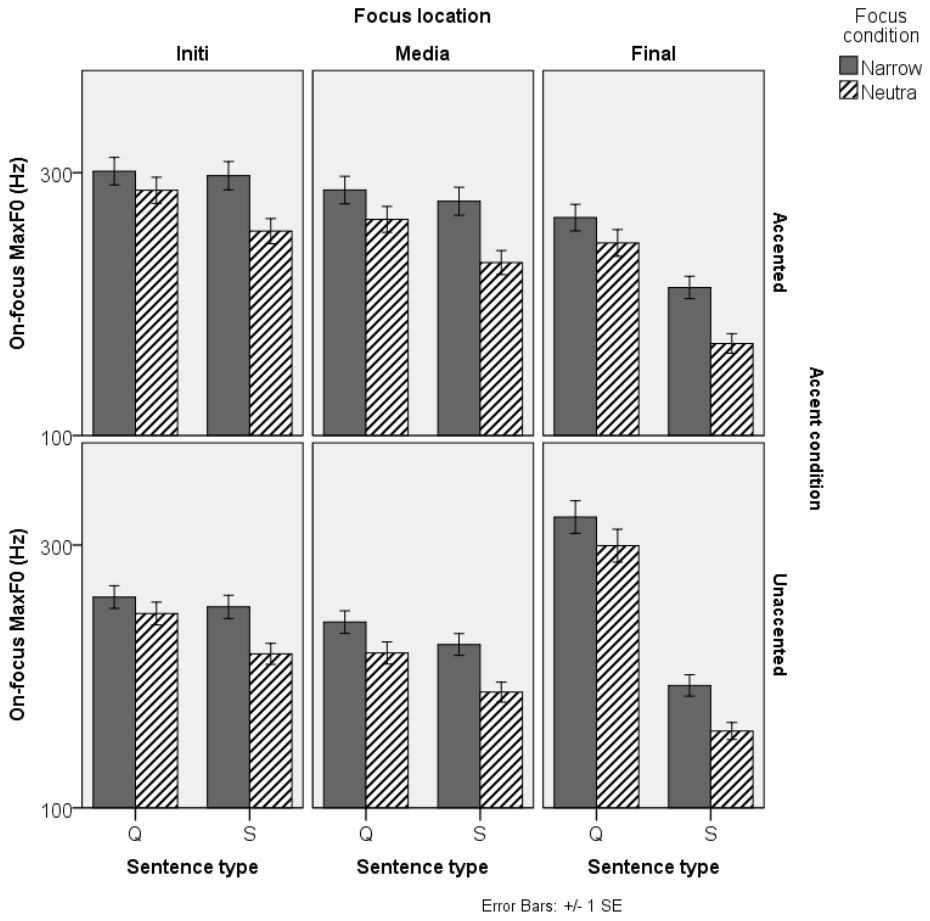


***Figure 39**. On-focus MaxF0 (in Hz) across sentence types, accent conditions and focus conditions.*

5.4 Discussion

This study investigated how questions differed from statements, as well as the roles of focus and pitch accent in interrogative utterances. For the effect of sentence type, it was found that questions have higher F0 peaks than statements, ceteris paribus. This echoes with comparable work on other languages (Eady & Cooper, 1986 for English; Ma, Ciocca, & Whitehill, 2006 for Cantonese; Yuan, 2006 for Mandarin). Also, PFC appears to be absent in questions, unlike in statements where PFC is present after a pitch accent. Moreover, intensity does not seem to play a role in marking focus, again unlike in statements where raised intensity is often associated with focus (also contra Kochanski, Grabe, Coleman, & Rosner, 2005).

If PFC is absent in questions, it would appear that only on-focus raising of F0-related focus markers are left at speakers' disposal. Although **Figure 35**, **Figure 36**, and **Figure 37** all show evidence of PFC after initial accented focus, as previously mentioned only some of the speakers actually employed this strategy, thus the non-significant results in ANOVA. For speakers who do use PFC, the F0 trajectory is similar to what we saw in statements — lowered

post-focus F0 peak, but with a sentential-final rise that marks questions and a generally raised F0 level. Where PFC is absent, the movement of F0 trajectory resembles declarative unaccented focus, guided by on-focus raising and carryover effect; the overlapping post-focus contours of narrow and neutral focus seem to suggest a common underlying target, just like in unaccented declarative focus in Chapter 4.

The effect of focus on duration in questions is largely confined to the focused item, standing in contrast to previous work on Japanese focus (Maekawa, 1997) where both pre-focus and post-focus shortening were observed. There is evidence of pre-focus lengthening, which is potentially attributable to Intonational Phrase boundary insertion before focus, causing final lengthening in the pre-focus item; however the small increase in duration (<4 ms for both medial and final foci in the data) is probably negligible. The discrepancy between the present findings and Maekawa's is reminiscent of the research literature on English focus prosody, where there are studies that found strictly local focus effect on duration (Eady & Cooper, 1986) as well as those that found extra-focus effects (Folkins, Miller, & Minifie, 1975; Weismer & Ingrisano, 1979). Eady and Cooper (1986) suggest that the discrepancy could be due to the length of the target sentence, but the present statistical analysis found no significant interaction between utterance length (8 morae vs. 11 morae) and focus on most of the pre-focus or post-focus variables. For the one interaction (Length*Focus) that turned out to be significant (duration before medial focus, $F(1,9) = 9.43$, $p = 0.013$), the effect size is small. A corpus with much longer sentences may produce different results, but given the present data I conclude that in interrogative focus prosody durational focus markers are strictly on-focus.

It thus appears that alongside on-focus lengthening the most robust cue to focus in interrogative prosody is on-focus raising of F0. The fact that PFC is absent is interesting in that in other studies (Taheri Ardali, Xu, & Rahmani, 2014) it has been shown that PFC is a salient cue, and that where PFC is naturally absent (i.e. final focus), focus is easily confused with neutral. This thus poses the issue whether focus can be accurately identified in questions as well as in statements. A perception study is under way to this effect.

From a cross-linguistic perspective, that there is no PFC in questions informs us that even in languages where PFC is typologically present (Xu et al., 2012; Xu, 2011a), its realisation is subject to certain functional conditions. In the case of Japanese, the realisation of PFC appears to be subject to at least accent condition and sentence type.

5.5 Chapter conclusion

In this Chapter the acoustic correlates of focus in yes-no questions were investigated. Although in some speakers PFC was used after an accented focus, the effect of focus on post-focus F0-related measurements did not reach statistical significance when pooling all 10 speakers' data together. On-focus raising of F0 was robustly observed across all conditions. Compared to statements, questions have the effect of raising general F0 level of the entire sentence,

potentially taking F0 peaks near one's pitch ceiling, but the statistical analysis shows that there are subtle differences among the F0 peaks in question narrow focus, question neutral focus, and statement narrow focus.

CHAPTER 6: SYNTHESIZING WORD AND SENTENTIAL PROSODY

6.1 Introduction

Analysis-by-synthesis is a robust way of showing how capable a model is in capturing variability in intonation (see Xu, 2011b for a systematic review on works in the analysis-by-synthesis approach). By assessing the accuracy of the synthesis, one can directly compare various models using the same dataset. In the midst of the 'lack of reference problem' (Xu, 2011b) in prosodic research, that is, the study of prosody does not have a reference like word identity in the case of studying segments (except in the case of lexical tone), this approach offers an additional guard against running into circularity.

The accuracy of a model can be assessed by measuring how much natural utterances differ from the synthetic ones. A good model that manages to capture the variability present in a given corpus should be able to resynthesise the entire corpus with good accuracy, in terms of such measures as Pearson's *r* and root-mean-square error (RMSE). That said, good synthesis accuracy alone does not necessarily guarantee a good model; an informative and useful model should also be based on assumptions that bear relevance to human speech mechanisms.

The assumptions that underlie a model thus tell good models from bad ones. One rule of thumb is Occam's razor, which states that in the face of rival models the one that has fewer assumptions should be selected. In other words, ceteris paribus, the more parsimonious a model (i.e. smaller degrees of freedom) is the better. However, when the additional assumptions are justified (e.g. physiologically motivated), a more elaborate model is not necessarily worse. Therefore, the ideal model for speech prosody should have assumptions that reflect human speech production / perception mechanisms (whether physical / physiological / neural), and only incorporate abstract notions when there are functional justifications to them (such as distinguishing communicative meanings). Only such a model would maximally resemble a human talker, and inform us about how speech prosody works.

To avoid resorting to unnecessary abstract notions in a model, one first needs to distinguish contextual[25] variability from non-contextual ones (Xu & Prom-on, 2014). In speech prosody, contextual variability includes tonal coarticulation (Xu & Wang, 2001), post-low bouncing (Prom-on et al., 2012), and pre-low raising (Gandour et al., 1994; Laniran & Clements, 2003); whereas non-contextual variability comes from the need to convey complex messages in speech. A few types of contextual variability are the result of physiological constraints and can thus be predicted, such as the carryover effect in tonal coarticulation, post-low bouncing, or the timing of surface F0 landmarks. These predictable features thus need not be specified as input

[25] Here 'context' means tonal context, i.e. surrounding tones.

parameters — algorithms can be incorporated into the model as underlying assumptions. Meanwhile, non-contextual variability should be functionally motivated, and established in controlled experiments. Ideally, variant realisations of a given communicative function should be represented as one invariant category. The conditions that give rise to the 'allophones' should be predicted by the model if they are contextual, or represented in parallel with other co-occurring functions which lead to such variants.

Such a stringent approach to the modeling of prosody is pursued by PENTA. Recent development in PENTA has focused on its application in F0 synthesis. Using PENTAtrainer2 (Xu & Prom-on, 2014), aspects of the prosody of several languages have been analyzed by synthesis, with good synthesis accuracy achieved (e.g. Liu et al., 2013 for focus and sentence type in English and Mandarin; Prom-on et al., 2012 for neutral tone in Mandarin). This Chapter contributes to this endeavor and reports two modelling studies using the corpora described in previous Chapters, to show that PENTA is suitable for accentual languages (Japanese) as well as tonal and stress languages (e.g. Mandarin and English). In §6.2, the corpus reported in Chapter 2 is resynthesised using PENTAtrainer1, PENTAtrainer2 and AMtrainer, to compare the performance of local parameters of AM and PENTA, and to compare the respective performance of PENTA's own local and global parameters. In §6.3 we look at how PENTAtrainer2 synthesises longer utterances (i.e. the corpus reported in Chapters 5 and 6) that involve communicative functions like focus and sentence type. The focus of the latter study is whether good synthesis accuracy in numerical terms reflects good synthesis to the native listeners' ears. The goal of this Chapter is to pave way for future modelling studies that compare multiple theories using the same dataset and evaluation protocols, and to illustrate that syntheised stimuli from PENTAtrainer2 are sufficiently accurate to be used in a variety of applications.

6.2 Synthesizing word prosody[26]

6.2.1 Introduction

The modelling of Japanese lexical prosody dates back to at least the 1960s when the Fujisaki Model was introduced (see introduction in Chapter 1). In his model (Fujisaki & Nagashima, 1969 and subsequent work) Fujisaki proposed that the intonation of a Japanese word could be modelled using an Accent Command and a Phrase Command. On the other hand, studies in the AM framework (notably Beckman & Pierrehumbert, 1986a; Pierrehumbert & Beckman, 1988) have proposed an elaborate model for Japanese intonation which, combined with relevant synthesis rules (Beckman & Pierrehumbert, 1986b; Pierrehumbert, 1981), could be used to synthesise F0 contours. This Chapter presents synthesis results from the PENTAtrainers, to

[26] A version of this Section was reported in Lee et al (2014).

show the workings of PENTA and that PENTA is compatible with a pitch accent language like Japanese. Further, I synthesise word intonation using AMtrainer, an synthesiser based on Pierrehumbert (1981), to attempt a direct comparison between AM and PENTA, which has never been done in previous research[27].

6.2.2 Methods

6.2.2.1 The corpus

The corpus used in the present study was described in §2.2.1. A total of 33 Japanese words were chosen as stimuli (see **Table 2**, Page 29). The target words varied in length (1~4 morae), accent condition (unaccented and initial/medial/ penultimate/final accent), and syllable structure (CVCV, CVn, CVV). From eight speakers, a total of 2,640 utterances (33 target words × 8 speakers × 5 repetitions × 2 speech rates) were collected. The target words were framed in the carrier sentence *jiten-ni X-mo nottemasu* 'The word X too is found in the dictionary'.

6.2.2.2 PENTAtrainer

As described in §3.2, the PENTAtrainers are two semi-automatic software packages for analysis and synthesis of speech melody based on communicative functions and Target Approximation model (Xu & Wang, 2001; Xu, 2005). They are both in the form of Praat (Boersma & van Heuven, 2001) scripts. The basic idea of the PENTAtrainers is to extract the underlying pitch targets defined in height (*b*), slope (*m*), and strength (*λ*) by means of automatic analysis-by-synthesis based on the quantitative Target Approximation (qTA) (Prom-on et al., 2009).

PENTAtrainer1 extracts target parameters locally unit by unit through an exhaustive search based on analysis-by-synthesis. For each target interval (typically the syllable), PENTAtrainer1 compares all possible combinations of *b, m*, and *λ* within the search ranges and finds the parameter combination that generates F0 contours with the least difference from the original. It also records learning accuracy in terms of RMSE and Pearson's *r* for each labeled interval, as well as the mean RMSE and global *r* for all the labeled intervals in the utterance.

[27] To the best of my knowledge, little work has been done to directly compare the effectiveness of different models, using comparable data and evaluation protocols, with a few exceptions (Raidt, Bailly, Holm, & Mixdorff, 2004; Sun, 2002).

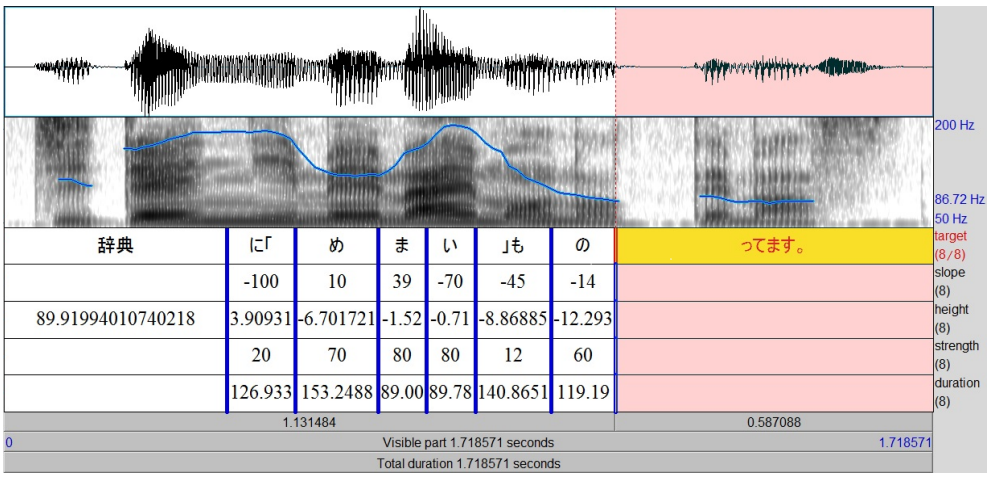


***Figure 40**. Extraction of model parameters for each labeled interval by PENTAtrainer1 (Jiten-ni me**ma**i-mo nottemasu 'The word "memai" too is found in the dictionary'). In order, the second to the fifth tiers show slope, height, strength, and duration of the labeled intervals. The parameter numbers in Tiers 2-5 are extracted rather than manually entered.*

PENTAtrainer2 extracts qTA targets globally from an entire corpus by means of analysis-by-synthesis based on simulated annealing[28] (Kirkpatrick, Gelatt, & Vecchi, 1983). This strategy is reminiscent of child language acquisition; through hearing the talker adjusts and refines his/her articulation and attains native adult-like pronunciation over time. In this experiment, PENTAtrainer2 is trained on the data through 1000 iterations to yield optimal global parameters (see Appendix 2). More description of PENTAtrainer2, including its comparison with PENTAtrainer1, can be found in Xu and Prom-on (2014).

To apply it, users need to annotate each interval with labels for the functions being modeled, as illustrated in **Figure 41**. The labels used and their respective meanings are listed in **Table 13** and **Table 14**. Note that underlying targets occupy the entire duration of a given tone-bearing unit (e.g. syllable), and that labels on different tiers have aligned boundaries (recall 'parallel encoding'). The result of the subsequent target extraction process will be globally optimal values of *b, m,* and *λ* for each of the functional combinations. Like PENTAtrainer1, PENTAtrainer2 records RMSE and *r* values as indicators of modeling performance.

[28] This is a strategy for efficiently obtaining articulatory parameters within a large search space. This is opposed to the exhaustive search strategy in PENTAtrainer1, which ensures the best possible outcome but takes a long time to compute.

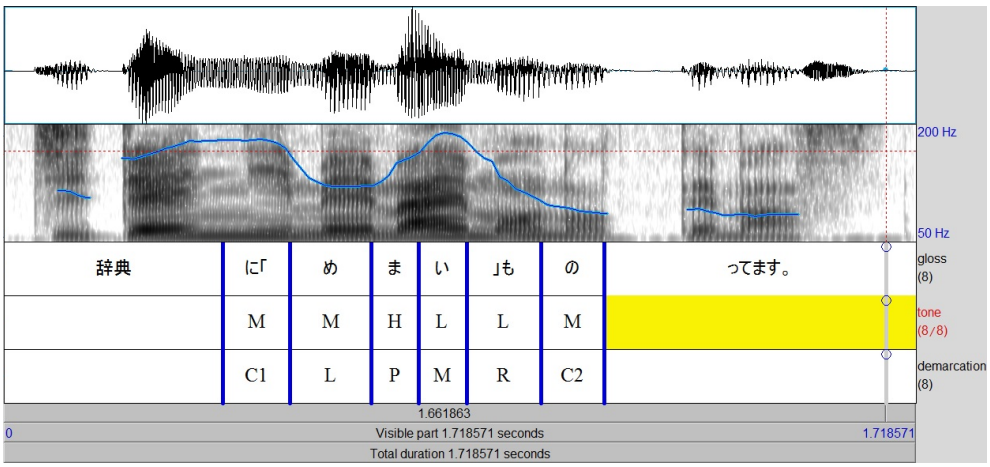


***Figure 41**. Functional annotation in PENTAtrainer2 (same sentence as in **Figure 40**). The labeled functions are Tone and Demarcation.*

With both PENTAtrainers, predictive F0 contour generation can be performed with averaged categorical target parameters. With PENTAtrainer2, the categorical parameters are extracted directly. With PENTAtrainer1, the categorical parameters are the mean parameters of all the individual tokens of the same category. F0 contours generated with the categorical parameters can then be compared to those of the natural utterances. See 6.2.3.3 below for the results of predictive synthesis.

	Mora	Syllable
Tone	H,M,L	H,M,L,F
Demarcation	C1,L,M,R,P,LP,C2	C1,L,M,R,P,LP,C2
TBU	Mora	Syllable

***Table 13**. Functional labels used in the three annotation schemes for PENTAtrainer1 and PENTAtrainer2.*

Tone		Demarcation	
H	The host mora of pitch accent (≈H* in J-ToBI)	C1, C2	Parts of the carrier sentence that respectively precede and follow the target word.
M	All the morae/syllables of an unaccented word (≈H- in J-ToBI)	L, R	Respectively the left edge and the right edge of target word that do not bear pitch accent (cf. P/LP below).
L	All the morae/syllables after the accent host (≈L in J-ToBI)	M	Word-medial positions that do not bear pitch accent (cf. P/LP below).
F	A heavy syllable (CVV/CVn, in the Syllabic segmentation scheme described below in 6.2.2.4) that hosts pitch accent (≈H*+L in J-ToBI)	P, LP	Accent-bearing word-medial positions and left edge of word, respectively.

***Table 14**. Definition of the functional labels in **Table 13**.*

6.2.2.3 AMtrainer

AMtrainer is a Praat-based training model developed by Yi Xu. It provides a similar user interface as PENTAtrainer1, but the parameters extracted are location and height. Built upon algorithms proposed in Pierrehumbert (1981), AMtrainer take point tier labels as input (see

Figure 42), which correspond to specific F0 turning points on the surface; the rest of the F0 contours are assumed to result from linear or sagging interpolation between the turning points.

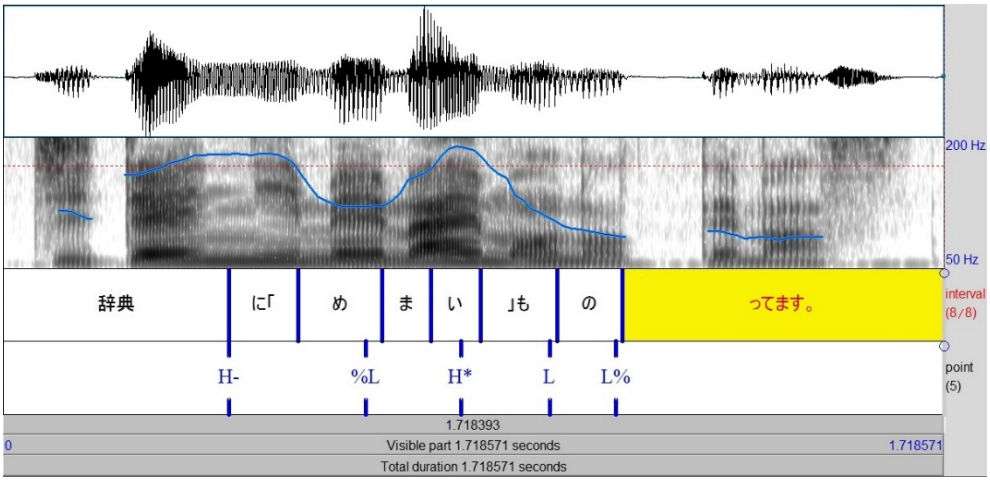


***Figure 42**. ToBI annotation for AMtrainer (Sentence as **Figure 40**)*

The present analysis follows the standard J-ToBI annotation convention (Venditti, 2005) for Japanese lexical prosody. The annotation of an unaccented word consists of the boundary tone (L%), and a phrasal tone (H-), whereas that of an accented word contains also a pitch accent (H*+L). Note that for simplicity's sake here the phrasal tone (H-) is omitted in cases where pitch accent occurs in the first or second mora of the word (in which case H* and H- would be too close to each other to add to discernable improvement in synthesis accuracy). See Venditti (2005) for more information regarding J-ToBI.

Annotation for AMtrainer was performed in three steps. First, continuous F0 contours were obtained from ProsodyPro (Xu, 2013), with vocal pulses manually checked and rectified. Then, from these data the F0 turning points corresponding to %L, H-, H*+L, and L% were identified for each utterance; and subsequently, converted to Praat TextGrid files to be used as the input for AMtrainer. The criteria for identifying F0 turning points were as follows:

Tone	Definition
H- (#1)	This tone corresponds to the beginning of the case marker *–ni*, which is part of the carrier sentence that precedes the target word. The inclusion of this tone is to allow for interpolation with the following %L.
%L	This tone corresponds to minimum F_0 in the first mora of the target word.
H- (#2)	This tone corresponds to the maximum F_0 velocity value in the second and third morae of the target word.
H*	This tone corresponds to maximum F_0 in the accent host mora and the ensuing one.
L	This tone corresponds to the minimum F_0 velocity value in the first two post-accent morae.
L%	This tone corresponds to minimum F_0 in the mora after the target word (i.e. *no-*).

***Table 15**. Label extraction criteria for AMtrainer*

6.2.2.4 Analyses

The accuracy of PENTAtrainers depends on both the target approximation algorithm and how well the annotation/categorisation scheme captures the variation in the data. Here I consider two schemes: Mora and Syllable.

The analysis will be presented over three subsections below. The PENTAtrainers assess the goodness of fit between the synthesised and original F0 contours using two measurements, namely, learning accuracy and synthesis accuracy. Although the two measurements are highly similar in nature, the design of AMtrainer renders it only possible to yield the former, in terms of which, in §6.2.3.1, I will first compare AMtrainer and the PENTAtrainers, before proceeding to the discussion of the synthesis accuracy results of PENTAtrainer1 and PENTAtrainer2. In §6.2.3.2, I consider the accuracy of speaker-dependent synthesis — synthesis of the F0 contours of a given speaker using the global parameters learned from his/her own utterances. In §6.2.3.3, the results of predictive synthesis accuracy is presented. Here I adopt the Jackknife procedure (Quenouille, 1956), where the global parametric values of all speakers save one are averaged and used to predict F0 contours of the speaker being left out. The procedure is repeated eight times such that all eight speakers' data are assessed.

6.2.3 Results

6.2.3.1 Learning accuracy

Table 16 shows the learning accuracy of AMtrainer, PENTAtrainer1, and PENTAtrainer2. Both annotation schemes under PENTAtrainer1 (second and third groups from top) yielded higher Pearson's *r* (0.998 and 0.994) and lower RMSE (0.101 and 0.122) than the other groups, suggesting that synthesised F0 contours from PENTAtrainer1 differed less from the original. AMtrainer reached similar learning accuracy to PENTAtrainer2.

	Segmentation	Accented		Unaccented		Overall	
		RMSE	*r*	RMSE	*r*	RMSE	*r*
AMTrainer		0.623	0.972	0.727	0.765	0.654	0.909
PENTAtrainer1	Mora	**0.112**	**0.998**	**0.075**	**0.992**	**0.101**	**0.996**
	Syllable	0.136	0.997	0.09	0.985	0.122	0.994
PENTAtrainer2	Mora	1.117	0.960	1.021	0.804	1.088	0.913
	Syllable	1.129	0.958	1.006	0.743	1.092	0.893

***Table 16**. Learning accuracy of AMtrainer, PENTAtrainer1 and PENTAtrainer2.*

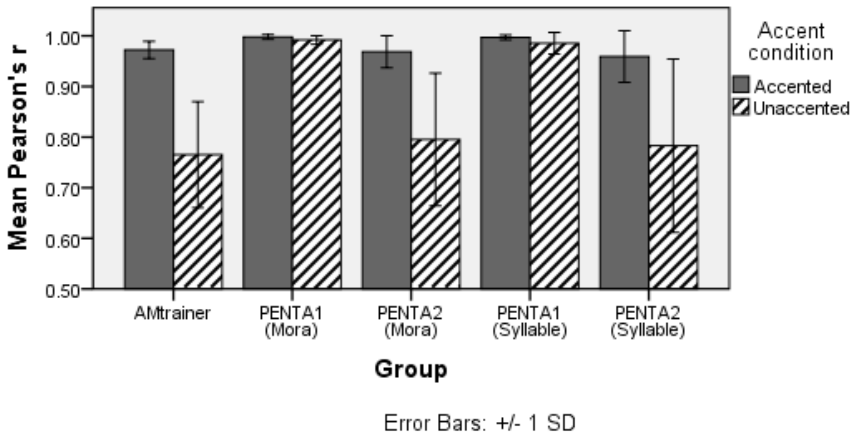


***Figure 43**. Mean Pearson's r of the three training tools.*

An interesting pattern emerged after subsetting the data according to accent conditions (accented vs. unaccented). As is obvious in **Table 16**, learning accuracy was considerably lower in unaccented words than in accented words for AMtrainer and PENTAtrainer2; for PENTAtrainer1 learning accuracy was similar in the two accent conditions. Moreover, for PENTAtrainer2, learning accuracy of unaccented words was higher with mora being the tone-bearing unit than otherwise.

6.2.3.2 Speaker-dependent predictive synthesis

In this subsection synthesis accuracy results are reported. The difference between this subsection and §6.2.3.1 is that here all resynthesised contours are generated from global parametric values, whereas in the previous subsection local parametric values were used for PENTAtrainer1. Assessment of accuracy is based on all the original F0 contours of a given speaker compared with those generated from the global parametric values learned from all the utterances of the same speaker. Note that since PENTAtrainer1 only extracts local parametric values of individual utterances, here the global values used for PENTAtrainer1 are the result of averaging over individual local values.

Table 17 shows that PENTAtrainer2 has an advantage over PENTAtrainer1 in synthesis accuracy. This difference is to be solely attributed to the sources of global parametric values used for generating F0 contours — for PENTAtrainer1, the global parametric values are the mean averages of local *b, m*, and *λ*, whereas for PENTAtrainer2, the global values are directly obtained through optimisations over an entire corpus. The present results thus show the effectiveness of global optimisation for predictive synthesis.

Figure 44 shows the synthesis accuracy of the PENTAtrainers under different accent conditions. Similar to what was observed in **Table 16**, unaccented words consistently achieved weaker Pearson's *r* than their accented counterparts. Note that PENTAtrainer1 achieved much

weaker *r* than it did in **Table 16**, because here F0 contours were synthesised using averaged global values, whereas in **Table 16** synthesis was based on local parametric values and did not have to capture cross-repetition variations.

	Segmentation	Accented		Unaccented		Overall	
		RMSE	*r*	RMSE	*r*	RMSE	*r*
PENTAtrainer1	Mora	1.786	0.924	1.796	0.674	1.789	0.849
	Syllable	1.746	0.941	2.054	0.628	1.839	0.846
PENTAtrainer2	Mora	**1.117**	**0.962**	1.021	**0.804**	**1.088**	**0.914**
	Syllable	1.129	0.960	**1.006**	0.748	1.092	0.896

***Table 17**. Accuracies of speaker-dependent synthesis by both PENTAtrainers.*

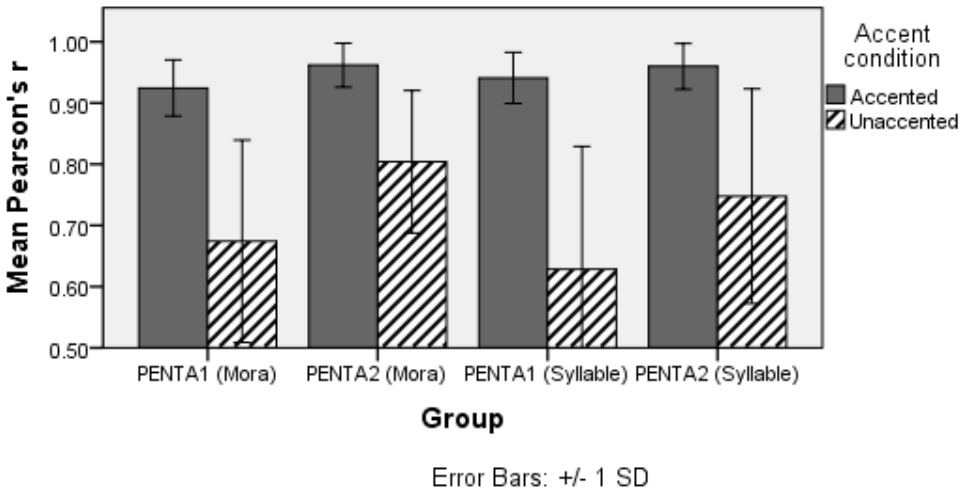


***Figure 44**. Mean Pearson's r of speaker-dependent synthesis by both PENTAtrainers.*

Finally, in an attempt to improve the synthesis accuracy of unaccented words, I tested an additional function 'Word Length' (i.e. 1~4 morae), alongside 'Tone' and 'Demarcation'. Despite using more predictors (i.e. 11→32 sets of global articulatory parameters [*m, b, λ*] for moraic segmentation, 15→41 for syllabic segmentation), and the known effect of word length on F0 in Japanese (Selkirk et al., 2004), synthesis accuracy deteriorated for unaccented words, with RMSE = 0.952, *r* = 0.797 (down from *r* = 0.804) for moraic segmentation, and RMSE = 0.958, *r* = 0.793 (from 0.748) for syllabic segmentation. This suggests that 'Word Length' was not effective in capturing the remaining variation in the data, echoing with the conclusion in §2.3.1 that word length is not a good predictor compared with accent condition and peak-to-end distance. In turn, it also lends further support to the observation in Chapter 2 that peak-to-end distance influences accent peak F0; the effect of peak-to-end distance would be meaningless if word length has a systematic effect on F0 in the data at the same time.

6.2.3.3 Speaker-independent Predictive synthesis

Using the Jackknife procedure, the predictive power of the global articulatory parameters of PENTAtrainers was assessed for each speaker in the corpus. As can be seen in **Table 18**, even though the recordings of the speaker being assessed were excluded from the training corpus, predictive synthesis still yielded satisfactory accuracy (overall r>0.8 in most cases). However, here PENTAtrainer2 no longer showed absolute advantage over PENTAtrainer1. This is especially the case with unaccented words, where both segmentation schemes under PENTAtrainer2 yielded r<0.7. The difference in accuracy between **Table 17** and **Table 18** shows that there is considerable cross-speaker variation in the present corpus.

	Segmentation	Accented		Unaccented		Overall	
		RMSE	*r*	RMSE	*r*	RMSE	*r*
PENTAtrainer1	Mora	1.761	0.925	1.807	0.668	1.775	0.847
	Syllable	1.938	**0.932**	2.307	0.585	2.050	0.826
PENTAtrainer2	Mora	**1.726**	0.921	**1.767**	**0.696**	**1.739**	**0.853**
	Syllable	2.088	0.877	2.547	0.608	2.227	0.796

***Table 18**. Accuracies of speaker-independent predictive synthesis*

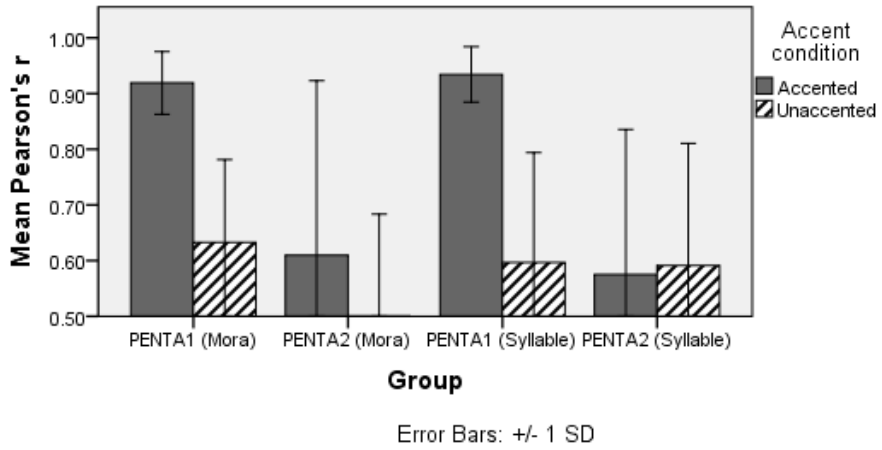


***Figure 45**. Mean Pearson's r of speaker-independent predictive synthesis (Jackknife procedure).*

6.2.4 Discussion

AM and PENTA differ in terms of several theoretical assumptions, which have implications for their predictive power. Notably, in the former a tone target is a point in the surface contour but an underlying linear trajectory in the latter. In addition, for AM the temporal alignment of a tone in relation to segments is flexible and has to be specified. For PENTA pitch targets and the tone-bearing units are synchronised to segmental units and so no further alignment specification is needed.

It is not possible to assess whether AM or PENTA is superior based on the present study because tone labelling for AMtrainer was performed post-hoc based on the actual location of acoustical landmarks whereas the categories used in PENTAtrainers were pre-defined. The goodness of fit for AMtrainer only reflects the effectiveness of linear and sagging interpolation

as an F0 contour generation mechanism (Pierrehumbert, 1981). In contrast, both PENTAtrainers perform predictive synthesis based on functional categories. That AM labels were extracted from the actual location of acoustical landmarks limits the comparability between the tools here. To render AM more comparable to models that take categorical input like PENTA, the height value and temporal alignment of its labels need to be predicted by an algorithm rather than added post-hoc. To do so, one must first find out the segmental anchoring behavior of each tone (T. Ishihara, 2006) and then calculate the temporal alignment of the tones in each utterance; whereas for scaling there is no simple way of prediction yet (but see Beckman & Pierrehumbert, 1986b). Once the issue of post-hoc annotation for AMtrainer is overcome, it would be desirable to test the assumptions of temporal alignment vs. target approximation using a dataset that controls for speech rate, like the one used in the present study.

Nonetheless, this Section has shown that both AM and PENTA can fit Japanese word prosody non-predictively with satisfactory accuracy. The fact that both models do almost equally well means that AMtrainer and PENTAtrainer can serve as a platform for fair and direct comparisons between the two theories if used properly. Collaborative efforts are needed in the future to reach this goal by investigating more types of speech data and devising a protocol of annotation for an unbiased comparison of the models.

A side issue that was unresolved in Chapter 2 is the transition between acoustical landmarks (peaks, valleys, and turning points). In a four-mora final accented word like in **Figure 6** (page 25), AM captures the upward movement between H- and H* elegantly by positing a linear interpolation between a lower H- and a higher H*. In terms of Target Approximation, one possible account would be that the H- sequence is a H tone spoken with weak articulatory effort, whereas the H* is the same H tone but subject to pre-low raising, and spoken at greater articulatory strength. This proposal is inspired by the findings in Prom-on et al (2012) where neutral tone is argued to be realised at weak articulatory strength, thus taking several syllables until the underlying target is reached. In this case, the gradual rise from H- (second mora) is due to target undershoot, and is only reached at H* where a stronger effort speeds up the achieving of the target. While hypothetical, this account finds support in the strength values learned by PENTAtrainer2 above (see Appendix 2). For both moraic segmentation and syllabic segmentation, the strength value that corresponds to H* (highlighted turquoise) is greater than that for H- (highlighted yellow). It is thus likely that what is usually described as a linear transition between two sparse targets is a series of static targets spoken at weak articulatory strength. More work is needed to confirm this account.

6.2.5 Interim conclusion

This Section has presented a user report of AMtrainer and the PENTAtrainers. I aimed to set a fair platform for further comparison between PENTA Model and AM Theory in modeling and predicting the F0 contours. Both PENTAtrainer1 and PENTAtrainer2 reached an accuracy of

predictive synthesis as high as r>0.9, showing that PENTA is an effective tool in F0 modeling. The high accuracy achieved by AMtrainer reflects the effectiveness of sagging and linear interpolation as a means of contour generation. Meanwhile, the similarity between the results from AMtrainer and PENTAtrainers suggests that there is a potential for AMtrainer to predictively generate F0 contours with functional and categorical input. But a more objective way of generating the input for AMtrainer is needed before a full comparison between the two theories is possible.

6.3 Synthesizing sentential prosody

6.3.1 Introduction

This Section takes PENTAtrainer2 further and reports the results of F0 synthesis of the corpus described in Chapters 4 and 5 (**Table 12**, Page 70). The goals of this study are to test whether PENTAtrainer2 is capable of synthesizing longer utterances (up to 11 morae in this case) that involve various communicative functions (e.g. focus and sentence type), as well as to assess whether the numerical synthesis accuracy reported by PENTAtrainer2 truly reflects its genuineness to the native listeners' ears. First the accuracy data of predictive and non-predictive synthesis will be presented to compare the goodness of fit between natural utterances and synthesised ones, followed by naturalness judgment ratings of the same synthesised stimuli by native listeners. The implications of these results will be discussed in §6.3.4.

6.3.2 Methods

6.3.2.1 The corpus

The corpus described in **Table 12** will be analyzed. There are 6,400 utterances (2 sentence lengths × 8 accented conditions × 2 sentence types × 4 focus conditions × 5 repetitions × 10 speakers). For each target sentence there are four possible focus conditions, namely initial, medial, final, and neutral. The sentence types are yes/no questions vs. statements. Each sentence is either eight or 11 morae in length. Focus was elicited by having the speaker produce the question and the (corrective) statement in pair. Of the 6,400 utterances collected, 149 had to be discarded due to mis-production of accent condition. A total of 6251 were retained.

6.3.2.2 Annotation

In this study, raw sound data were first chunked into individual utterances, and subsequently segmented by mora and by syllable in Praat. Under moraic segmentation, a heavy syllable is segmented into two intervals equal in duration. Then, the segmented data were functionally labelled like in **Figure 46**.

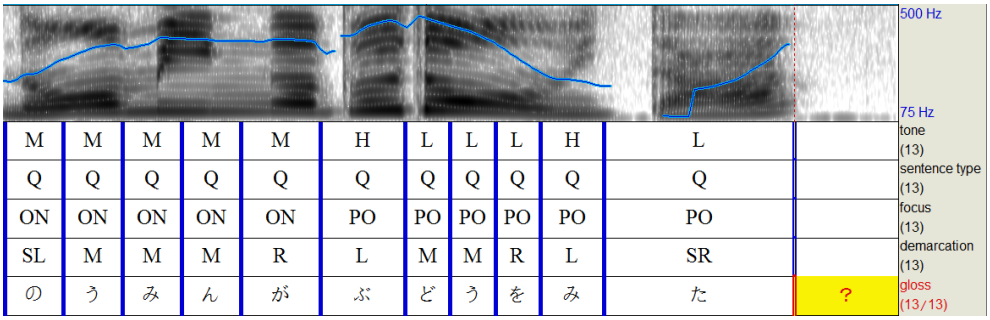


***Figure 46**. Functional annotation in PENTAtrainer2. The labeled functions are Tone, Sentence Type, Demarcation and Focus. The fifth tier (gloss) is not included in actual analysis.*

On the Tone tier each interval is marked H, M, or L, following §6.2. In the case of syllabic segmentation an accented heavy syllable is labelled F (see Chapter 3). H represents the high target in an accented word (cf. H* in Venditti, 2005), whereas M stands for the high target elsewhere (cf. H- in Venditti, 2005). The low target in an accented word is marked L. Under syllabic segmentation, pitch accent as hosted in a heavy syllable is hypothesised as bearing a falling target, thus the F label. Sentence Type is either Q(uestions) or S(tatements). Note that these labels provided no phonetic guidance to PENTAtrainer2, as they are treated simply as category identifiers. On the Focus tier, intervals in a focused sentence are labeled as on-focus, pre-focus, or post-focus (Xu et al., 2004), and those in a neutral sentence are all labelled N (neutral). The Demarcative tier contains information on the position of an interval in the word and the sentence, comprising five categories–left/right edge of word, middle of word, and left/right edge of sentence. Usually this tier contains other morpho-syntactic information as well, but since there is only one type of syntactic structure (simple SOV) in this corpus five categories sufficed for our purpose. These four tiers combined give rise to 72 unique communicative conditions for the corpus, which means that the entire corpus will be synthesised using 72 sets of qTA parameters (*b, m, λ*).

Globally optimal parametric values were obtained (see Appendix 3) after 1,000 reiterations of the learning process. §6.3.3.1 reports the accuracy of speaker-dependent synthesis, that is, synthesis of the F0 contours of a given speaker using the global parameters learned from his/her own utterances In §6.3.3.2, the results of predictive synthesis accuracy are presented. Like in §6.2, here I adopted the Jackknife procedure, where the global parametric values of all speakers save one are averaged and used to predict the F0 contours of the speaker being left out. The procedure is repeated ten times such that all ten speakers' data are assessed.

6.3.2.3 Naturalness judgment

The synthesis quality was also assessed perceptually in a naturalness judgment test. Sixteen native Japanese listeners (3 male) were recruited as subjects. They were all born and raised in the Greater Tokyo area (Tokyo, Saitama, Kanagawa, and Chiba), and aged between 23 and 37 years old (mean age = 27.9, S.D. = 3.9, see **Table 19**). Most subjects had arrived in the UK less than a year ago, except YK who had arrived for 12 months, and MK who had spent two years in the USA. None reported any (history of) speech or hearing impairment.

Initial	Age	Sex	Born	Grew up	Father from	Mother from
MK	26	F	Tokyo	Tokyo	Tochigi	Aichi
KS	23	F	Saitama	Chiba	Chiba	Niigata
ER	29	F	Tokyo	Tokyo	Tokyo	Tokyo
AK	23	F	Chiba	Chiba	Kanagawa	Fukushima
YK	31	F	Tokyo	Tochigi	Fukushima	Tokyo
MF	24	F	Chiba	Chiba	Chiba	Fukushima
MY	24	F	Chiba	Chiba	Chiba	Chiba
SN	27	F	Kanagawa	Kanagawa	Osaka	Osaka
WW	30	M	Kanagawa	Kanagawa	Osaka	Osaka
TM	34	M	Tokyo	Tokyo	Kanagawa	Tokyo
AT	27	F	Tokyo	Tokyo	Nagoya	Nagoya
RO	28	F	Tokyo	Tokyo	Kagoshima	Kagoshima
YS	26	F	Tokyo	Tokyo	Tokyo	Tokyo
HF	37	F	Kanagawa	Saitama	Gunma	Gunma
YF	32	F	Kanagawa	Kanagawa	Kanagawa	Kanagawa
KK	26	M	Tokyo	Tokyo	Tokyo	Yamagata

***Table 19**. Information of participants in natural judgment test*

The listening test took place in a quiet room in University College London. Subjects were seated in front of a laptop computer, which displayed the Praat MFCC interface, and stimuli were presented over circumaural headphones. They listened to each stimulus and rated the naturalness on a 1~5 scale, with 5 being the most natural. Each stimulus could be replayed up to three times. All participants were briefed about the experiment and granted their written consent to being tested. Speakers were then interviewed about their linguistic background and history of speech and hearing impairment. All speakers were remunerated for their time. In total each speaker listened to 128 utterances.

6.3.3 Results

6.3.3.1 Speaker-dependent synthesis accuracy

Table 20 shows the respective mean synthesis accuracy under moraic and syllabic segmentation in terms of RMSE and *r*. Here the articulatory parameters used to synthesise the F0 contours of a given speaker are obtained through training on the utterances of the same speaker. Across sentence types and focus conditions, synthesis accuracy is high with $r>0.9$ and RMSE<1.8 in most cases. Although *r* appears to be greater in certain contexts, RMSE values are not smaller in those cases, suggesting that no particular condition is more accurately

modelled than the others. I also carried out visual inspection, like in **Figure 47**, and found that the resynthesised contours were highly similar to their original counterpart.

Sentence type	Focus	Mora		Syllable	
		RMSE	*r*	RMSE	*r*
Question	Initial	1.662	0.917	1.776	0.9
	Medial	1.555	0.921	1.644	0.902
	Final	1.495	0.915	1.514	0.913
	Neutral	1.603	0.908	1.711	0.89
	Sub-avg	**1.579**	**0.915**	**1.662**	**0.901**
Statement	Initial	1.725	0.928	1.823	0.915
	Medial	1.513	0.923	1.625	0.906
	Final	1.416	0.911	1.391	0.91
	Neutral	1.657	0.891	1.678	0.876
	Sub-avg	**1.578**	**0.913**	**1.63**	**0.901**
Grand average		**1.578**	**0.914**	**1.646**	**0.901**

***Table 20**. Accuracy of speaker-dependent synthesis.*

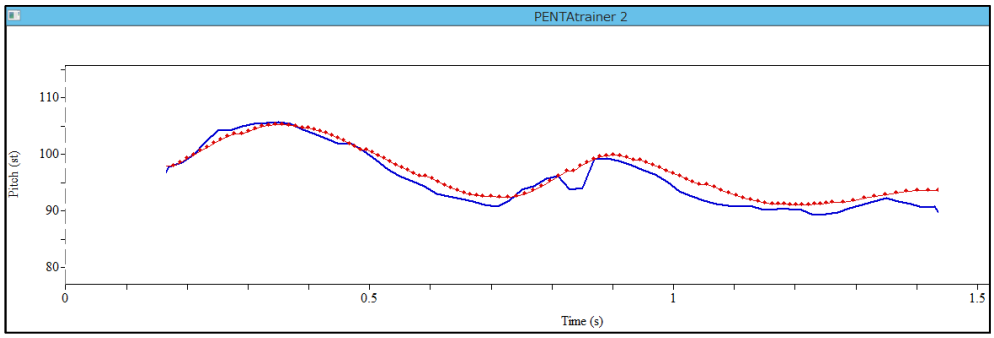


***Figure 47**. An interface of PENTAtrainer2 for visual inspection of synthesis accuracy. The blue curve is the F0 contour of a natural utterance whereas the red dotted curve represents the corresponding resynthesis. The target sentence in this example is **mu**umin-ga **bu**dou-o **mi**ta 'Moomin watched martial arts', with focus on the first word.*

6.3.3.2 Speaker-independent synthesis accuracy

Table 21 shows mean predictive synthesis accuracy under the Jackknife procedure, where the speaker being modelled was excluded from the training. The resynthesis deviated more from natural utterances (cf. **Table 20** above). The overall accuracy is RMSE = 2.733 *r* = 0.8 for moraic segmentation, and RMSE = 3.024 *r* = 0.702 for syllabic segmentation. The comparative advantage of the mora segmentation is greater here than in **Table 20**.

Sentence type	Focus	Mora		Syllable	
		RMSE	*r*	RMSE	*r*
Question	Initial	2.94	0.787	3.13	0.736
	Medial	2.639	0.804	2.947	0.721
	Final	2.468	0.849	2.658	0.769
	Neutral	2.554	0.842	2.979	0.732
	Sub avg	**2.65**	**0.821**	**2.929**	**0.739**
Statement	Initial	3.242	0.793	3.327	0.751
	Medial	2.687	0.8	2.975	0.662
	Final	2.766	0.739	3.018	0.663
	Neutral	2.578	0.785	3.154	0.581
	Sub avg	**2.817**	**0.779**	**3.118**	**0.664**
Grand average		**2.733**	**0.8**	**3.024**	**0.702**

***Table 21**. Synthesis accuracy of PENTAtrainer2 under Jackknife procedure by sentence type and focus condition.*

6.3.3.3 Naturalness judgment results

Results of the naturalness judgment test are presented in **Table 22**. I am interested in whether Type of stimuli (original vs. synthesised) affects how a listener rates the naturalness of stimuli, or whether its interaction with other effects reaches statistical significance. A repeated measures ANOVA was conducted to compare the effects of Focus condition, Sentence length, Accent conditions of Wd1, Wd2, Wd3, Sentence type, and Stimulus type on naturalness rating. Results show no significant main effect of Type of stimuli on naturalness scores. This suggests that the two types of stimuli sounded equally natural to the native listeners. The grand mean rating of natural stimuli is 3.688, which is close to that of synthesised stimuli (3.658, out of a 1~5 scale).

On the whole, statements (mean = 3.817) are judged to sound more natural than questions (mean = 3.528) in both natural and synthetic stimuli. The effect of Sentence type is, albeit small, statistically significant $F(1,7) = 5.64$, $p = 0.049$. This discrepancy is possibly due to the effect of polarity focus mentioned in Chapter 5 as an artefact of the design of the stimuli; indeed, there is a small but significant interaction between Focus condition and Sentence type $F(3,7) = 3.144$, $p = 0.047$, lending support to the view that polarity focus might be in play. This issue will be further discussed in §6.4.

Sentence type	Focus	Original	Synthesis	Average
Question	Initial	3.588	3.691	3.64
	Medial	3.566	3.456	3.511
	Final	3.397	3.441	3.419
	Neutral	3.507	3.581	3.544
	Sub-avg	**3.515**	**3.542**	**3.528**
Statement	Initial	3.772	3.669	3.721
	Medial	3.801	3.743	3.772
	Final	3.816	3.809	3.813
	Neutral	4.051	3.875	3.963
	Sub-avg	**3.86**	**3.774**	**3.817**
Grand average		**3.688**	**3.658**	**3.673**

***Table 22**. Mean naturalness ratings by focus condition and sentence type.*

6.3.3.4 Post-focus compression

There is one final issue left unresolved in §4.4.1. I argued that the F0 trajectory after an unaccented focus is guided by carryover effect and a weak articulatory strength. Recall that for statements with only unaccented words in **Figure 33** (Page 79), post-focus F0 contours converged with neutral focus after several syllables, reminiscent of neutral in Mandarin. To verify this hypothesis I revised the functional annotation (syllabic segmentation) such that the 'Post-focus' category is split into PU (post-focus not preceded by an accented word) and PO (post-focus elsewhere), and retrained the model using the same method as described above. This revised scheme captures the distinction between **Figure 33** and all others target sentences, and tells us if such distinction involves a difference in articulatory strength.

Results in **Figure 48** (also Appendix 3.4) appear to agree with this hypothesis. The upper panel represents questions, and the lower panel represents statements. In the statement panel, the PU condition (post-focus and not preceded by any accented word) has a much weaker strength than the PO condition (all other post-focus conditions), supporting the hypothesis that the post-focus F0 contours in **Figure 33** is a result of weak articulatory strength.

Note that the present result is intended to test the articulatory strength hypothesis for **Figure 33**. A more elaborate annotation scheme that differentiates all possible combinations, including a new condition of 'accent condition of preceding word' is necessary, but this example in **Figure 48** will suffice as a preliminary support for my proposal. On a side note, it may be counter-intuitive to some that the ON condition has a weaker articulatory strength than the PO condition. One possible explanation is that after a sharp accentual fall of the focused item, one needs to maintain sufficient muscle stiffness (i.e. strength) to prevent post-low bouncing, a phenomenon observed in Mandarin neutral tone (Prom-on et al., 2012) but not in Japanese PFC. Specifically, in Mandarin after a Low tone F0 bounces back to a high level, presumably due to a temporary loss of balance in laryngeal control, left unchecked under weak articulatory strength. In Japanese, F0 stays low (i.e. no bouncing) after a sharp accentual fall of the focused item. Thus it is possible that in cases where there is PFC, PO targets have great articulatory strength. However, again, this explanation needs to be verified by an appropriate annotation scheme; the present scheme was devised to compare **Figure 33** vs. other conditions.

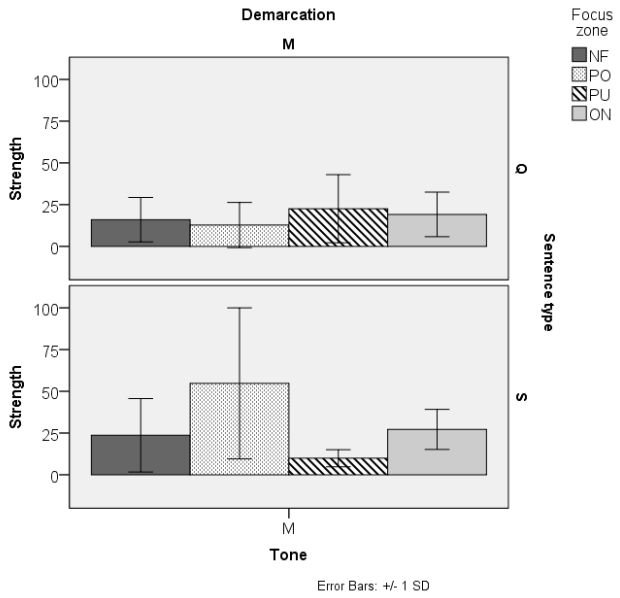


***Figure 48**. Mean strength values under different communicative function conditions (revised annotation).*

6.3.4 Discussion

The present study has shown that Japanese sentential prosody can be modelled with parametric representations based on an articulatory-functional model. Compared to §6.2.3 (Speaker dependent [mora] RMSE = 1.088, *r* = 0.914; [syllable] RMSE = 1.092 *r* = 0.896), results in **Table 20** are very similar. On the other hand, synthesis accuracy is much lower under Jackknife procedure (**Table 21**), compared to **Table 18** where [mora] RMSE = 1.739, *r* = 0.853; [syllable] RMSE = 2.227, *r* = 0.796, suggesting that there is more cross-speaker variability in sentential prosody than in lexical prosody. This observation echoes with (A. Lee & Xu, 2012) where some focus cues like pre-focus F0 lowering was found to be optionally used by some speakers, whereas other cues like post-focus compression were consistently used by all; for word prosody, such freedom is less common owing to the need to mark lexical contrasts.

The results agree with §6.2 where moraic segmentation yielded better synthesis accuracy. This may seem to suggest that the mora is the true tone-bearing unit in Japanese for tonal target approximation, contra the findings in Chapter 3 as well as other languages like Mandarin and English where tonal targets are hosted in the syllable. However, I am hesitant to come to such a conclusion because a heavy syllable under moraic segmentation comprises two intervals, but only one interval under syllabic segmentation. This means that by nature the former involves more degrees of freedom, leading to better ability to capture variability. Thus these results cannot be taken as an answer to what the domain of target approximation of Japanese is; the question needs to be tackled through better controlled experiments, which take into account the confounds from degrees of freedom, as was done in Chapter 3. Though

unexpected, it is not totally surprising that moraic segmentation naturally led to better accuracy in this Chapter.

The high synthesis accuracy is supported by naturalness judgment ratings by native listeners. This means that the synthetic stimuli do not sound different from the natural ones to the participants. By extension, the remaining errors not captured by PENTAtrainer2 do not make the resynthesis any less natural-sounding. This means that the key information has been successfully encoded in the learned parameters. Therefore, PENTAtrainer2 can offer, for purposes like perception tests, natural sounding stimuli which are free of cross-repetition inconsistency common in natural stimuli.

A further implication of the results is that the PENTA model as well as its prosodic representation are well suited for Japanese. Syllable-by-syllable target specification, as has been shown, is adequate for a corpus with numerous non-contextual (i.e. functional) variations. The encoding schemes of all functions jointly determine a unique articulatory target for each syllable. By incorporating articulatory factors (Xu & Sun, 2002), there is thus no need to specify temporal alignment of tone. Whether my approach is superior to other frameworks is still an open question, but I have shown that PENTA representation is at least as suitable for Japanese as for other languages like Mandarin and English (Liu et al., 2013; Xu & Prom-on, 2014).

The relatively low naturalness rating of questions is interesting, and is believed to be a result of polarity focus (i.e. verbs in yes/no questions tend to see raised F0, thus confusable with final focus) introduced by the design of the stimuli. Without any context given, an interrogative intonation pattern that sounded as if bearing multiple foci would understandably sound odd to a native listener. Even though previous research has shown that in a production experiment prosodic focus can be effectively elicited merely by boldfacing the word of interest (Fung & Mok, 2014), the naturalness of such utterances is not guaranteed. While having successfully controlled for utterance length and microprosodic effects on F0, admittedly our experimental design has taken toll on the naturalness of our interrogative stimuli. In this regard, this experiment has also served to clearly demonstrate the compromise between stringent experimental control in lab speech and its perceived naturalness.

On a side note, the training process of PENTAtrainer2 is also reminiscent of child language acquisition. The development of infant speech relies on audition — deaf children cannot learn to speak by themselves (Oller & Eilers, 1988; Raphael, Borden, & Harris, 2011). Over the course of repetitions PENTAtrainer2 refines its articulatory parameters in order to generate F0 contours that are more similar to the original, just as infants gradually refine their articulation over time by listening to themselves during practice. The role of feedback is now known to be of great importance and has been incorporated into recent models of speech production (e.g. the DIVA Model, Guenther & Perkell, 2004), in line with the workings of PENTAtrainer2.

6.4 Chapter conclusion

Two sets of F0 synthesis results were presented in this Chapter. In §6.2, I compared the performance of local vs. global articulatory parameters in synthesizing word prosody, and made a novel attempt at comparing AM and PENTA in terms of their local fitting accuracy. Naturally, local fitting by PENTAtrainer1 is superior to synthesis from global parameters as the former does not need to handle cross-repetition variation. In the meantime, that the AMtrainer yielded comparable synthesis accuracy is encouraging, as it shows that my approach offers a possibility of fair and direct comparison between the two models in future studies. F0 synthesis results were presented in §6.3. I showed that PENTAtrainer2 is capable of synthesizing longer Japanese utterances that involve multiple communicative functions like focus and sentence type. More importantly, it was shown that the RMSE and *r* values reported by PENTAtrainer2 truly reflected the genuineness of the synthesised stimuli, which were perceived by native listeners to be as natural-sounding as natural stimuli. To fully understand Japanese prosody, work needs to be done to directly compare PENTA with other theories, and ideally eliminate unfit models that do not reflect the real mechanisms of human speech.

CHAPTER 7: GENERAL CONCLUSIONS

This dissertation explored several aspects of Japanese prosody through a number of production experiments and modelling studies. I set out with two overarching goals: (1) to revisit numerous issues in the prosody of Japanese without resorting to abstract notions that can be otherwise accounted for in mechanistic terms; and (2) to illustrate that PENTA is suitable for Japanese.

In Chapter 2, I showed that in Japanese the F0 peak associated with a pitch accent varies with its following low tone. It was argued that the variable F0 peak height is the result of pre-low raising. Pearson's *r* revealed an inverse relation between accent peak and the following low tone, and that such a relationship became more pronounced when the peak was further away from word end. These results suggested that in Japanese a low tone raises the preceding high tone, which is consistent with our current understanding of the physiology of vocal fold tension control in F0 production.

The goal of Chapter 3 was to answer the mora/syllable dichotomy in Japanese from a Target Approximation perspective. Three pieces of evidence were found in support of the syllable as the domain of tonal target approximation — namely F0 peak timing, bimodal distribution of target slope and height, and PENTAtrainer1 learning accuracy. The results suggested that Japanese is like such languages as Mandarin and English, where the syllable bears tones.

Chapters 4 and 5 reported a production study of Japanese focus prosody that controlled for accent condition, focus condition, sentence type and word length. My findings generally agreed with previous research, i.e. that there is on-focus enhancement and post-focus reduction in various forms. It was proposed that PFC is only realised after a pitch accent, and does not include cases of compressed initial rise in an exclusively unaccented utterances. Those cases are, I argued, a result of a carryover effect from the preceding on-focus item, thus the apparent 'post-focus raising'. It was also observed that the raised contour after an unaccented focus gradually converged with the neutral focus contour, and posited that these two focus conditions have the same underlying pitch target, although the post-focus contour approximates this target at a low articulatory strength. Compared to statements, questions have the effect of raising the general F0 level of the entire sentence.

Two sets of F0 synthesis results were presented in Chapter 6. In §6.2, I compared the performance of local vs. global articulatory parameters in synthesizing word prosody, and made a novel attempt at comparing AM and PENTA in terms of their local fitting accuracy. The AMtrainer yielded comparable synthesis accuracy to the PENTAtrainers, suggesting that my approach can offer a possibility of fair and direct comparison between the two models in future studies. In §6.3, I showed that PENTAtrainer2 is capable of synthesizing longer Japanese utterances that involve multiple communicative functions like focus and sentence type. More importantly, it was shown that the RMSE and *r* values reported by PENTAtrainer2 truly reflected the genuineness of the synthesised stimuli, which were perceived by native listeners to be as

natural-sounding as natural stimuli. To fully understand Japanese prosody, work needs to be done to directly compare PENTA with other theories, and ideally eliminate unfit models that do not reflect the real mechanisms of human speech.

The findings in this dissertation have several theoretical and general implications to linguists and non-linguists alike. First of all, I have demonstrated how to survey a language from the PENTA perspective. Not only is this about assuming PENTA's theoretical framework (i.e. the mechanism of Target Approximation, parallel encoding, and full specification of tone targets) per se, but also the research principles that guide a PENTA-style study. Specifically, I have chosen to use strict minimal pairs as far as possible, though apparently at the expense of speech naturalness as discussed in §6.3.4. I have also shown how one can study speech prosody through multiple angles, namely acoustic analysis and analysis-by-synthesis. Insisting on such stringency means labour-intensive experiments that cannot answer many questions at once, but it also ensures robustness of the results and hopefully saves future researchers from having to repeat the same steps.

With regards to multi-angle analysis, the PLR proposal in Chapter 2 will ultimately need to be verified by physiological data. Whether using invasive electromyography or surface electromyography, it is necessary to investigate the coordination of laryngeal muscles at different speech rates to unveil the underlying mechanism of pre-low raising. If the physiological data turn out to disagree with my proposal, then all the data in Chapter 2 will require a drastically different explanation.

The account in Chapter 4 is of typological significance. Whereas Japanese is found to be a language that marks focus with PFC, I have argued that its realisation is conditional upon factors like pitch accent. This is the first study to show that even if a language is 'PFC-positive', PFC does not have to be used across the board. The findings in Chapter 4 thus enrich our understanding of PFC as a typological feature, and opens up a whole new research topic of language-specific conditions on PFC realisation. On a more indirect level, Chapter 4 offers explicit details of how focus is realised in Japanese, which may be useful for L2 learners who want to acquire a more native-like intonation.

In Chapter 1 I pointed out that PENTA is a theory of articulation, then in subsequent Chapters I attempted to show the relationship between surface acoustics and underlying articulatory targets (especially in Chapter 3). However, the link between articulatory planning and phonological grammar was beyond the scope of this dissertation, and remains to be explored. Perhaps the only possible conclusion at this stage is that underlying targets and phonological units (lexical accent) belong to separate domains. This is shown in Chapter 3, where the proposed [Falling] target is backed by acoustic and modelling evidence, and is otherwise ungrounded in terms of phonology (at least in currently prevailing theories). This mismatch suggests that phonology and articulatory targets are linked in intonation, but are not equivalent. This link, in turn, may be analogous to the relationship between morphemes and surface forms (Liu et al., 2013). That is, the pitch accent in Japanese phonology (i.e. a prosodic

function) has two ‘allophones’ in articulation, i.e. [High] and [Falling] (surface variants). Perception studies could be devised to shed more light on this.

As has been discussed in the preceding Chapters, this dissertation suffers from several limitations. The design of the stimuli, while successfully avoiding confounds from microprosody and global planning, led to unnatural production by the speakers. This has been reflected in the naturalness judgment ratings of interrogative focus. The syntax-prosody interface is another area that the present work has not touched upon due to time constraints. This is both an exciting topic and one that is less studied in PENTA, and thus could yield important insights for this line of research.

Needless to say, this work is far from complete. Other communicative functions that await investigation include emotion, alongside demarcation (phrasing and syntax). Another interesting topic would be the strategies that Japanese speakers use to enhance the attractiveness of their own voice (cf. Xu, Lee, Wu, Liu, & Birkholz, 2013). Once the prosodic manifestation of more communicative functions is better understood, it would be possible to use PENTA to model and predictively synthesise even spontaneous utterances in a wide range of contexts.

BIBLIOGRAPHY

't Hart, J., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.

Arisaka, H. (1941). アクセントの型の本質について [On the nature of accent shape]. *Gengo Kenkyu [言語研究]*, *7/8*, 83–92.

Arvaniti, A., & Ladd, D. R. (1995). Tonal alignment and the representation of accentual targets. In *Proceedings of the 13th International Congress of Phonetic Sciences (ICPhS 1995)*. Stockholm, Sweden.

Arvaniti, A., & Ladd, D. R. (2009). Greek wh-questions and the phonology of intonation. *Phonology*, *26*(01), 43–74.

Arvaniti, A., Ladd, D. R., & Mennen, I. (2006). Phonetic effects of focus and "tonal crowding" in intonation: Evidence from Greek polar questions. *Speech Communication*, *48*(6), 667–696.

Atkinson, J. E. (1978). Correlation analysis of the physiological factors controlling fundamental voice frequency. *Journal of the Acoustical Society of America*, *63*(1), 211–222.

Beckman, M. E. (1986). *Stress and non-stress accent*. Dordrecht: Foris.

Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 9–54). New York, NY: Oxford University Press.

Beckman, M. E., & Pierrehumbert, J. B. (1986a). Intonational structure in Japanese and English. *Phonology Yearbook*, *3*, 255–309.

Beckman, M. E., & Pierrehumbert, J. B. (1986b). Japanese prosodic phrasing and intonation synthesis. In *Proceedings of the 24th Annual Meeting on Association for Computational Linguistics (ACL1986)* (pp. 173–180). New York, NY.

Boersma, P. P. G., & van Heuven, V. J. J. P. (2001). Speak and unSpeak with PRAAT. *Glot International*, *5*(9/10), 341–347.

Calhoun, S. (2004). Overloaded ToBI and what to do about it: An Argument for function-based phonological intonation categories. In *Papers presented at the Linguistics Postgraduate Conference, University of Edinburgh*. Edinburgh.

Campbell, W. N., & Venditti, J. J. (1995). J-ToBI: An intonation labelling system for Japanese. In *Proceedings of the 1995 Autumn Meeting of the Acoustical Society of Japan [日本音響学会 1995 年秋季研究発表会講演論文集]* (pp. 317–318). Utsunomiya, Japan.

Caspers, J., & van Heuven, V. J. J. P. (1993). Effects of time pressure on the phonetic realization of the Dutch accent-lending pitch rise and fall. *Phonetica*, *50*(3), 161–171.

Chen, Y., & Xu, Y. (2006). Production of weak elements in speech: Evidence from F0 patterns of neutral tone in Standard Chinese. *Phonetica*, *63*(1), 47–75.

Cheng, C. (2012). *Mechanism of extreme phonetic reduction: Evidence from Taiwan Mandarin*. *PhD Thesis*. University College London, London.

Cheng, C., & Xu, Y. (2013). Articulatory limit and extreme segmental reduction in Taiwan Mandarin. *Journal of the Acoustical Society of America*, *134*(6), 4481–4495.

Cheng, C., & Xu, Y. (2014). Mechanism of disyllabic tonal reduction in Taiwan Mandarin. *Language and Speech*, 1–34.

Cho, H. (2011). Phrase positional effects on F0 peak timing in Tokyo Japanese. *Journal of the Korean Society of Speech Sciences [말소리와 음성과학]*, *3*(3), 69–75.

Chomsky, N. (1972). Deep structure, surface structure and semantic interpretation. In Danny D. Steinberg & L. A. Jakobovits (Eds.), *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology* (pp. 62–119). Cambridge: Cambridge University Press.

Chuenwattanapranithi, S., Xu, Y., Thipakorn, B., & Maneewongvatana, S. (2008). Encoding emotions in speech with the size code. *Phonetica*, *65*, 210–230.

Connell, B., & Ladd, D. R. (1990). Aspects of pitch realisation in Yoruba. *Phonology*, *7*(1), 1–29.

Cooper, W. E., & Sorensen, J. M. (1977). Fundamental frequency contours at syntactic boundaries. *Journal of the Acoustical Society of America*, *62*(3), 683–692.

Cooper, W. E., & Sorensen, J. M. (1981). *Fundamental frequency in sentence production*. New York, NY: Springer-Verlag.

Deguchi, M., & Kitagawa, Y. (2002). Prosody and wh-questions. In M. Hirotani (Ed.), *Proceedings of the 32nd Annual Meeting of the North Eastern Linguistic Society (NELS 32)* (pp. 73–92). Amherst, MA: GLSA Publications.

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

Eady, S. J., & Cooper, W. E. (1986). Speech intonation and focus location in matched statements and questions. *Journal of the Acoustical Society of America*, *80*(2), 402–415.

Erickson, D. M. (1976). *A physiological analysis of the tones of Thai*. *PhD Thesis*. University of Connecticut, Storrs, CT.

Folkins, J. W., Miller, C. J., & Minifie, F. D. (1975). Rhythm and syllable timing in phrase level stress patterning. *Journal of Speech and Hearing Research*, *18*, 739–753.

Fougeron, C., & Jun, S.-A. (1998). Rate effects on French intonation: Prosodic organization and phonetic realization. *Journal of Phonetics*, *26*(1), 45–69.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, *27*(4), 765–768.

Fujisaki, H. (1977). Functional models of articulatory and phonatory dynamic: Temporal organization of articulatory and phonatory controls in realization of Japanese word accent. In M. Sawashima & F. S. Cooper (Eds.), *Dynamic aspects of speech production: Current results, emerging problems, and new instrumentation*. Tokyo: University of Tokyo Press.

Fujisaki, H., & Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan [日本音響学会誌]*, *5*(4), 233–242.

Fujisaki, H., & Nagashima, S. (1969). A model for the synthesis of pitch contours of connected speech. *Annual Report of the Engineering Research Institute, University of Tokyo*, *28*, 53–60.

Fujisaki, H., Wang, C., Ohno, S., & Gu, W. (2005). Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command–response model. *Speech Communication*, *47*(1-2), 59–70.

Fung, S. H. H., & Mok, P. K. P. (2014). Realization of narrow focus in Hong Kong English declaratives: A pilot study. In *Proceedings of the 7th International Conference on Speech Prosody (SP2014)* (pp. 964–968). Dublin.

Gandour, J. T., Potisuk, S., & Dechongkit, S. (1994). Tonal coarticulation in Thai. *Journal of Phonetics*, *22*, 477–492.

Goldsmith, J. A. (1976). *Autosegmental Phonology*. *PhD Thesis*. Massachusetts Institute of Technology, Cambridge, MA.

Gu, W., & Lee, T. (2007). Effects of tonal context and focus on Cantonese F0. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007)* (pp. 1033–1036). Saarbrücken, Germany.

Guenther, F. H., & Perkell, J. S. (2004). A neural model of speech production and its application to studies of the role of auditory feedback in speech. In B. Maassen, R. D. Kent, H. Peters, P. van Lieshout, & W. Hulstijn (Eds.), *Speech motor control in normal and disordered speech* (pp. 29–49). Oxford: Oxford University Press.

Gussenhoven, C. H. M. (1999). On the limits of focus projection in English. In P. Bosch & R. van der Sandt (Eds.), *Focus: Linguistic, cognitive, and computational perspectives* (pp. 43–55). Cambridge University Press.

Gussenhoven, C. H. M. (2004). *The phonology of tone and intonation*. New York, NY: Cambridge University Press.

Haraguchi, S. (2002). Accent. In N. Tsujimura (Ed.), *The Handbook of Japanese Linguistics* (pp. 1–30). Malden, MA: Blackwell Publishers.

Hasegawa, Y., & Hata, K. (1995). The function of F0-peak delay in Japanese. In *Proceedings of the 21st Annual Meeting of the Berkeley Linguistics Society (BLS21)* (pp. 141–151). Berkeley, CA.

Higashikawa, M., Nakai, K., Sakakura, A., & Takahashi, H. (1996). Perceived pitch of whispered vowels--Relationship with formant frequencies: A preliminary study. *Journal of Voice*, *10*(2), 155–158.

Hirano, H., Nakamura, I., Minematsu, N., Suzuki, M., Nakagawa, C., Nakamura, N., … Hashimoto, H. (2013). A free online accent and intonation dictionary for teachers and learners of Japanese. In *Proceedings of Interspeech 2013* (pp. 1875–1876). Lyon, France.

Hirschberg, J. (2004). Pragmatics and intonation. In L. R. Horn & G. Ward (Eds.), *The handbook of pragmatics* (pp. 515–537). Malden, MA: Blackwell Publishing.

Hombert, J.-M., Ohala, J. J., & Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language*, *55*(1), 37–58.

Hwang, H. K. (2011). Distinct types of focus and wh-question intonation. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS 2011)* (pp. 922–925). Hong Kong.

Hyman, L. M., & Schuh, R. G. (1974). Universals of tone rules: Evidence from West Africa. *Linguistic Inquiry*, *5*(1), 81–115.

Ipek, C. (2011). Phonetic realization of focus with no on-focus pitch range expansion in Turkish. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS 2011)* (pp. 140–143). Hong Kong.

Ishihara, S. (2002). Invisible but audible wh-scope marking: Wh-constructions and deaccenting in Japanese. In L. Mikkelsen & C. Potts (Eds.), *Proceedings of the 21st West Coast Conference on Formal Linguistics (WCCFL 21)* (pp. 180–193). Somerville, MA: Cascadilla.

Ishihara, S. (2003). *Intonation and interface conditions*. *PhD Thesis*. Massachusetts Institute of Technology, Cambridge, MA.

Ishihara, S. (2007). Major phrase, focus intonation, multiple spell-out. *The Linguistic Review*, *24*, 137–167.

Ishihara, S. (2011a). Focus prosody in Tokyo Japanese wh-questions with lexically unaccented wh-phrases. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS 2011)* (pp. 946–949). Hong Kong.

Ishihara, S. (2011b). Japanese focus prosody revisited: Freeing focus from prosodic phrasing. *Lingua*, *121*(13), 1870–1889.

Ishihara, S. (2015). Syntax-phonology interface. In H. Kubozono (Ed.), *Handbook of Japanese Phonetics and Phonology*. Berlin: Mouton de Gruyter.

Ishihara, T. (2006). *Tonal alignment in Tokyo Japanese*. *PhD Thesis*. University of Edinburgh, Edinburgh.

Ito, K. (2002a). Ambiguity in broad focus and narrow focus interpretation in Japanese. In *Proceedings of the 1st International Conference on Speech Prosody (SP2002)* (pp. 411–414). Aix-en-Provence, France.

Ito, K. (2002b). *The interaction of focus and lexical pitch accent in speech production and dialogue comprehension: Evidence from Japanese and Basque*. *PhD Thesis*. University of Illinois, Urbana-Champaign, Urbana, IL.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680.

Kochanski, G. P., Grabe, E., Coleman, J. S., & Rosner, B. S. (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America*, *118*(2), 1038–1054. doi:10.1121/1.1923349

Kori, S. (2013). 判定要求の質問文における疑問型上昇調とその音声的特徴 [Interrogative rise in Tokyo Japanese]. *Studies in Language and Culture [言語文化研究]*, *39*, 221–244.

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica*, *55*(3), 243–276.

Kubozono, H. (1993). *The organization of Japanese prosody*. Tokyo: Kurosio.

Kubozono, H. (2007). Focus and intonation in Japanese: Does focus trigger pitch reset? *Interdisciplinary Studies on Information Structure (ISIS)*, *9*, 1–27.

Labrune, L. (2012). Questioning the universality of the syllable: Evidence from Japanese. *Phonology*, *29*(01), 113–152.

Laniran, Y. O. (1992). *Intonation in tone languages: The phonetic implementation of tones in Yoruba*. *PhD Thesis*. Cornell University.

Laniran, Y. O., & Clements, G. N. (2003). Downstep and high raising: Interacting factors in Yoruba tone production. *Journal of Phonetics*, *31*(2), 203–250.

Laniran, Y. O., & Gerfen, C. (1997). High raising, downstep and downdrift in Igbo. In *Paper presented at the 71st annual meeting of the Linguistic Society of Americ*. Chicago, IL.

Lee, A., & Xu, Y. (2012). Revisiting focus prosody in Japanese. In *Proceedings of the 6th International Conference on Speech Prosody (SP2012)* (pp. 274–277). Shanghai.

Lee, A., Xu, Y., & Prom-on, S. (2013). Mora-based pre-low raising in Japanese pitch accent. In *Proceedings of Interspeech 2013* (pp. 3532–3536). Lyon, France.

Lee, A., Xu, Y., & Prom-on, S. (2014). Modeling Japanese F0 contours using the PENTAtrainers and AMtrainer. In *Proceedings of the 4th International Symposium on Tonal Aspects of Languages (TAL 2014)* (pp. 164–167). Nijmegen.

Lee, Y.-C., & Xu, Y. (2010). Phonetic realization of contrastive focus in Korean. In *Proceedings of the 5th International Conference on Speech Prosody (SP2010)*. Chicago, IL.

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

Liu, F., & Xu, Y. (2005). Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica*, *62*, 70–87.

Liu, F., & Xu, Y. (2007). Question intonation as affected by word stress and focus in English. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007)* (pp. 1189–1192). Saarbrücken, Germany.

Liu, F., Xu, Y., Prom-on, S., & Yu, A. C. L. (2013). Morpheme-like prosodic functions: Evidence from acoustic analysis and computational modeling. *Journal of Speech Sciences*, *3*(1), 85–140.

Ma, K. Y. J., Ciocca, V., & Whitehill, T. L. (2006). Quantitative analysis of intonation patterns in statements and questions in Cantonese. In *Proceedings of the 3rd International Conference on Speech Prosody (SP2006)* (pp. 277–280). Dresden.

Maekawa, K. (1991). Perception of intonational characteristics of WH and non-WH questions in Tokyo Japanese. In *Proceedings of the 12th International Congress of Phonetic Sciences (ICPhS 2003)* (pp. 202–205). Aix-en-Provence, France.

Maekawa, K. (1997). Effects of focus on duration and vowel formant frequency in Japanese. In Y. Sagisaka, W. N. Campbell, & N. Higuchi (Eds.), *Computing prosody: Computational models for processing spontaneous speech* (pp. 129–153). New York, NY: Springer.

Maekawa, K. (1998). Phonetic and phonological characteristics of paralinguistic information in spoken Japanese. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 1998)*. Sydney.

Maekawa, K., Kikuchi, H., Igarashi, Y., & Venditti, J. J. (2002). X-JToBI: An extended J-ToBI for spontaneous speech. In *Proceedings of Interspeech 2002* (pp. 1545–1548). Denver, CO.

McCawley, J. D. (1968). *The phonological component of a grammar of Japanese*. The Hague: Mouton.

Myers, S. (2003). F0 timing in Kinyarwanda. *Phonetica*, *60*(2), 71–97.

Nakamura, I., Minematsu, N., Suzuki, M., Hirano, H., Nakagawa, C., Nakamura, N., … Hashimoto, H. (2013). Development of a web framework for teaching and learning Japanese prosody: OJAD (Online Japanese Accent Dictionary). In *Proceedings of Interspeech 2013* (pp. 2554–2558). Lyon, France.

Neustupný, J. V. (1966). 日本語のアクセントは高低アクセントか [Is Japanese accent a pitch accent?]. *Bulletin of the Phonetic Society of Japan [音声学会会報]*, *121*, 1–7.

O'Shaughnessy, D. D., & Allen, J. (1983). Linguistic modality effects on fundamental frequency in speech. *Journal of the Acoustical Society of America*, *74*(4), 1155–1171.

Ohala, J. J. (1978). Production of tone. In V. A. Fromkin (Ed.), *Tone: A linguistic survey* (pp. 5–39). New York, NY: Academic Press.

Ohde, R. N. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. *Journal of the Acoustical Society of America*, *75*(1), 224–230.

Öhman, S. E. G. (1967). Word and sentence intonation: A quantitative model. *Speech, Music and Hearing Quarterly Progress and Status Report (STL-QPSR)*, *8*(2-3), 20–54.

Oller, D. K., & Eilers, R. E. (1988). The role of audition in infant babbling. *Child Development*, *59*(2), 441–449.

Ostry, D. J., & Munhall, K. G. (1985). Control of rate and duration of speech movements. *Journal of the Acoustical Society of America*, *77*(2), 640–648.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. *PhD Thesis*. Massachusetts Institute of Technology, Cambridge, MA.

Pierrehumbert, J. B. (1981). Synthesizing intonation. *Journal of the Acoustical Society of America*, *70*(4), 985–995.

Pierrehumbert, J. B., & Beckman, M. E. (1988). *Japanese Tone Structure*. Cambridge, MA: Massachusetts Institute of Technology.

Poser, W. J. (1984). *The phonetics and phonology of tone and intonation in Japanese*. *PhD Thesis*. Massachusetts Institute of Technology, Cambridge, MA.

Prom-on, S., Liu, F., & Xu, Y. (2012). Post-low bouncing in Mandarin Chinese: Acoustic analysis and computational modeling. *Journal of the Acoustical Society of America*, *132*(1), 421–432.

Prom-on, S., & Xu, Y. (2012). PENTATrainer2: A hypothesis-driven prosody modeling tool. In *Proceedings of the 5th IESL Conference on Experimental Linguistics* (pp. 93–100). Athens, Greece.

Prom-on, S., Xu, Y., & Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America*, *125*(1), 405–424.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, *43*(3/4), 353–360.

Raidt, S., Bailly, G., Holm, B., & Mixdorff, H. (2004). Automatic generation of prosody: Comparing two superpositional systems. In *Proceedings of the 2nd International Conference on Speech Prosody (SP2004)*. Nara, Japan.

Raphael, L. J., Borden, G. J., & Harris, K. S. (2011). *Speech science primer: Physiology, acoustics, and perception of speech* (6th ed.). Baltimore, MD: Wolters Kluwer.

Rialland, A. (1981). Le système tonal du gurma (langue gur de Haute-Volta). *Journal of African Languages and Linguistics*, *3*, 39–64.

Rietveld, A. C. M., & Gussenhoven, C. H. M. (1987). Perceived speech rate and intonation. *Journal of Phonetics*, *15*, 273–285.

Saitou, T., Unoki, M., & Akagi, M. (2005). Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis. *Speech Communication*, *46*(3-4), 405–417.

Sapir, S. (1989). The intrinsic pitch of vowels: Theoretical, physiological, and clinical considerations. *Journal of Voice*, *3*(1), 44–51.

Sato, Y. (1993). The duration of syllable-final nasals and the mora hypothesis in Japanese. *Phonetica*, *50*, 44–67.

Sawashima, M., Kakita, Y., & Hiki, S. (1973). Activity of the extrinsic laryngeal muscles in relation to Japanese word accent. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics [東京大学医学部音声言語医学研究施設年報]*, *7*, 19–25.

Selkirk, E. O., Shinya, T., & Kawahara, S. (2004). Phonological and phonetic effects of Minor Phrase length on F0 in Japanese. In *Proceedings of the 2nd International Conference on Speech Prosody (SP2004)* (pp. 183–186). Nara, Japan.

Shadle, C. H. (1985). Intrinsic fundamental frequency of vowels in sentence context. *Journal of the Acoustical Society of America*, *78*(5), 1562–1567.

Simada, Z., & Hirose, H. (1970). The function of the laryngeal muscles in respect to the word accent distinction. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics [東京大学医学部音声言語医学研究施設年報]*, *4*, 27–40.

Snider, K. L. (1998). Phonetic realisation of downstep in Bimoba. *Phonology*, *15*, 77–101.

Steele, S. A. (1986). Interaction of vowel F0 and prosody. *Phonetica*, *43*, 92–105.

Sugahara, M. (2002). Conditions on post-focus dephrasing in Tokyo Japanese. In *Proceedings of the 1st International Conference on Speech Prosody (SP2002)*. Aix-en-Provence, France.

Sugahara, M. (2003a). *Downtrends and post-focus intonation in Tokyo Japanese*. *PhD Thesis*. University of Massachusetts, Amherst, Amherst, MA.

Sugahara, M. (2003b). The tone-bound declination slope: Evidence from Japanese. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003)* (pp. 2613–2616). Barcelona.

Sugito, M. (1968). 東京二拍語尾高と平板アクセント考 [Comparison between “high final” and “level” tone in Tokyo Japanese]. *Bulletin of the Phonetic Society of Japan [音声学会会報]*, *129*, 1–4.

Sugito, M. (1982). *日本語アクセントの研究 [Research on Japanese accent]*. Tokyo: Sanseido [三省堂].

Sugito, M. (2003). Timing relationships between prosodic and segmental control in Osaka Japanese word accent. *Phonetica*, *60*(1), 1–16.

Sugiyama, Y. (2012). *The production and perception of Japanese pitch accent*. Newcastle upon Tyne, England: Cambridge Scholars Publishing.

Sugiyama, Y., & Moriyama, T. (2013). Do formant frequencies correlate with Japanese accent? In *Proceedings of the 21st International Congress on Acoustics (ICA2013)* (Vol. 19, p. 060196). Montréal.

Sun, X. (2002). *The determination, analysis, and synthesis of fundamental frequency*. *PhD Thesis*. Northwestern University, Evanston, IL.

Taheri Ardali, M., Xu, Y., & Rahmani, H. (2014). The perception of prosodic focus in Persian. In *Proceedings of the 7th International Conference on Speech Prosody (SP2014)* (pp. 515–519). Shanghai.

Taylor, P. (1994). The rise/fall/connection model of intonation. *Speech Communication*, *15*(1-2), 169–186.

Truckenbrodt, H. (2004). Final lowering in non-final position. *Journal of Phonetics*, *32*(3), 313–348.

Turco, G., Dimroth, C., & Braun, B. (2012). Intonational means to mark verum focus in German and French. *Language and Speech*, *56*(4), 461–491.

Vance, T. J. (1995). Final accent vs. no accent: Utterance-final neutralization in Tokyo Japanese. *Journal of Phonetics*, *23*(4), 487–499.

Venditti, J. J. (2005). The J_ToBI model of Japanese intonation. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 172–200). New York, NY: Oxford University Press.

Venditti, J. J., Maekawa, K., & Beckman, M. E. (2008). Prominence marking in the Japanese intonation system. In S. Miyagawa & M. Saito (Eds.), *The Oxford handbook of Japanese linguistics* (pp. 456–512). New York, NY: Oxford University Press.

Venditti, J. J., & van Santen, J. P. H. (2000). Japanese intonation synthesis using superposition and linear alignment models. In *Proceedings of Interspeech 2000* (pp. 605–608). Beijing.

Wang, B., & Xu, Y. (2011). Differential prosodic encoding of topic and focus in sentence-initial position in Mandarin Chinese. *Journal of Phonetics*, *39*(4), 595–611.

Warner, N. (1997). Japanese final-accented and unaccented phrases. *Journal of Phonetics*, *25*(1), 43–60.

Warner, N., & Arai, T. (2001). Japanese mora-timing: A review. *Phonetica*, *58*, 1–25.

Weismer, G. G., & Ingrisano, D. (1979). Phrase-level timing patterns in English: Effects of emphatic stress location and speaking rate. *Journal of Speech and Hearing Research*, *22*, 516–533.

Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics*, *27*(2), 349–366.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, *25*(1), 61–83.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics*, *27*(1), 55–105.

Xu, Y. (2001). Fundamental frequency peak delay in Mandarin. *Phonetica*, *58*(1-2), 26–52.

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, *46*(3-4), 220–251.

Xu, Y. (2006). Principles of tone research. In *Proceedings of the 2nd International Symposium on Tonal Aspects of Languages (TAL 2006)* (pp. 3–13). La Rochelle, France.

Xu, Y. (2007). Speech as articulatory encoding of communicative functions. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007)* (pp. 25–30). Saarbrücken, Germany.

Xu, Y. (2011a). Post-focus compression: Cross-linguistic distribution and historical origin. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS 2011)* (pp. 152–155). Hong Kong.

Xu, Y. (2011b). Speech prosody: A methodological review. *Journal of Speech Sciences*, *1*(1), 85–115.

Xu, Y. (2013). ProsodyPro: A tool for large-scale systematic prosody analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)* (pp. 7–10). Aix-en-Provence, France.

Xu, Y., Chen, S.-W., & Wang, B. (2012). Prosodic focus with and without post-focus compression: A typological divide within the same language family? *The Linguistic Review*, *29*(1), 131–147.

Xu, Y., Lee, A., Wu, W.-L., Liu, X., & Birkholz, P. (2013). Human vocal attractiveness as signaled by body size projection. *PLoS One*, *8*(4), e62397. Retrieved from http://www.plosone.org/article/info:doi/10.1371/journal.pone.0062397

Xu, Y., & Prom-On, S. (2010-2015). PENTAtrainer1.praat. Available from: <http://www.phon.ucl.ac.uk/home/yi/PENTAtrainer1/>.

Xu, Y., & Prom-on, S. (2014). Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Communication*, *57*, 181–208.

Xu, Y., & Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America*, *111*(3), 1399–1413.

Xu, Y., & Wang, M. (2009). Organizing syllables into groups: Evidence from F0 and duration patterns in Mandarin. *Journal of Phonetics*, *37*(4), 502–520.

Xu, Y., & Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, *33*(4), 319–337.

Xu, Y., & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*, *33*(2), 159–197.

Xu, Y., Xu, C. X., & Sun, X. (2004). On the temporal domain of focus. In *Proceedings of the 2nd International Conference on Speech Prosody (SP2004)* (pp. 81–84). Nara, Japan.

Yip, M. J. W. (2002). *Tone*. New York, NY: Cambridge University Press.

Yoshida, S. (1990). A Government-based analysis of the “mora” in Japanese. *Phonology*, *7*(2), 331–351.

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

Yuan, J. (2006). Mechanisms of question intonation in Mandarin. In *Proceedings of the 5th International Symposium on Chinese Spoken Language Processing (ISCSLP 2006)* (pp. 19–30). Singapore.

APPENDICES

Appendix 1: F0 contours - statement vs. questions

1.1 Description of data

In Appendice 1.2, 1.3, 1.4, averaged F0 contours of 48 target sentences spoken in question vs. statement and narrow focus vs. neutral focus are presented. Below is the naming scheme of the target sentences used in Appendix 1 and Chapters 4 and 5. The name of a given panel consists of five digits, for example I111Q:

I	Focus condition
1	Word I (1~4)
1	Word II (1~4)
1	Word III (1~2)
Q	Sentence type (question or statement)

The gloss to Words I, II, III is as follows:

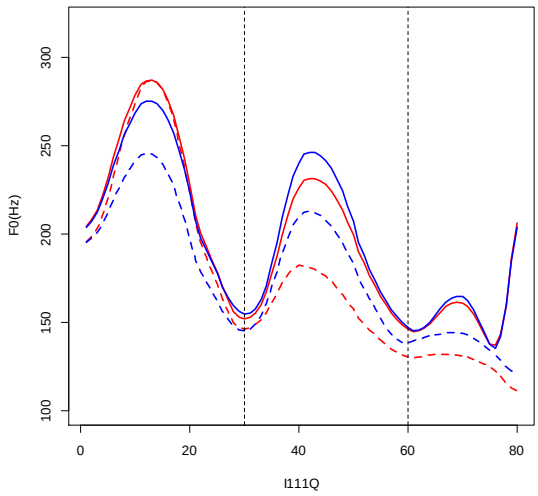
			Word I			Word II			Word III
Short	Accented	**1**	**me**i-ga May が May-NOM	×	**1**	**mo**mo 腿 thigh	×	**1**	-o **mi**ta を見た -ACC saw
	Unaccented	2	mei-ga 侄が Niece-NOM		2	momo 桃 peach		2	-ni nita に似た -DAT resembled
Long	Accented	**3**	**muu**min-ga ムーミンが Moomin-NOM	×	**3**	**bu**dou 武道 martial arts	×	**1**	-o **mi**ta を見た -ACC saw
	Unaccented	4	noumin-ga 农民が Farmer-NOM		4	budou 葡萄 grapes		2	-ni nita に似た -DAT resembled

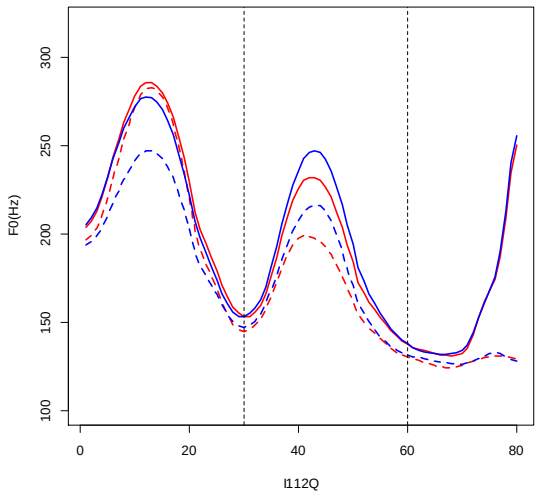
Thus panel I111Q stands for the question ***me****i-ga* ***mo****mo-o* ***mi****ta?* under initial focus. In Appendices 1.2, 1.3, 1.4, the panel of a given question contains also the corresponding question under neutral focus, as well as the corresponding statements under narrow and neutral focus. The color of the curves in Appendices 1.2, 1.3, 1.4 represents:

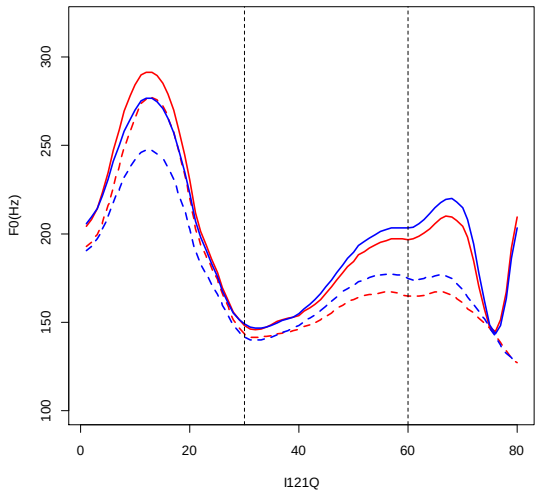
---Statement, narrow focus
—Question, narrow focus
---Statement, neutral focus
—Question, neutral focus

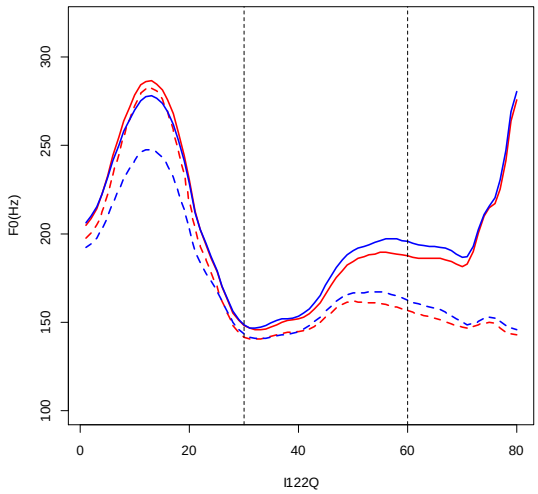
Panels 111Q and 222Q below were also presented in **Figure 38**.

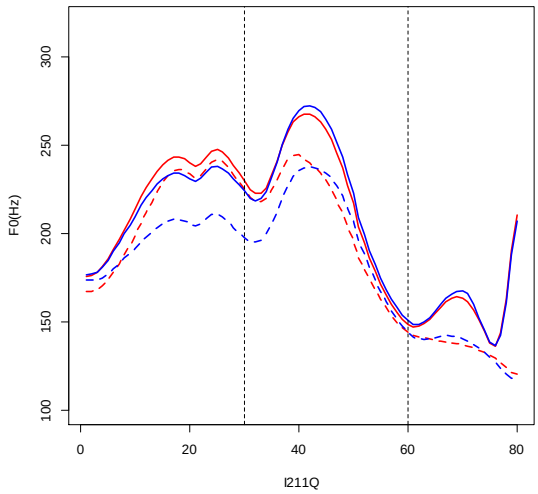
1.2 Initial focus

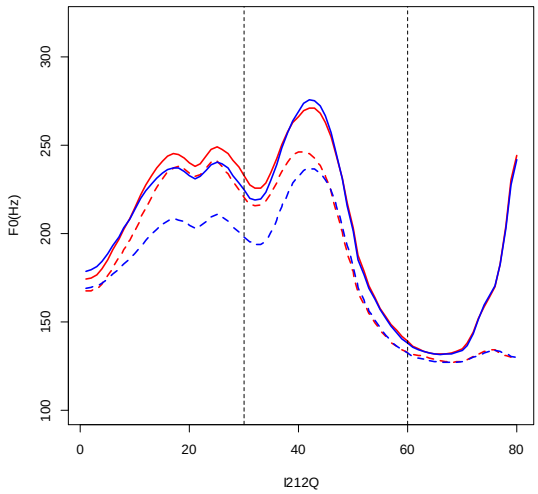


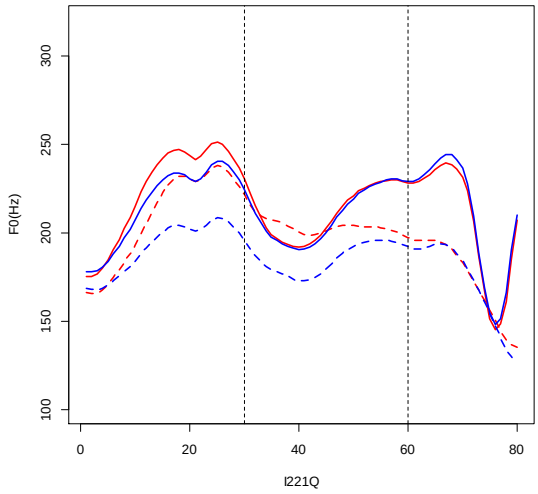


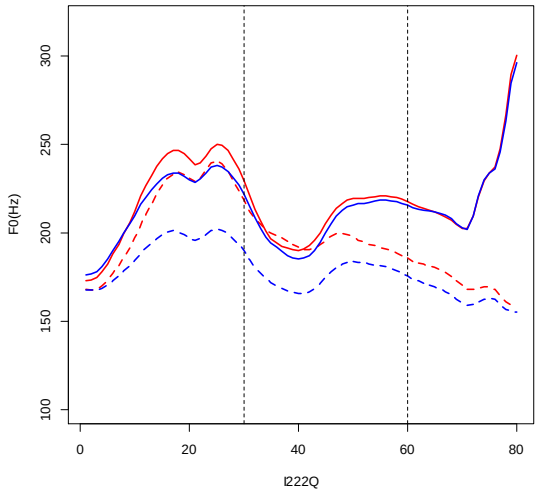


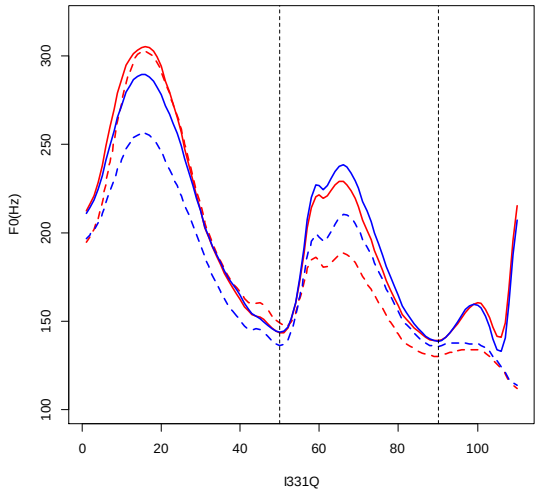


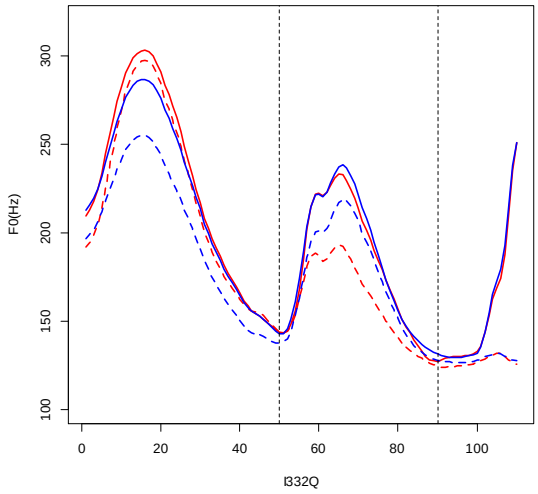


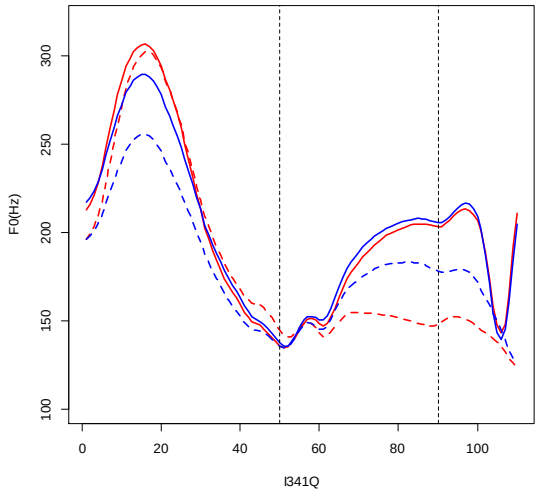


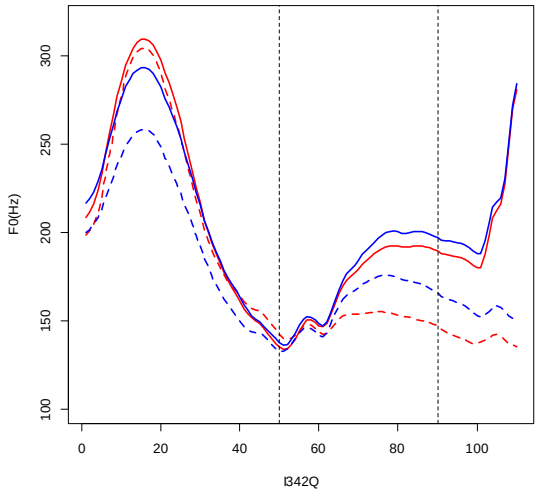

F0(Hz)
100
150
200
250
300
0
20
40
60
80
I221Q

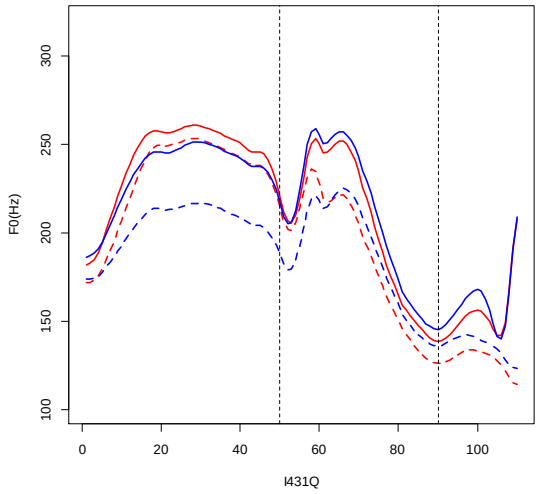

F0(Hz)
100
150
200
250
300
0
20
40
60
80
I222Q

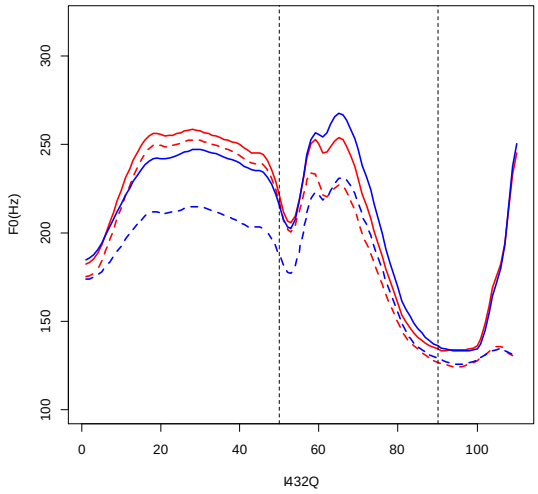

F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
I331Q

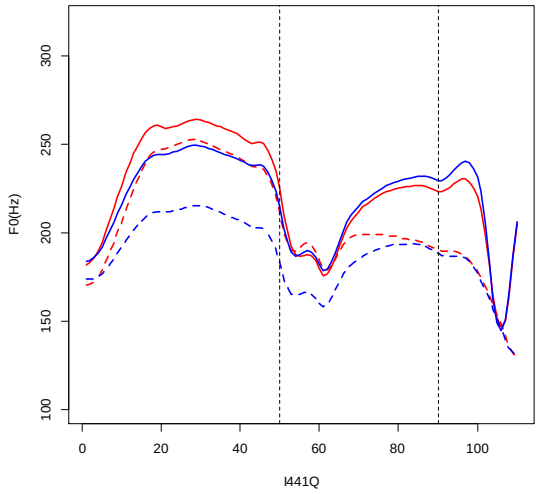

F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
I332Q

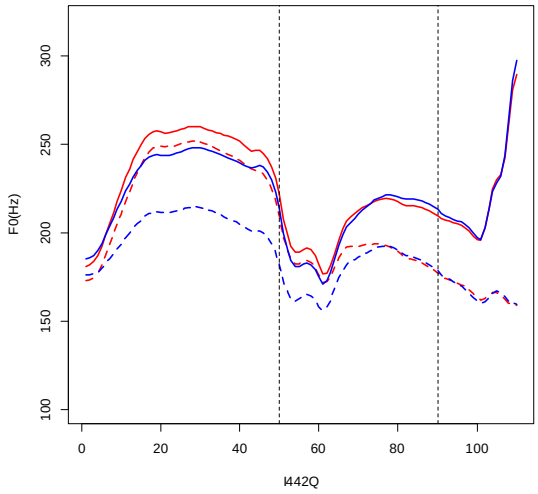

F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
I341Q


F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
I342Q

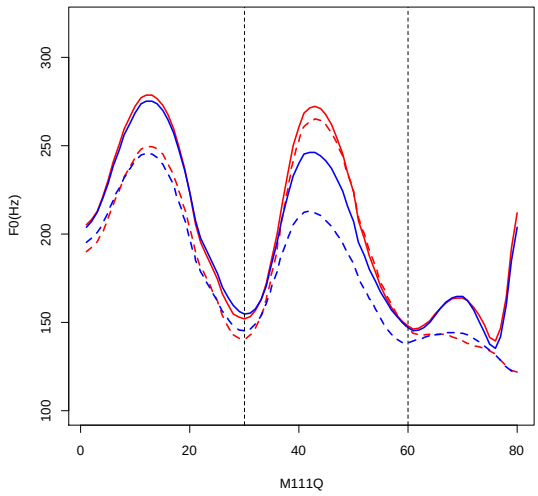

F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
l431Q

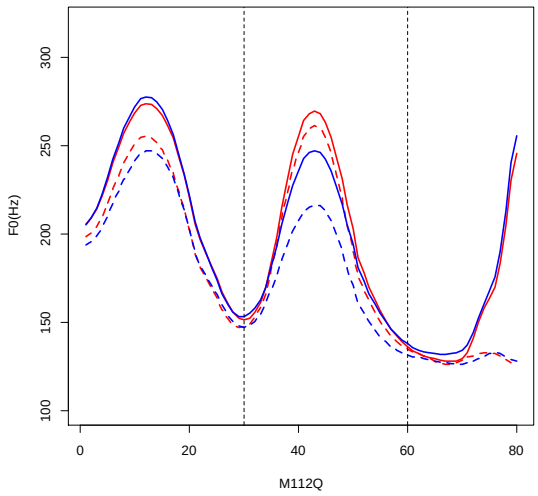

F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
l432Q

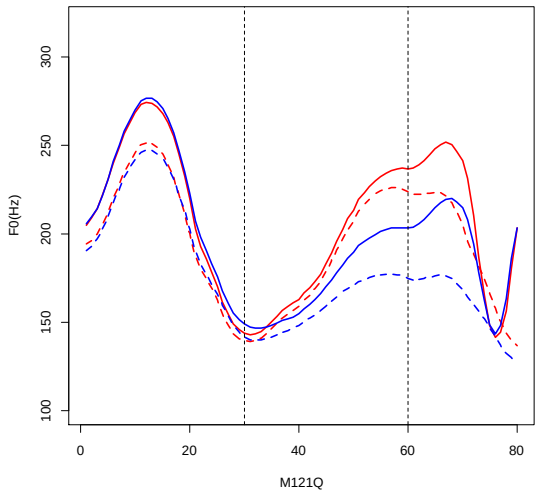

F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
l441Q

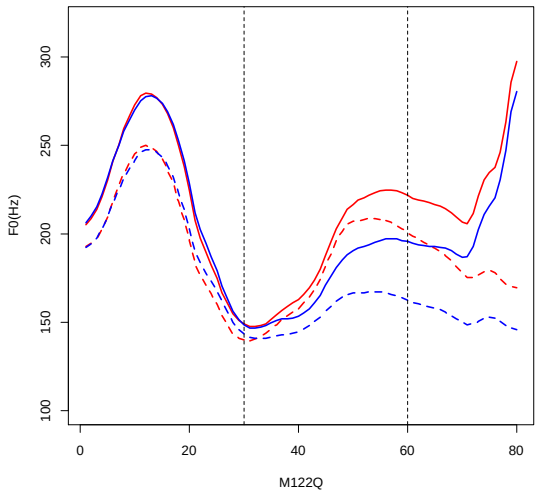

F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
l442Q

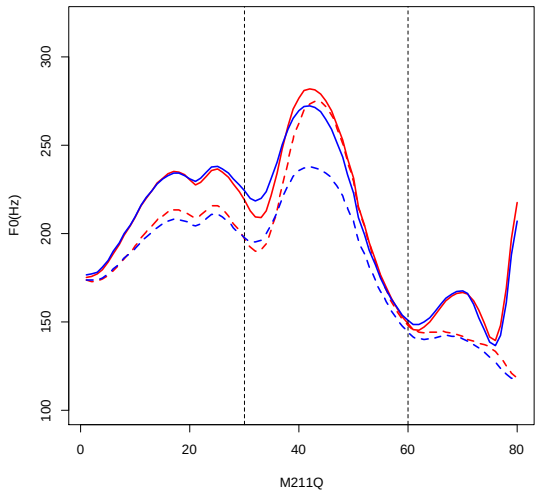
1.3 Medial focus

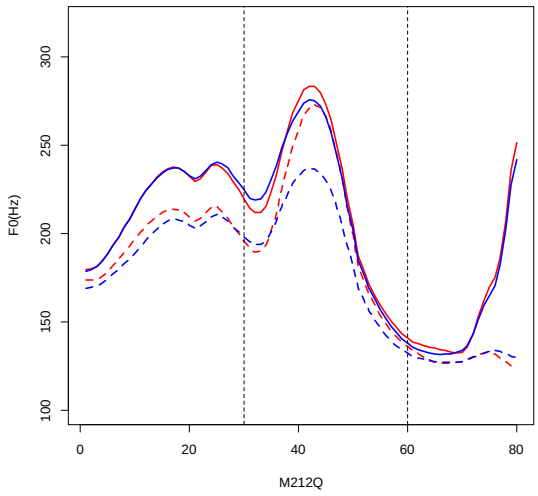


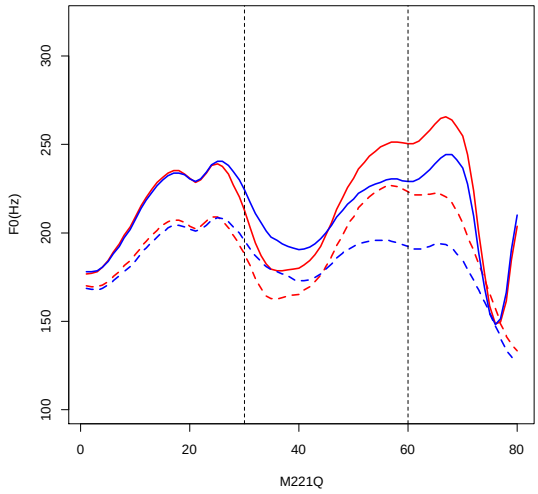


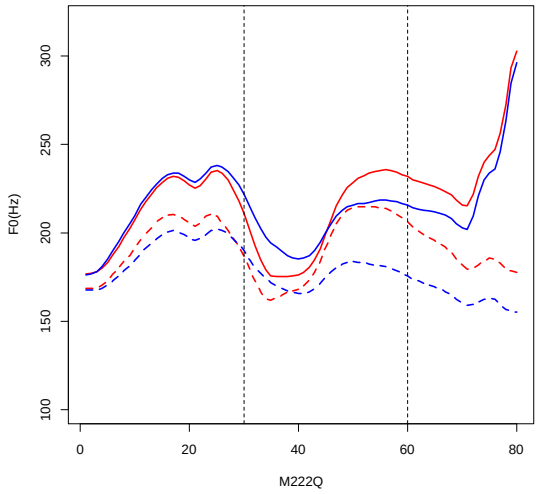


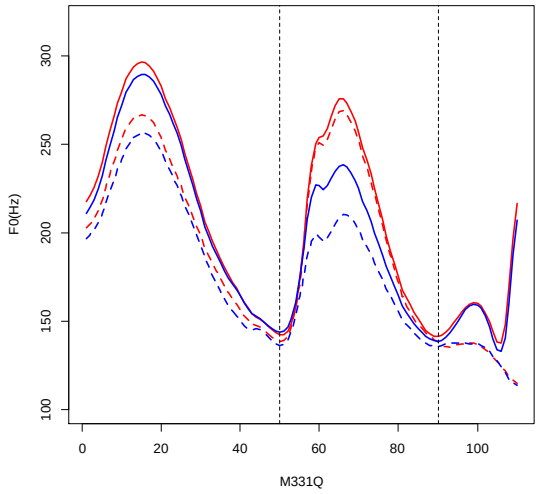


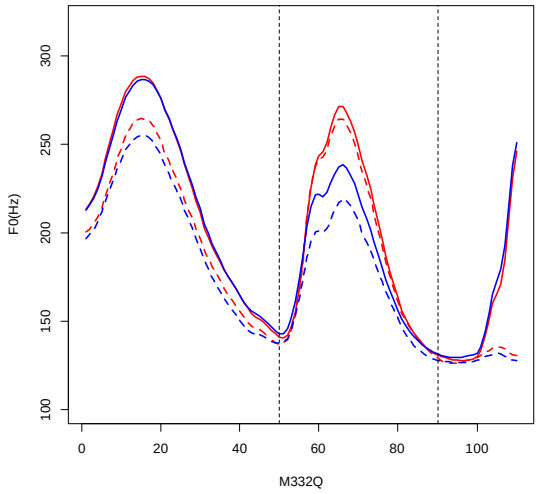


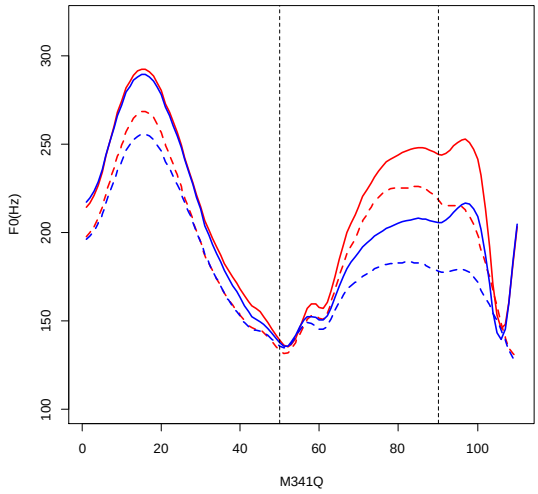


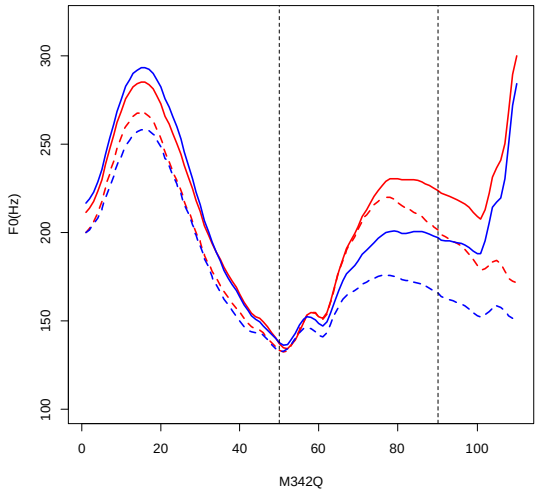

F0(Hz)
100
150
200
250
300
0
20
40
60
80
M221Q

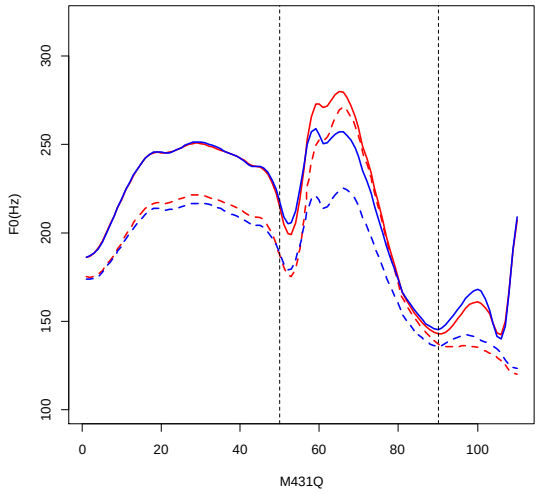

F0(Hz)
100
150
200
250
300
0
20
40
60
80
M222Q

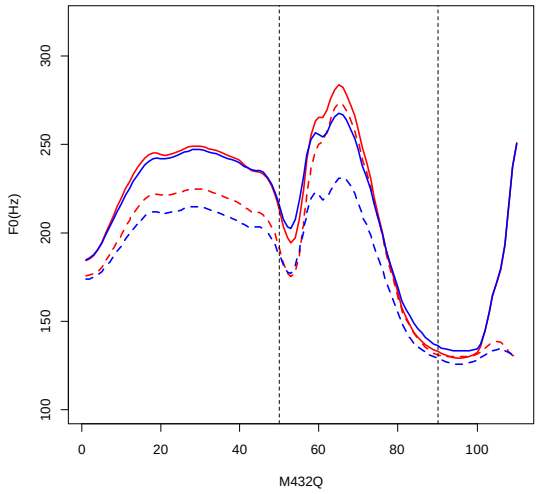

F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
M331Q

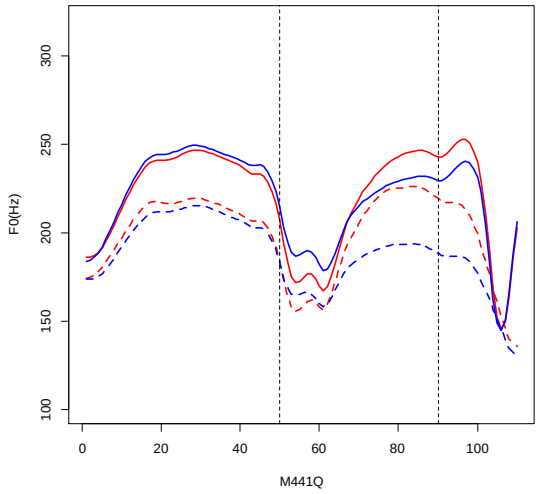

F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
M332Q

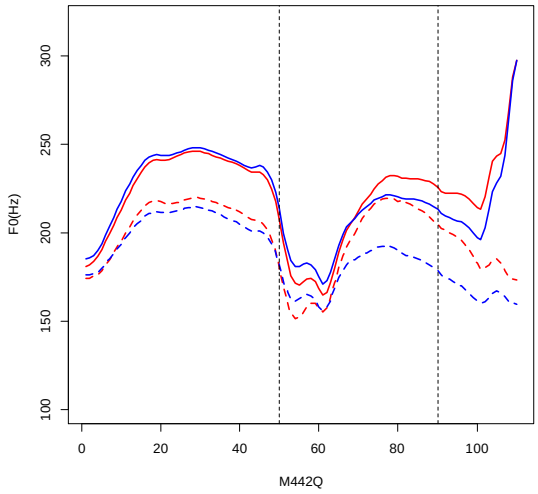

F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
M341Q


F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
M342Q

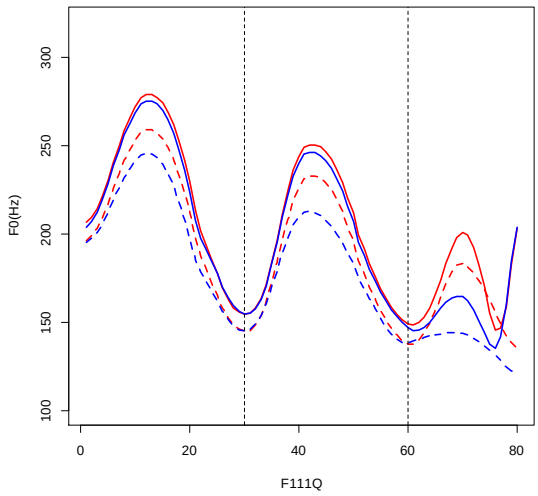

F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
M431Q

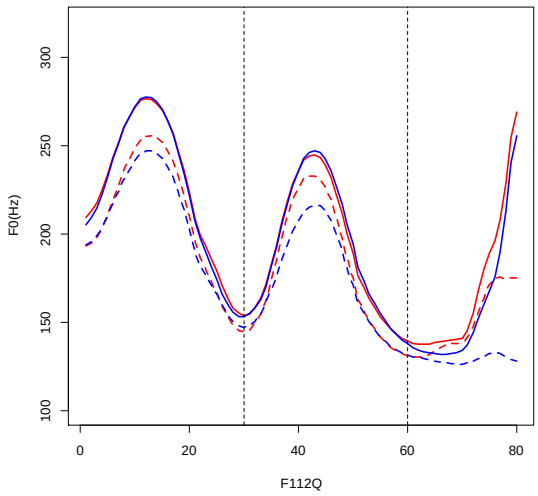

F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
M432Q

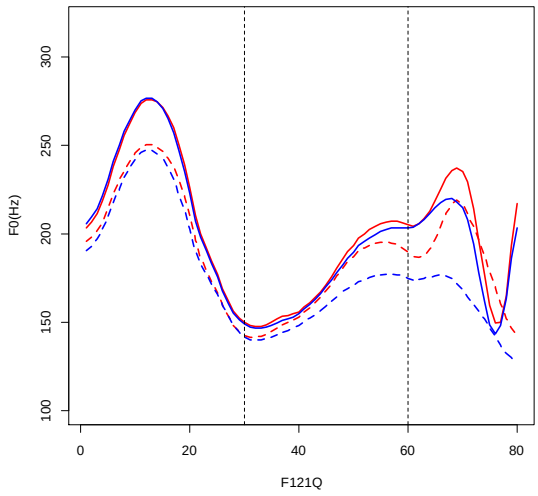

F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
M441Q

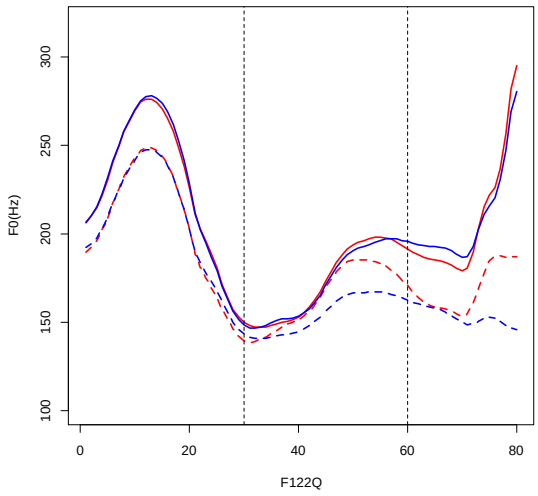

F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
M442Q

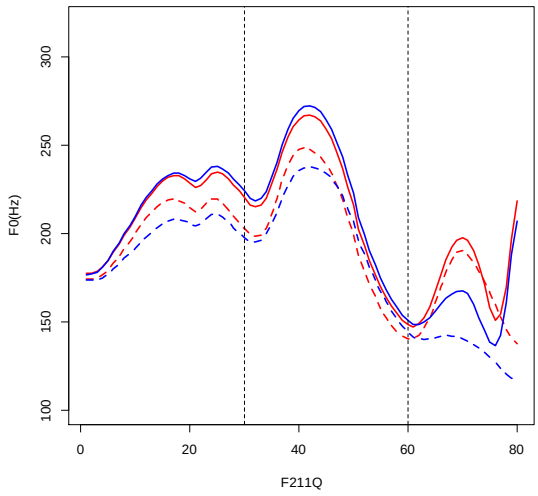
1.4 Final focus

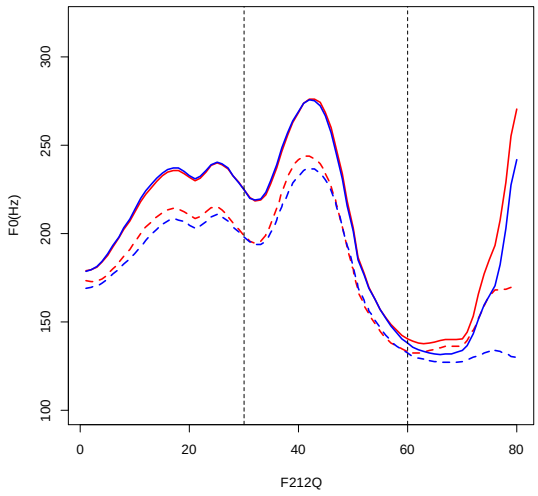


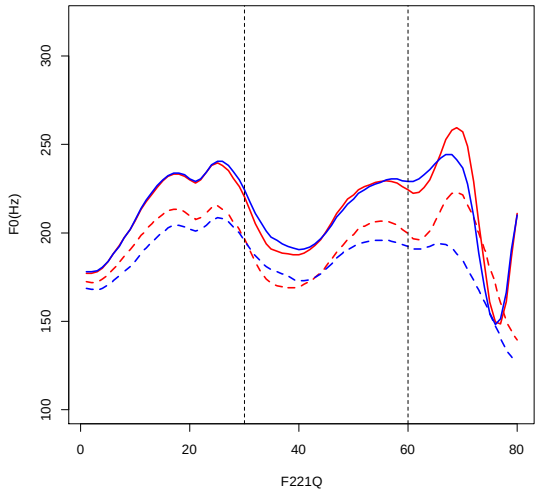


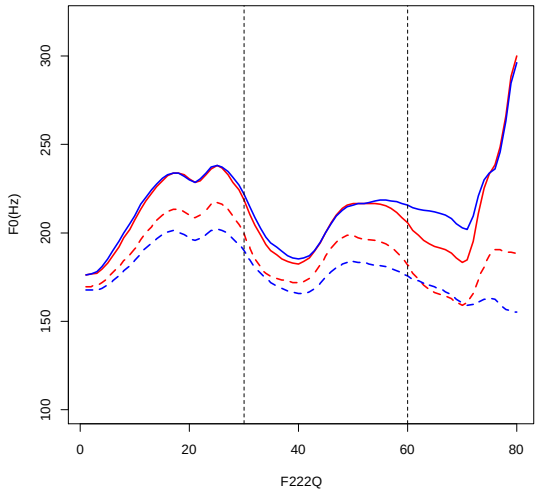


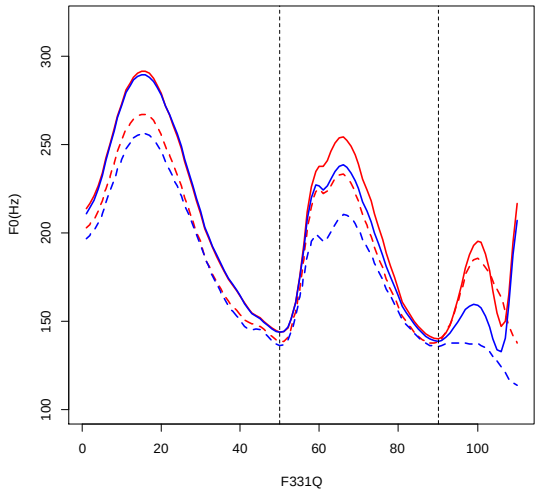


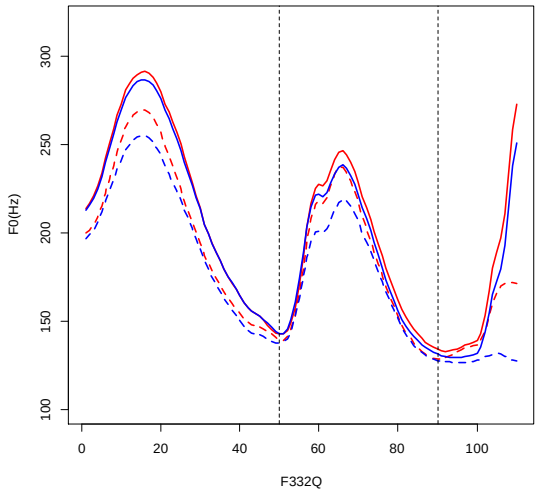


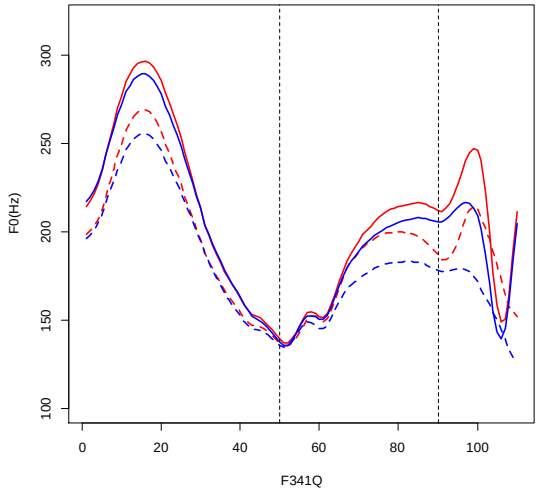


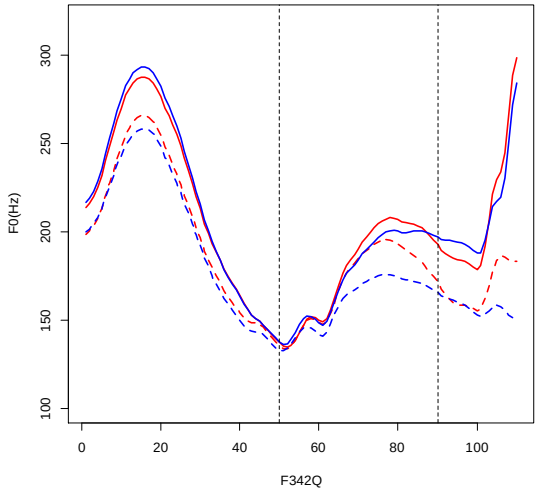

F0(Hz)
100
150
200
250
300
0
20
40
60
80
F221Q

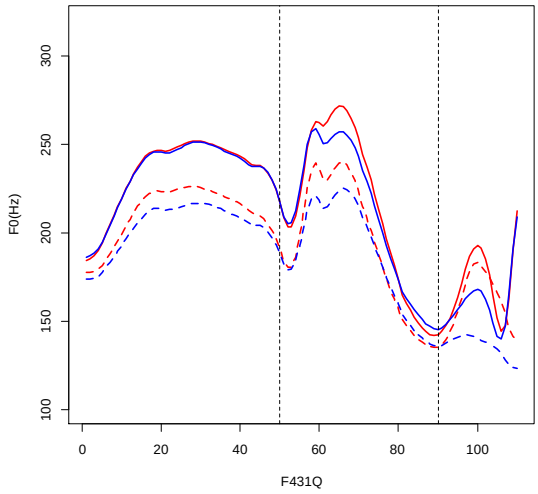

F0(Hz)
100
150
200
250
300
0
20
40
60
80
F222Q

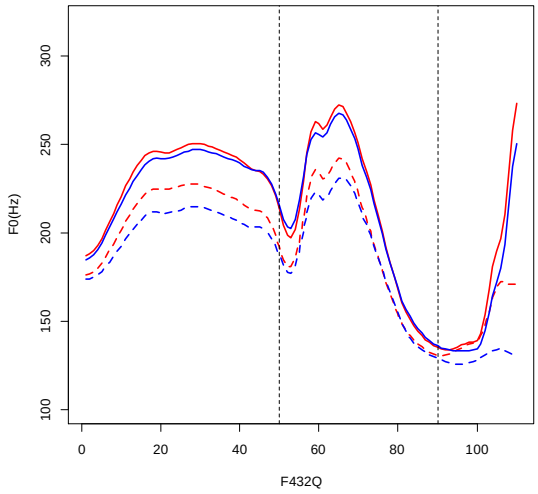

F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
F331Q

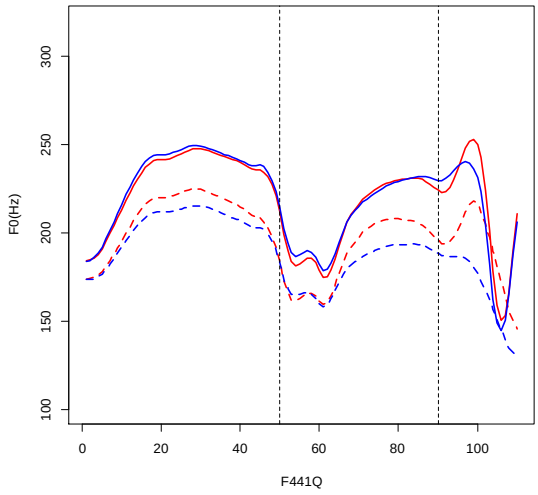

F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
F332Q

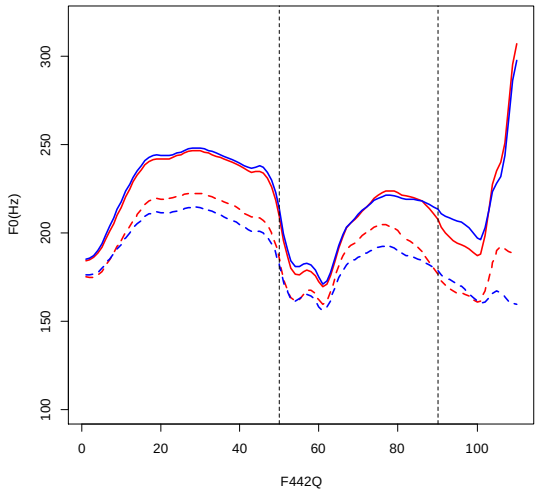

F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
F341Q


F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
F342Q


F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
F431Q


F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
F432Q


F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
F441Q


F0(Hz)
100
150
200
250
300
0
20
40
60
80
100
F442Q

Appendix 2: Global articulatory parameters of lexical corpus

2.1 Description of data

Below are the mean articulatory parameters averaged for each combination of communicative functions in the corpus described in Chapters 2. Parametric values presented below were extracted by PENTAtrainer2 and averaged across eight speakers. Mean parametric values from PENTAtrainer1 are omitted due to poor predictive synthesis performance in §6.2.3.3. See §6.2.2 for details on the meaning of the functional labels and how the model was trained.

2.2 Moraic segmentation

Demarcation	Tone	Slope	Height	Strength
C1	M	-34.5863	0.21875	13.0025
C2	M	30.0025	-10.6138	12.0475
L	H	8.85625	0.4925	19.915
L	M	34.4025	-4.2975	41.73375
LP	H	37.725	1.08	28.14875
M	H	-25.19	3.86625	***19.34***
M	L	-24.0338	-5.19125	17.115
M	M	-1.3025	-0.66	31.19
P	H	-0.115	-2.1925	***35.215***
R	L	-8.14875	-8.52	31.48375
R	M	8.46625	-6.04125	22.3
R	M	8.46625	-6.04125	22.3

2.3 Syllabic segmentation

Demarcation	Tone	Slope	Height	Strength
C1	M	20.225	-8.1025	9.34625
C2	M	29.91625	-13.0925	9.61375
L	H	-10.4425	5	32.40375
L	M	31.53625	-4.92375	46.54625
LL	F	-56.9313	7.29	12.12375
LL	H	26.02	1.1325	19.51375
LL	M	27.9025	5.5	40.1525
LP	H	48.92	-1.01625	15.26375
M	H	0.22375	-0.0075	***25.13625***
M	L	-17.4225	-6.9925	18.01
M	M	2.53875	-2.39375	19.26375
P	F	-51.6938	0.54	20.5325
P	H	10.0375	-2.9125	***37.51875***
R	L	-6.805	-8.64	30.73875
R	M	-17.7275	-2.58375	10.8075

Appendix 3: Global articulatory parameters of sentential corpus

3.1 Description of data

Below are the mean articulatory parameters averaged for each combination of communicative functions in the corpus described in Chapters 4 and 5. Parametric values presented below were extracted by PENTAtrainer2 and averaged across 10 speakers. See §6.3.2 for details on the meaning of the functional labels and how the model was trained.

3.2 Moraic segmentation

Sentence type	Demarcation	Tone	Focus	Slope	Height	Strength
Q	L	H	NF	-77.091	9.958	27.407
Q	L	H	ON	-28.584	8.461	29.174
Q	L	H	PO	-48.539	9.326	22.289
Q	L	H	PR	-52.503	8.856	34.075
Q	L	M	NF	42.145	-2.073	12.698
Q	L	M	ON	38.241	-0.826	15.925
Q	L	M	PO	46.142	-4.982	15.084
Q	L	M	PR	17.271	0.633	16.977
Q	M	L	NF	-50.339	-0.042	19.359
Q	M	L	ON	-39.648	-1.703	18.03
Q	M	L	PO	-41.961	-2.873	25.765
Q	M	L	PR	-39.42	1.079	16.522
Q	M	M	NF	16.346	1.402	18.439
Q	M	M	ON	19.73	1.278	23.976
Q	M	M	PO	-25.123	-2.981	12.792
Q	M	M	PR	6.24	2.957	20.17
Q	R	L	NF	-10.676	-6.158	62.201
Q	R	L	ON	1.323	-5.546	53.725
Q	R	L	PO	-34.014	-4.854	30.18
Q	R	L	PR	7.777	-6.889	50.349
Q	R	M	NF	-16.821	-0.204	14.483
Q	R	M	ON	6.013	-2.496	22.056
Q	R	M	PO	-4.569	4.052	21.619
Q	R	M	PR	-29.811	-0.467	20.586
Q	SL	H	NF	27.12	5.416	23.515
Q	SL	H	ON	35.437	8.293	27.324
Q	SL	H	PR	9.538	8.36	22.344
Q	SL	M	NF	36.867	3.922	21.113
Q	SL	M	ON	17.013	7.574	21.354
Q	SL	M	PR	18.071	4.953	17.665
Q	SR	L	NF	56.877	2.147	60.973
Q	SR	L	ON	53.712	0.323	52.207
Q	SR	L	PO	65.414	4.066	48.895
Q	SR	M	NF	72.217	4.671	15.298
Q	SR	M	ON	57.505	7.391	36.07
Q	SR	M	PO	44.368	4.044	16.215
S	L	H	NF	-39.089	8.01	14.656
S	L	H	ON	-20.457	7.176	29.718
S	L	H	PO	-39.262	11.224	15.129
S	L	H	PR	-25.483	6.282	37.626
S	L	M	NF	-12.866	-1.699	22.268
S	L	M	ON	37.878	3.458	23.585

S	L	M	PO	16.876	-6.27	17.214
S	L	M	PR	29.173	2.336	20.647
S	M	L	NF	-37.222	-1.918	20.627
S	M	L	ON	-39.972	-0.957	16.455
S	M	L	PO	-61.589	-2.835	22.721
S	M	L	PR	-28.891	-3.117	20.5
S	M	M	NF	-0.143	-0.006	19.677
S	M	M	ON	4.796	3.29	30.384
S	M	M	PO	-30.675	1.347	24.35
S	M	M	PR	-10.098	2.116	24.914
S	R	L	NF	8.822	-6.879	53.004
S	R	L	ON	4.726	-8.527	60.761
S	R	L	PO	-20.412	1.554	23.095
S	R	L	PR	-23.059	-4.55	47.736
S	R	M	NF	-5.981	0.046	22.627
S	R	M	ON	6.288	0.29	27.81
S	R	M	PO	-9.283	2.844	19.568
S	R	M	PR	-36.26	-4.482	17.998
S	SL	H	NF	29.183	3.82	16.984
S	SL	H	ON	21.061	8.503	26.26
S	SL	H	PR	24.748	6.282	23.43
S	SL	M	NF	27.173	2.556	16.903
S	SL	M	ON	50.348	2.798	17.202
S	SL	M	PR	12.147	4.182	17.107
S	SR	L	NF	1.475	-5.605	90
S	SR	L	ON	-15.898	-6.478	92.239
S	SR	L	PO	-0.419	-7.524	74.856
S	SR	M	NF	10.208	-7.743	36.486
S	SR	M	ON	-7.979	-0.904	51.263
S	SR	M	PO	-29.379	-6.574	22.15

3.3 Syllabic segmentation

Sentence type	Demarcation	Tone	Focus	Slope	Height	Strength
Q	L	H	NF	-48.656	7.293	27.323
Q	L	H	ON	-31.128	8.123	27.854
Q	L	H	PO	-48.808	10.895	22.478
Q	L	H	PR	-20.381	8.858	33.362
Q	L	M	NF	16.308	2.808	19.22
Q	L	M	ON	27.25	3.691	11.332
Q	L	M	PO	18.603	1.645	14.598
Q	L	M	PR	46.634	-3.695	23.409
Q	M	L	NF	-63.942	-2.134	49.06
Q	M	L	ON	-60.735	-3.114	32.559
Q	M	L	PO	-41.759	-1.972	31.471
Q	M	L	PR	-52.766	-4.766	22.6
Q	M	M	NF	3.313	0.618	14.521
Q	M	M	ON	12.242	0.436	24.75
Q	M	M	PO	0.062	1.406	24.504
Q	M	M	PR	-13.231	5.273	15.762
Q	R	L	NF	-10.666	-4.102	58.557
Q	R	L	ON	-13.549	-4.087	59.608
Q	R	L	PO	-16.603	-6.789	42.654
Q	R	L	PR	-7.135	-4.252	64.78
Q	R	M	NF	-15.858	-1.389	15.859
Q	R	M	ON	-15.41	0.108	23.905
Q	R	M	PO	-36.245	6.89	24.216
Q	R	M	PR	-35.37	-0.095	19.227
Q	SL	F	NF	-45.674	2.385	19.717

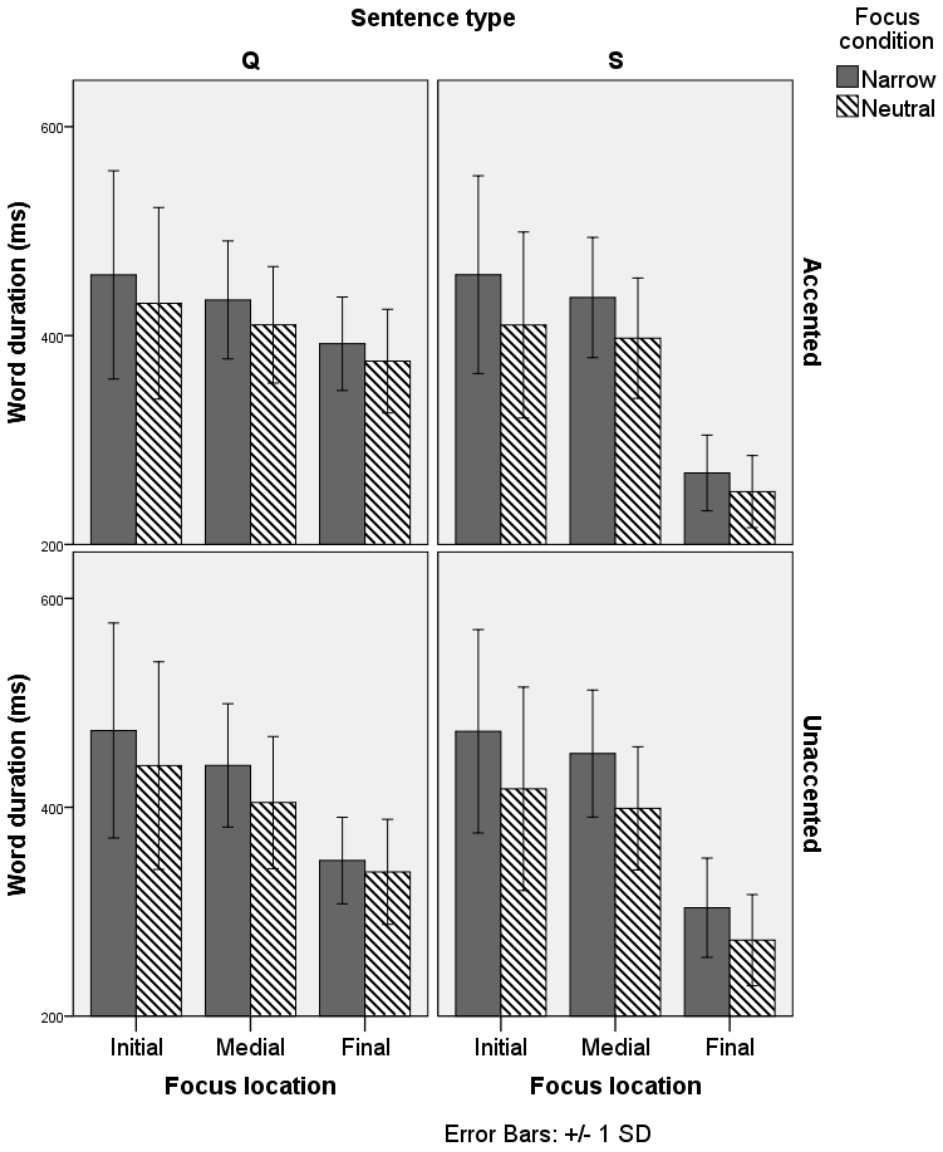
Q	SL	F	ON	-58.124	5.902	26.601
Q	SL	F	PR	-40.661	2.312	15.899
Q	SL	M	NF	-7.107	6.119	18.442
Q	SL	M	ON	-7.287	6.455	17.985
Q	SL	M	PR	6.576	1.935	14.093
Q	SR	L	NF	69.715	4.684	42.752
Q	SR	L	ON	59.412	4.726	46.45
Q	SR	L	PO	61.841	3.814	35.863
Q	SR	M	NF	76.007	6.2	13.82
Q	SR	M	ON	42.552	7.835	25.528
Q	SR	M	PO	39.47	5.973	13.483
S	L	H	NF	-42.006	15.935	18.335
S	L	H	ON	-33.093	11.848	23.822
S	L	H	PO	-51.425	11.939	15.455
S	L	H	PR	2.901	5.41	39.636
S	L	M	NF	10.306	4.893	13.869
S	L	M	ON	30.221	1.313	29.611
S	L	M	PO	-2.101	-0.461	10.812
S	L	M	PR	6.369	0.713	16.065
S	M	L	NF	-58.617	-3.952	39.403
S	M	L	ON	-52.744	-4.688	31.436
S	M	L	PO	-37.278	-6.365	30.581
S	M	L	PR	-51.815	-1.667	29.104
S	M	M	NF	-27.203	-2.354	19.25
S	M	M	ON	4.766	-1.231	25.149
S	M	M	PO	-23.658	1.831	15.25
S	M	M	PR	-18.383	2.546	21.464
S	R	L	NF	-4.872	-3.778	42.079
S	R	L	ON	-4.901	-6.542	48.431
S	R	L	PO	1.884	-9.903	22.715
S	R	L	PR	10.485	-2.501	74.578
S	R	M	NF	-15.431	0.692	20.369
S	R	M	ON	11.937	1.436	22.441
S	R	M	PO	-28.526	6.071	16.136
S	R	M	PR	-40.863	0.244	17.028
S	SL	F	NF	-41.256	-3.388	18.43
S	SL	F	ON	-57.451	2.982	18.413
S	SL	F	PR	-40.682	1.933	18.945
S	SL	M	NF	16.541	-0.28	14.595
S	SL	M	ON	7.507	5.238	17.165
S	SL	M	PR	1.869	1.776	14.595
S	SR	L	NF	10.016	-6.995	89.702
S	SR	L	ON	-19.873	-4.938	95.971
S	SR	L	PO	4.141	-7.544	58.906
S	SR	M	NF	8.576	-8.364	35.548
S	SR	M	ON	-8.157	-1.914	29.877
S	SR	M	PO	3.685	-5.951	14.914

3.4 Syllabic segmentation (revised for §6.3.3.4)

Sentence type	Demarcation	Tone	Focus	Slope	Height	Strength
Q	M	M	NF	-1.544	-1.759	30.35
Q	M	M	ON	0.081	-3.3	36.244
Q	M	M	PO	0.719	-3.505	21.128
Q	M	M	PR	-6.908	-1.868	46.531
Q	M	M	PU	14.962	-3.066	19.058

S	M	M	NF	-6.821	1.324	35.002
S	M	M	ON	-6.083	-1.505	43.732
S	M	M	PO	-9.942	-5.214	37.595
S	M	M	PR	-5.269	-2	42.859
S	M	M	PU	-6.872	-10.308	24.461

Appendix 4: On-focus vs. neutral focus word duration



Appendix 5: F0 contours – effect of pre-low raising

5.1 Description of data

In Appendice 5.2, 5.3, 5.4, and 5.5, averaged F0 contours of 33 target words spoken at normal and slow speed are presented. The name of a given F0 curve shown on individual panels below consists of four elements, for example n21n:

n	Speech rate (n = normal, s = slow)
2	Word length (morae)
1	Accent condition (0-4)
n	Syllable structure (n = CVn, l = CVV, s = CV)

The gloss to the target words is as follows:

1-μ	10s	値	ne
	11s	根	**ne**
2-mora	20s	真似	mane
	20l	舞い	mai
	21s	メモ	**me**mo
	21l	May	**me**i
	21n	面	**me**n
	22s	旨	mu**ne**
3-mora	30s	実物	mimono
	30la	根芋	neimo
	30lb	未明	mimei
	30na	木綿	momen
	31s	女波	**me**nami
	31la	迷雰	**me**imu
	31lb	二名	**ni**mei
	31na	任务	**ni**nmu
	32s	斜め	na**na**me
	32lb	眩晕	me**ma**i
	32nb	二万	ni**ma**n
	33s	見もの	mimo**no**
	33la	縫い目	nui**me**
4-mora	40s	物真似	monomane
	40la	命名	meimei
	40na	年々	nennen
	41la	ムーミン	**mu**umin
	41na	何年	**na**nnen
	42s	皆々	mi**na**mina
	43s	生生	nama**na**ma
	43la	铭々	mei**me**i
	43na	面々	men**me**n
	44s	あのまま	anoma**ma**
	44lb	二枚目	nimai**me**
	44nb	二年目	ninen**me**

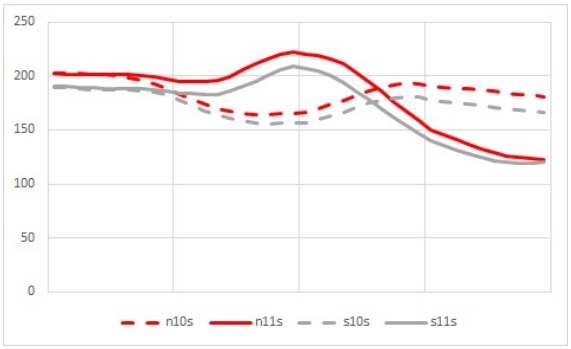
---Unaccented, normal speed

—Accented, normal speed

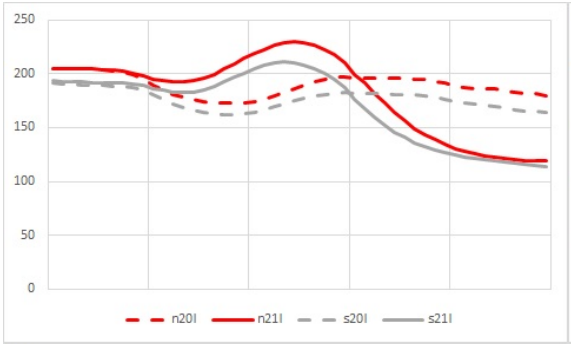
---Unaccented, slow speed

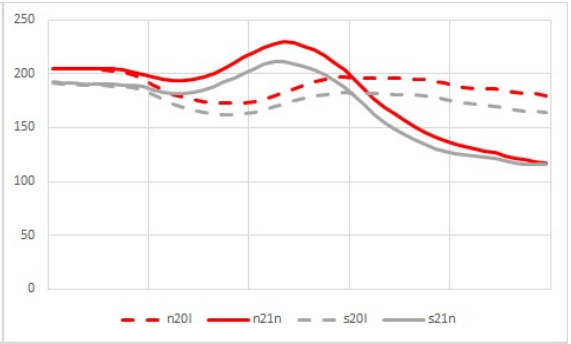
—Accented, slow speed

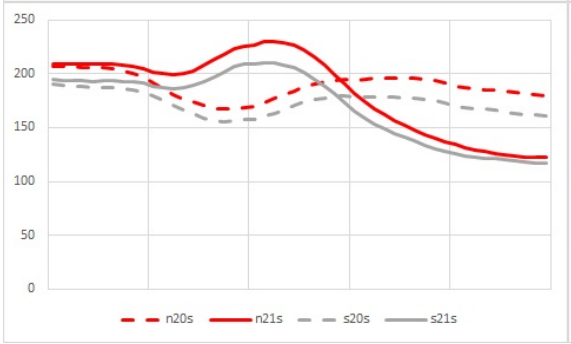
5.2 One-mora words

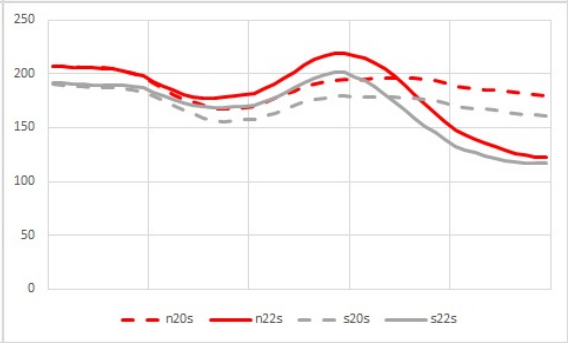


5.3 two-mora words

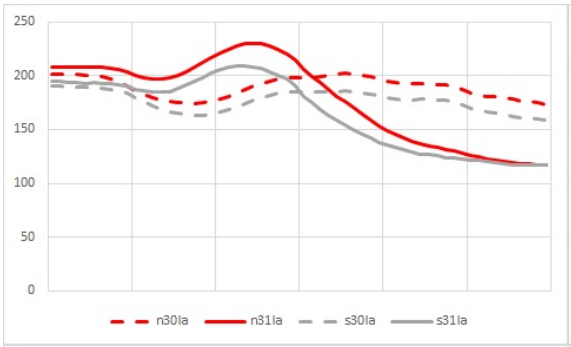


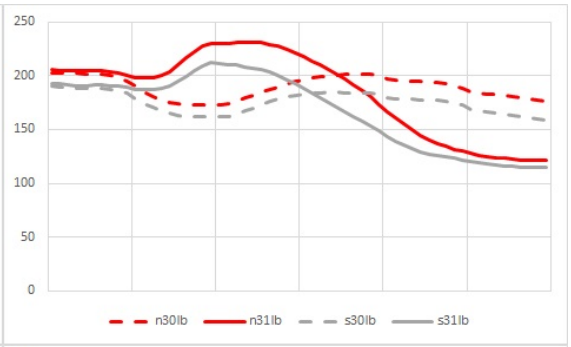


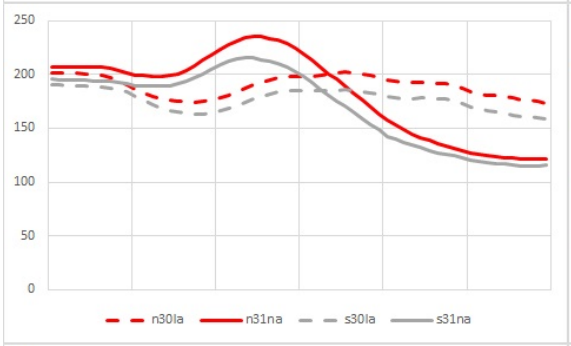


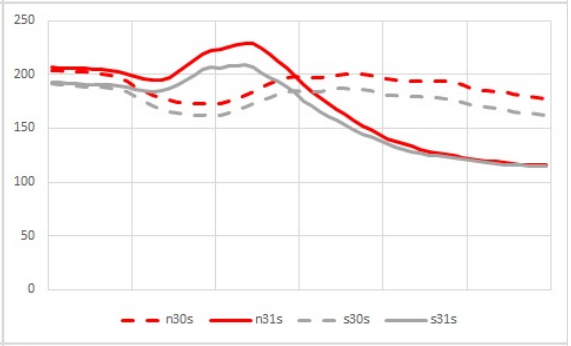


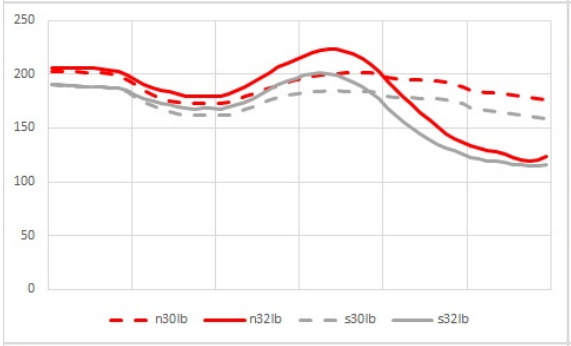
5.4 Three-mora words

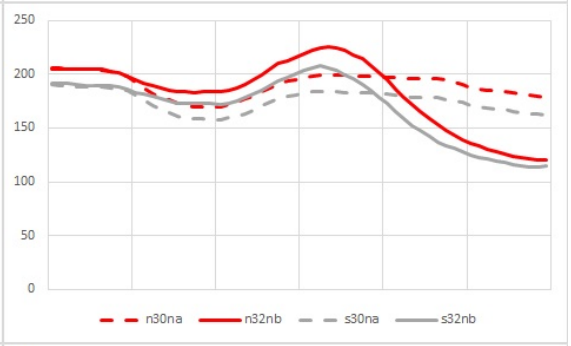


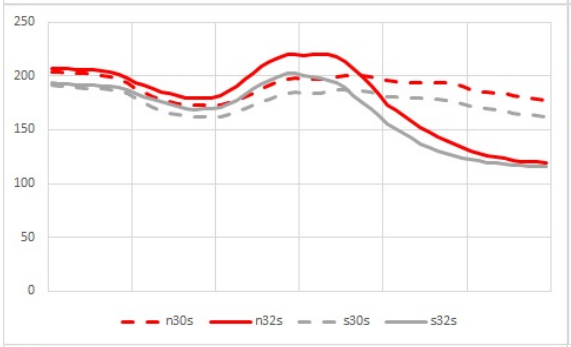


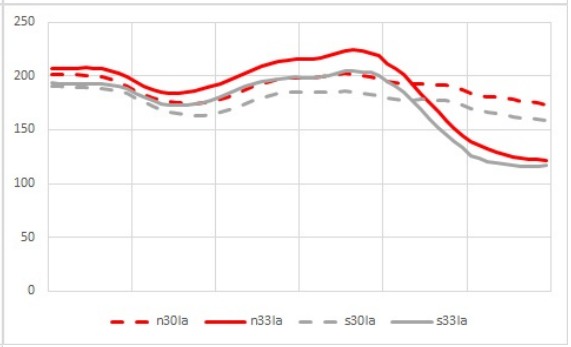


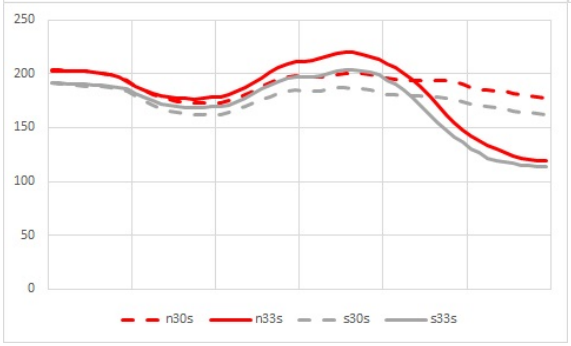




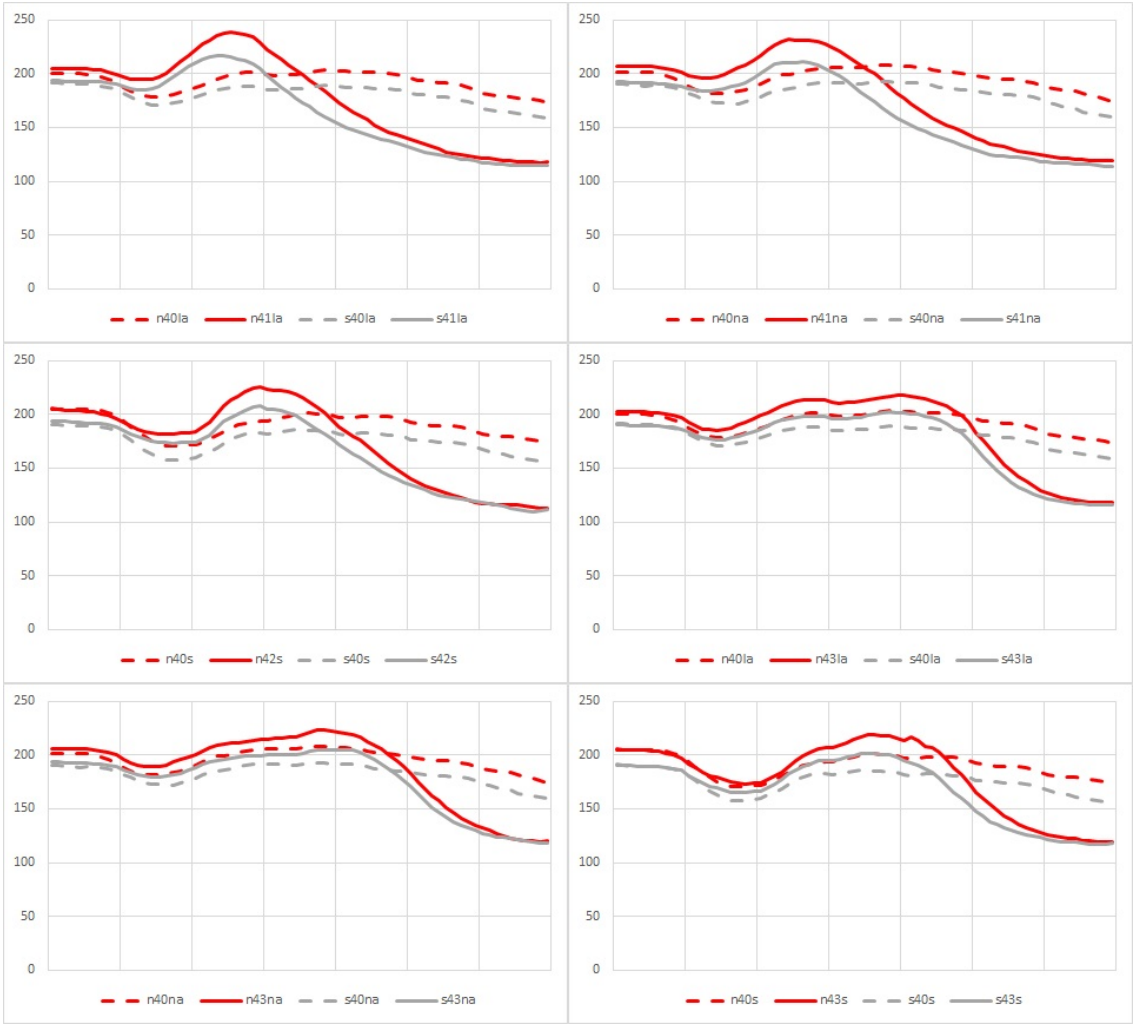


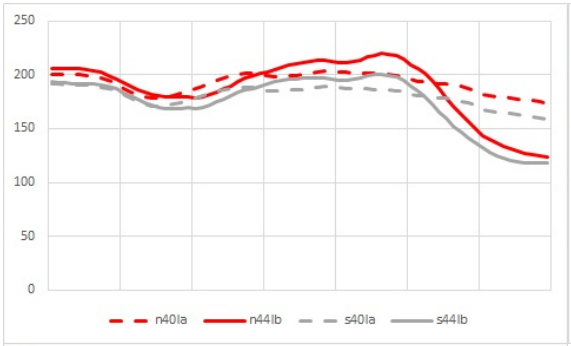


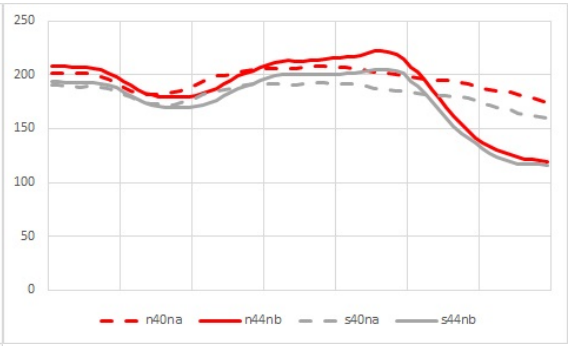


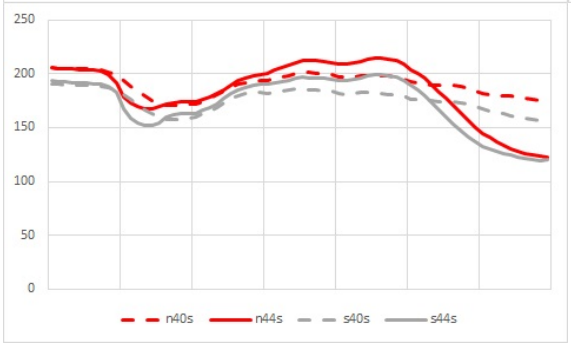


5.5 Four-mora words




250
200
150
100
50
0
n40la
n44lb
s40la
s44lb


250
200
150
100
50
0
n40na
n44nb
s40na
s44nb


250
200
150
100
50
0
n40s
n44s
s40s
s44s