UNIVERSITY OF CALIFORNIA SAN DIEGO

Phonetics of Period Doubling

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Linguistics and Cognitive Science

by

Yaqian Huang

Committee in charge

Professor Marc Garellek, Chair
Professor Sarah Creel
Professor Jody Kreiman
Professor Sharon Rose
Professor William Styler

2023

Phonetics of Period Doubling

The dissertation of Yaqian Huang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my parents.

TABLE OF CONTENTS

Dissertation Approval Page iii

Dedication iv

List of Tables viii

List of Figures xi

Acknowledgement xviii

Vita xxi

Abstract of the Dissertation xxii

Chapter 1: Introduction 1

1.1 How period doubling influences pitch and tone production and perception 2

1.2 Definition of period doubling adopted in dissertation 4

1.3 Other definitions of related phenomena 10

1.4 Period doubling in typical speech 12

1.5 Other studies on period doubling or related phenomena 16

1.6 Specifics and structure of dissertation 19

Chapter 2: How is period doubling articulated? – An electroglottographic study of Mandarin 21

2.1 Introduction 21

2.2 Background 22

2.3 Methods 27

2.4 Results: Distribution and defining characteristics of PD 34

2.5 Results: EGG waveform analysis of PD, vocal fry, and modal voice 45
2.6 Results: EGG spectrum analysis of different types of period doubling 75
2.7 Discussion 87
2.8 Chapter summary 91

CHAPTER 3: Production of period doubling: acoustic characteristics and linguistic distribution 93
3.1 Introduction 93
3.2 Background 95
3.3 Methods 102
3.4 Results: Acoustic analysis of PD, vocal fry, and modal voice 107
3.5 Computational classification 127
3.6 Linguistic distributions 151
3.7 Discussion 163
3.8 Chapter summary 166

CHAPTER 4: Perception of period doubling: pitch, voice quality, and tone 168
4.1 Introduction 168
4.2 Experiment 1: Pitch perception during resynthesized period-doubled tones 175
4.3 Experiment 2: Pitch shadowing of resynthesized period-doubled tones 192
4.4 Discussion 210
4.5 Chapter summary 216

CHAPTER 5: General discussion 218
5.1 What are the articulatory aspects of period doubling found in typical, non-disordered, speech? 219
5.2 What are the acoustic characteristics of period doubling that distinguish it from other voicing types? 222

5.3 Where does period doubling occur linguistically? Specifically, does it occur in specific tonal and phrasal environments? 225
5.4 How do listeners perceive pitch, voice quality, and tone during period doubling? . 229
5.5 Implications for voice and linguistic theories and beyond 232
5.6 Future directions 235
5.7 Concluding remarks 237

APPENDIX 239

BIBLIOGRAPHY 246

LIST OF TABLES

TABLE 1.1. Descriptions of period doubling in literature. 8

TABLE 2.1. The distribution of period-doubled tokens across speakers. 34

TABLE 2.2. Mean frequencies of the fundamental cycles and two glottal sub-cycles in period doubling, the frequency ratio RT, and the mean (SD) ratio between the first glottal period and the fundamental period in each subject. 36

TABLE 2.3. Mean amplitudes of the two glottal cycles and their ratio RA of each subject. 43

TABLE 2.4. Mean (SD) of frequency and amplitude ratio pooled by speakers. 'AM' stands for amplitude-modulated and 'FM' stands for amplitude and frequency-modulated period-doubled tokens. 44

TABLE 2.5. Summary of mean (SD) of different CQ measures during period doubling calculated based on four algorithms within each subject. 47

TABLE 2.6. Mean (SD) of CQ measures in PD, vocal fry, and modal voice. 54

TABLE 2.7. Mean (SD) SQ in PD, vocal fry, and modal voice. 65

TABLE 2.8. Mean (SD) PIC in PD, vocal fry, and modal voice. 70

TABLE 3.1. Acoustic and articulatory measures extracted using VoiceSauce and EGG-Works. 97

TABLE 3.2. Mean (SD) H1*–H2* in PD, vocal fry, and modal voice by gender. . . . 109

TABLE 3.3. Mean (SD) H1*–A1* in PD, vocal fry, and modal voice by gender. . . . 110

TABLE 3.4. Mean (SD) H1*–A2* in PD, vocal fry, and modal voice by gender. . . . 111

TABLE 3.5. Mean (SD) H1*–A3* in PD, vocal fry, and modal voice by gender. . . . 112

TABLE 3.6. Mean (SD) H2*–H4* in PD, vocal fry, and modal voice by gender. . . . 114

TABLE 3.7. Mean (SD) H4*–H2K* in PD, vocal fry, and modal voice by gender. . . 115

TABLE 3.8. Mean (SD) H2K*–H5K* in PD, vocal fry, and modal voice by gender. . 116
TABLE 3.9. Mean (SD) H1* in PD, vocal fry, and modal voice by gender. 117
TABLE 3.10. Mean (SD) H2* in PD, vocal fry, and modal voice by gender. 118
TABLE 3.11. Correlation across CQ measures and H1*–H2*. 119
TABLE 3.12. Mean (SD) HNR ($<500Hz$) in PD, vocal fry, and modal voice by gender. 120
TABLE 3.13. Mean (SD) CPP in PD, vocal fry, and modal voice by gender. 121
TABLE 3.14. Mean (SD) SHR in PD, vocal fry, and modal voice by gender. 123
TABLE 3.15. Mean (SD) SoE in PD, vocal fry, and modal voice by gender. 124
TABLE 3.16. Mean (SD) Energy in PD, vocal fry, and modal voice by gender. 125
TABLE 3.17. Summary of acoustic measures across three voice types by gender. . . . 127
TABLE 3.18. Distributions of PD, vocal fry, and modal voice in training and test sets. 129
TABLE 3.19. Distributions of PD, vocal fry, and modal voice in women and men in the acoustic dataset. 132
TABLE 3.20. Distributions of PD, vocal fry, and modal voice in women and men in the acoustic and articulatory dataset. 135
TABLE 3.21. Confusion matrix of the predicted and the actual voice types in the test set of acoustic features using logistic regression with Lasso regularization. 139
TABLE 3.22. Non-zero coefficients of different acoustic features predicting PD, vocal fry, and modal voice using Lasso regularized logistic regression. 141
TABLE 3.23. Confusion matrix of the predicted and the actual voice types in the test set of acoustic and articulatory features using logistic regression with Lasso regularization. 143
TABLE 3.24. Non-zero coefficients of different acoustic and articulatory features predicting PD, vocal fry, and modal voice using Lasso regularized logistic regression. 144
TABLE 3.25. Confusion matrix of the predicted and the actual voice types in the test set of acoustic features using random forest. 145

TABLE 3.26. Confusion matrix of the predicted and the actual voice types in the test set of acoustic and articulatory features using random forest. 146
TABLE 3.27. Confusion matrixes of the predicted and the actual voice types in the test set of acoustic features (a) and combined features (b) using radial SVM. 150
TABLE 3.28. Summary of overall accuracy, macro average precision and recall scores using logistic regression, random forest, and radial SVM. 150
TABLE 3.29. Coded prosodic positions in the distribution analysis. 152
TABLE 3.30. Classification of vowels used in the distribution analysis. 158

TABLE 4.1. Amplitude and frequency ratios used for modulation based on empirical data. 178
TABLE 4.2. Gaussian distributions (mean, SD) used to generate pitch values of training tokens in f0 conditions of 200Hz and 300Hz. 180
TABLE 4.3. Perceived pitch in ten least ambiguous (top) and ten most ambiguous (bottom) period-doubled tokens based on standard deviation of response. 189
TABLE 4.4. Significant fixed effects in predicting mean, max, and min imitated f0. . 197
TABLE 4.5. Significant fixed effects in predicting mean imitated f0 imitated in subsets of modulation type. 198
TABLE 4.6. Imitated f0 (semitone) in ten least ambiguous (top) and ten most ambiguous (bottom) period-doubled tokens based on standard deviation of f0. 200
TABLE 4.7. Significant fixed effects in predicting imitated voice quality correlates. . 201
TABLE 4.8. Significance of independent and interactional predictors involving perceptual responses. 205

TABLE 3.A1. Word list used for elicitation of the production study in Huang (ur). . . 239

LIST OF FIGURES

FIGURE 1.1. Theoretical framework relating voice, pitch, and tone. Second row is specifically for period doubling, the voice quality focused on in this dissertation. 3

FIGURE 1.2. Types of period doubling (a-c) and vocal fry (d). 6

FIGURE 2.1. EGG waveform (a) and spectrum (b) of amplitude-modulated period doubling. 32

FIGURE 2.2. EGG waveform (a) and spectrum (b) of amplitude- and frequency-modulated period doubling. 33

FIGURE 2.3. Distribution of frequency ratio RT by speaker. 37

FIGURE 2.4. Scatter plot of the frequency ratio RT as a function of the fundamental frequency. 38

FIGURE 2.5. EGG waveform and spectrum of vocal fry, with mean CQ (hybrid) = 0.77. 40

FIGURE 2.6. Comparison of the f0 in vocal fry, modal voice, and the three frequencies in period doubling. 41

FIGURE 2.7. Distribution of amplitude ratio RA by speaker. 44

FIGURE 2.8. Diagram of EGG (black) overlaid by dEGG (blue line) illustrating CQ measures: speed quotient (SQ), and peak increase in contact (PIC), based on EGGWorks (Tehrani, 2009) 46

FIGURE 2.9. Pearson correlation coefficients among CQ measures during period doubling. 48

FIGURE 2.10. Distributions of Contact Quotient (derivative) in period doubling, vocal fry, and modal voice. 49

FIGURE 2.11. Distributions of Contact Quotient (threshold) in period doubling, vocal fry, and modal voice. 50
FIGURE 2.12. Distributions of Contact Quotient (hybrid) in period doubling, vocal fry, and modal voice. 51
FIGURE 2.13. Distributions of Contact Quotient (threshold) in period doubling, vocal fry, and modal voice, faceted by gender. 52
FIGURE 2.14. Distributions of Contact Quotient (Henry Tehrani method) in period doubling, vocal fry, and modal voice. 53
FIGURE 2.15. Alternation seen in Contact Quotient (derivative) during tokens of period doubling. 56
FIGURE 2.16. Alternation seen in Contact Quotient (threshold) during tokens of period doubling 57
FIGURE 2.17. Alternation seen in Contact Quotient (hybrid) during tokens of period doubling. 58
FIGURE 2.18. Contact Quotient (derivative) of each cycle during tokens of modal voice. 59
FIGURE 2.19. Contact Quotient (derivative) of each cycle during tokens of vocal fry. 60
FIGURE 2.20. Distributions of median coefficients of the FFT of Contact Quotient (derivative, threshold, hybrid, and HT) in modal voice, vocal fry, and period doubling. 61
FIGURE 2.21. Density distributions of CQ measures of every other pulse during period doubling, vocal fry, and modal voice. 63
FIGURE 2.22. Distributions of Speed Quotient (log-transformed) values in period doubling, vocal fry, and modal voice. 65
FIGURE 2.23. Alternation seen in Speed Quotient of each cycle during tokens of period doubling. 66
FIGURE 2.24. Speed Quotient of each cycle during tokens of modal voice. 67

FIGURE 2.25. Speed Quotient of each cycle during tokens of vocal fry. 67
FIGURE 2.26. Distributions of median FFT coefficients of Speed Quotient in period doubling, vocal fry, and modal voice. 68
FIGURE 2.27. Density distributions of Speed Quotient (log-transformed) of every other pulse during period doubling, vocal fry, and modal voice. 69
FIGURE 2.28. Distributions of Peak Increase in Contact in period doubling, vocal fry, and modal voice. 70
FIGURE 2.29. Peak increase in contact of each cycle in sustained period doubling tokens. 71
FIGURE 2.30. Peak increase in contact of each cycle in modal tokens. 72
FIGURE 2.31. Peak increase in contact of each cycle in sustained vocal fry tokens. . 72
FIGURE 2.32. Distributions of median FFT coefficients of Peak Increase in Contact in modal voice, vocal try, and period doubling. 73
FIGURE 2.33. Density distributions of Peak Increase in Contact of every other pulse during period doubling, vocal fry, and modal voice. 74
FIGURE 2.34. EGG spectrum of amplitude-modulated period doubling. 76
FIGURE 2.35. Waveform and spectrum of amplitude modulated period doubling (from speaker F04). 77
FIGURE 2.36. EGG spectrum of amplitude and frequency-modulated period doubling. 78
FIGURE 2.37. Waveform and spectrum of amplitude and frequency modulated period doubling (from speaker F37). 79
FIGURE 2.38. Waveform and spectrum of amplitude modulated period doubling (from speaker F37). 80
FIGURE 2.39. Waveform and spectrum of amplitude and frequency modulated period doubling (from speaker F30). 81
FIGURE 2.40. Waveform and spectrum of amplitude modulated period doubling showing interharmonics (from speaker F04). 82

FIGURE 2.41. Waveform and spectrum of amplitude modulated period doubling showing interharmonics in higher frequencies (from speaker F30). 83
FIGURE 2.42. Waveform and spectrogram of amplitude modulated period doubling showing bifurcation of f0 (from speaker F30). 85
FIGURE 2.43. Waveform and spectrogram of amplitude and frequency modulated period doubling showing bifurcation of f0 (from speaker F30). 86
FIGURE 2.44. Waveforms and spectra of vocal fry (first row) and modal voice (second row) (from speaker F04). 87

FIGURE 3.1. EGG waveform (a) and audio waveform (b) of period doubling. 105
FIGURE 3.2. EGG waveform (a) and audio waveform (b) of vocal fry. 106
FIGURE 3.3. Distributions of fundamental frequency (f0) in period doubling, vocal fry, and modal voice, faceted by gender. 107
FIGURE 3.4. Distributions of H1*–H2* in period doubling, vocal fry, and modal voice, faceted by gender. 109
FIGURE 3.5. Distributions of H1*–A1* in period doubling, vocal fry, and modal voice, faceted by gender. 110
FIGURE 3.6. Distributions of H1*–A2* in period doubling, vocal fry, and modal voice, faceted by gender. 111
FIGURE 3.7. Distributions of H1*–A3* in period doubling, vocal fry, and modal voice, faceted by gender. 112
FIGURE 3.8. Distributions of H2*–H4* in period doubling, vocal fry, and modal voice, faceted by gender. 113
FIGURE 3.9. Distributions of H4*–H2K* in period doubling, vocal fry, and modal voice, faceted by gender. 114
FIGURE 3.10. Distributions of H2K*–H5K in period doubling, vocal fry, and modal voice, faceted by gender. 115

FIGURE 3.11. Distributions of H1* in period doubling, vocal fry, and modal voice, faceted by gender. 117
FIGURE 3.12. Distributions of H2* in period doubling, vocal fry, and modal voice, faceted by gender. 118
FIGURE 3.13. Scatter plot of CQ (hybrid) and H1*–H2* values varying by voice types. 120
FIGURE 3.14. Distributions of HNR ($<500Hz$) in period doubling, vocal fry, and modal voice, faceted by gender. 121
FIGURE 3.15. Distributions of CPP in period doubling, vocal fry, and modal voice, faceted by gender. 122
FIGURE 3.16. Distributions of SHR in period doubling, vocal fry, and modal voice, faceted by gender. 123
FIGURE 3.17. Distributions of SoE in period doubling, vocal fry, and modal voice, faceted by gender. 125
FIGURE 3.18. Distributions of root-mean-squared energy (log-transformed) in period doubling, vocal fry, and modal voice, faceted by gender. 126
FIGURE 3.19. t-SNE visualization of tokens of period doubling, vocal fry, and modal voice using all the acoustic parameters. 133
FIGURE 3.20. t-SNE visualization of tokens of period doubling, vocal fry, and modal voice using all the acoustic and articulatory parameters. 134
FIGURE 3.21. Pearson correlation coefficients among acoustic features. 137
FIGURE 3.22. Pearson correlation coefficients among acoustic and articulatory features. 138
FIGURE 3.23. Lasso selection process of lambda based on misclassification error of period doubling, vocal fry, and modal voice using the acoustic features. 139
FIGURE 3.24. Lasso selection process of lambda based on misclassification error of period doubling, vocal fry, and modal voice using the acoustic and articulatory features. 142

FIGURE 3.25. Top 15 important acoustic features in the training set of the random forest model. 146
FIGURE 3.26. Multidimensional scaling plot using acoustic features in the training set of the random forest model. 147
FIGURE 3.27. Top 15 important acoustic features in the training set of the random forest model. 148
FIGURE 3.28. Multidimensional scaling plot using acoustic and articulatory features in the training set of the random forest model. 149
FIGURE 3.29. Waveform and spectrogram of a sample stimuli carrier sentence: 'I teach you blackboard drawing how to say' with coded prosodic positions. 153
FIGURE 3.30. Bar plot of raw count of period doubling and vocal fry across tones in sentence-medial positions. 154
FIGURE 3.31. Bar plot (in percentage) of period doubling and vocal fry across different utterance positions: initial, medial, and final by gender. 155
FIGURE 3.32. Bar plot (in percentage) of period doubling and vocal fry across different phrasal positions throughout the sentence by gender. 156
FIGURE 3.33. Bar plot (in percentage) of monophthongs by vowel height (a, b), diphthongs and triphthongs (c) across tokens of period doubling and vocal fry 159
FIGURE 3.34. The distribution of the distance between $F1$ and $f0$ (Hz) (a) and between $F1$ and $2 \times f0$ (Hz) (b) in [a, i, u] across tokens of period doubling and vocal fry. 161
FIGURE 3.35. The distribution of H1*–A1* (a) and H2*–A1* (b) in [a, i, u] across tokens of period doubling and vocal fry. 162
FIGURE 4.1. Experimental procedure from left to right: familiarization, training, testing phases. 176

FIGURE 4.2. Resynthesized period-doubled pulses of 200 Hz: original unmodified token (a), amplitude-modulated at 2.4 (b), frequency-modulated at 2.4 (c), and amplitude- (ratio: 2.4) plus frequency-modulated (2.4) (d). . . 178
FIGURE 4.3. Proportion of 'down' responses as a function of modulation types. . . 183
FIGURE 4.4. Proportion of 'down' responses as a function of the varying degrees of amplitude modulation in tokens of pure amplitude modulation (a) and combined modulation (b). 185
FIGURE 4.5. Proportion of 'down' responses as a function of the varying degrees of frequency modulation in tokens of pure frequency modulation (a) and combined modulation (b). 186
FIGURE 4.6. Imitated mean f0 (semitone) varied by modulation types. 197
FIGURE 4.7. Imitated mean f0 (semitone) varied by amplitude (a) or frequency (b) modulation degrees. 199
FIGURE 4.8. Imitated f0 across different f0 conditions in Experiment 2 as a function of perceptual responses in Experiment 1. 204
FIGURE 4.9. H1*–H2* (a), HNR (b), SHR (c), and SoE (d) in imitation across different f0 conditions in Experiment 2 vary as a function of perceptual responses and modulation type. 206

FIGURE 3.A1. Shrunk coefficients of acoustic features in classifying period doubling (a), vocal fry (b), and modal voice (c) using different lambda values during the Lasso selection process. 242
FIGURE 3.A2. Shrunk coefficients of acoustic and articulatory features in classifying period doubling (a), vocal fry (b), and modal voice (c) using different lambda values during the Lasso selection process. 244

ACKNOWLEDGEMENTS

This dissertation would not be made possible without my advisor Marc Garellek. I grew interested in voice quality because it was a 'minor' area when I applied for graduate programs in linguistics. I learned about 'pitch doubling' when segmenting Mandarin words using PRAAT to avoid pitch jumps and halves in my master's. I almost would miss being here if I did not email or receive from Marc years ago from the other side of the country, but his enthusiasm and entrusted personality helped me see my potential and made me believe that this would be the right place for me and it surely has been.

Marc is the most organized and disciplined person I have ever met in my professional career. He is always patient, insightful, attending to details, and passionate and thorough on topics he knows well. He cares not only about academic development but personal well-being. He is amiable, and is always there whenever I needed his advice, guidance, or support. Even if it were last-minute request, he always comes to the rescue. I have learned from Marc tremendously throughout my stay at UCSD. I see myself grow as a scholar with his mentorship: I learned from how to come up with a solid research idea, how to design experiments, how to write abstracts, reviews, and recommendation letters, to how to develop my own research program as an independent researcher, and importantly, how to be a mentor, and above and beyond.

I would like to thank my committee members. Will Styler, for he is always so open-minded and welcomes whatever ideas I have. He prioritizes the thinking process over the results – even if there lacks an apparent conclusion towards the end of our discussion, he would not seem bothered but encouraging and inspiring. I learned by exhausting all the possibilities and always left with confidence about the choices I had to make. Sharon Rose, for her humor and adorable personality. I enjoy a lot getting along with her and listening to her stories and experiences besides the busy academic life. She is quick-witted, wise, and always sees through the surface; at the same time, she is charismatic and gives sage advice. Sarah Creel, for she always asks the right questions and never holds back in giving out critiques and sharp views, especially from a cognitive science

perspective. Jody Kreiman, for her kindness and patience in walking me through my preliminary ideas on voice research and reminding me of the big picture behind the nitty-gritty details.

This dissertation is funded by the National Science Foundation (NSF) (Linguistics Program - Doctoral Dissertation Research Improvement Award, BCS-2141433, PI Marc Garellek, co-PI Yaqian Huang). I am grateful to my wonderful undergraduate research assistants: Shogo Nishimura, Hira Rizvi, Stephanie Luong, Theodore Jones, Xutong Zhang, Xinyi Liu, and Shihong Weng. Without them, this work would not be accomplished in its current shape. I would like to thank the audiences at *LSA 2019*, *ICPhS 2019*, *Speech Prosody 2022*, and *ASA 2022* for their helpful feedback and comments on different parts of this work at different stages.

My stay at UCSD, as a member of the Phonetics Lab, the PhonCompany, the Linguistics Department, and the interdisciplinary community across Psychology, Cognitive Science, Music, Neuroscience, and Computer Science has been a delight. One of my life goals is to be educated. I am glad that I have engaged with such vibrant research and learning communities and am lucky to have taken courses for my broad interests across multiple disciplines. In particular, I wish to thank my teachers and mentors who I have taken classes with and who have inspired and made my professional life enjoyable. Thank you, Eric Bakovic, Ed Vul, Eric Halgren, Eran Mukamel, Florian Meyer, Saharnaz Baghdadch, Emily Clem, Robert Kluender, Leon Bergen, Tamara Smyth, and Gary Cottrell.

I also would like to thank my teachers in my singing, drawing, dance, and Japanese classes who bring me back to simple pleasant moments in life and offer small meaningful breaks from sometimes tedious and laborious work.

I must thank the friends and colleagues I have got to know during graduate school including my MA and PhD programs. I am picky about friends, but I am glad that I have made some that will likely maintain for a lifetime. Big thanks to Lexi Geibler, Alba Fano, Xinglong Zhang, Xinqi Guo, Corey Zhou, Qi Cheng, and my cohorts. I also want to thank my friends back home who I have known since high school: Liying Ye, Lin Yang, Siyu Chen, and Yuling Lin.

My most sincere gratitude extends to my parents, Andi Yang and Wei Huang. Without them, my graduate and professional careers will not be possible. They are always my backup, my unconditional love and support.

VITA

1993	Born, Hengyang, Hunan.
2015	B.A. English and Finance, University of International Business and Economics
2017	M.A. Linguistics and Cognitive Science, University of Delaware
2017 - 2022	Teaching Assistant, Linguistics Department, University of California San Diego
2021	Research Assistant, Linguistics Department, University of California San Diego
2022	Ph.D. Linguistics and Cognitive Science, University of California San Diego

PUBLICATIONS

Huang, Y. (Under review). F0 and voice quality of coarticulated Mandarin tones. *Language and Speech*.

Huang, Y. (To appear). Phonetic and phonological analysis of the Rere vowel height system. *Proc. of Annual Conference on African Linguistics 51.52*. Florida.

Huang, Y. (2022). Articulatory properties of period-doubled voice in Mandarin. *Proceedings of Speech Prosody 2022*, pages 545-549. doi:10.21437/SpeechProsody.2022-111

Garellek, M., Chai, Y., Huang, Y., and Van Doren, M. (2021). Voicing of glottal consonants and non-modal vowels. *Journal of the International Phonetic Association*, pages 1-28. doi: 10.1017/S0025100321000116

Huang, Y. (2020). Different attributes of creaky voice distinctly affect Mandarin tonal perception. *The Journal of Acoustical Society of America*, 147(3), 1441-1458. doi:10.1121/10.0000721

Lau, S. H.*; Huang, Y.*; Ferreira, Victor S.; Vul, Edward (2019). Perceptual features predict word frequency asymmetry across modalities. *Attention, Perception, and Psychophysics*, 81(4), 1076-1087. doi:10.3758/s13414-019-01682-y (*equally contributed)

Huang, Y. (2019). The role of creaky voice attributes in Mandarin tonal perception. *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 1465-1469.

Huang, Y. (2019). Revisiting non-idempotency in Tibetan vowel harmony. *San Diego Linguistic Papers*, 7:2-16.

Huang, Y., Athanasopoulou, A., and Vogel, I. (2018). The effect of focus on creaky phonation in Mandarin Chinese tones. *University of Pennsylvania Working Papers in Linguistics*, 24(1):12.

FIELDS OF STUDY

Major Field: Linguistics

Studies in Phonetics

Secondary Field: Cognitive Science

ABSTRACT OF THE DISSERTATION

Phonetics of Period Doubling

by

Yaqian Huang

Doctor of Philosophy in Linguistics and Cognitive Science

University of California San Diego, 2023

Professor Marc Garellek, Chair

The human voice is the most common 'carrier' of speech, but how does linguistic voice quality affect speech production and perception? Typical 'modal' voice possesses a single fundamental frequency (f0), identified as the voice's pitch. Period doubling, known as a commonly-occurring type of creaky voice, consists of alternating glottal pulses with different periods and/or amplitudes for which multiple fundamental frequencies (f0s) co-exist. Thus, the pitch during period doubling is often indeterminate, and so it is unclear whether linguistic tone is identifiable, and how linguistic tone is identified, in this voice. Although period doubling has been mostly

studied in voice disorders and singing styles, it frequently occurs in non-pathological voices, and its defining characteristics remain to be determined.

This dissertation contributes three studies to characterize the physical, distributional, and perceptual aspects of period-doubled voice. Simultaneous electroglottography (EGG) and audio recordings of a Mandarin read speech corpus were analyzed to capture properties of the articulation and acoustics of period doubling in Chapters 2 and 3; artificial language learning and shadowing experimentation were used to probe perception of period doubling in Chapter 4.

The EGG study in Chapter 2 finds that period doubling is articulated as two alternating pulses with distinct pitches as well as voice qualities. Specifically, I show that the glottal cycles in period doubling are not generally constricted, but instead oscillate between degrees of constriction shown by alternating contact quotient, pulse shape, and speed of vocal fold contact. This in addition to the alternating frequencies likely leads to the indeterminate pitch and quality percept in period doubling. The results also pose challenges to the existing taxonomy of creaky voice subtypes based on the established acoustic attributes.

The acoustic analysis in Chapter 3 finds that period doubling is characterized acoustically via lower spectral tilt due to a stronger second harmonic from the original f0 (the first harmonics is derived from subharmonics), which distinguishes period doubling from vocal fry (another creaky-like voice quality) and modal voice. The results of the prosodic distribution show that, in Mandarin, period doubling occurs most frequently at the ends of utterances whereas vocal fry occurs at a post-focal position. This suggests that period doubling reflects vocal instability at the beginning and end of phonation, whereas vocal fry may be marking a weak prosodic element.

The perception study in Chapter 4 finds that both Mandarin and English listeners hear a ‘low-tone’ during period doubling, which is driven by the strength of frequency modulation more than that of amplitude modulation. When frequency modulation is at extremes, pitch is heard unambiguously as a lower tone. When frequency modulation is weak, pitch is often heard as ambiguous – both high and low tones are possible. Further, listeners are able to imitate the period-doubled tones not only by adjusting f0, but by also modulating their voice quality. It is

predicted that period doubling is used to signal low tones and could interfere with perception of tone of a high pitch.

Together, this dissertation establishes period doubling not only as a phonetic category distinct from other voicing types such as modal voice and vocal fry, but also serves a distinct linguistic role based on its phonetic aspects and role in perception. The findings provide insight into speech production, perception, and processing, with implications for how period doubling can be synthesized and used to convey linguistic meaning.

Chapter 1

Introduction

The human voice is an essential and integral part of speech. The use of voice is foundational and versatile. It is most known for serving to identify and distinguish talkers (Kreiman and Sidtis, 2011), signal emotion and affect in extra-linguistic settings (Laver, 1980; Esling et al., 2019), and convey relationships of prestige and dominance in speech as a sociolinguistic variable (Pittam, 1987; Esling, 1978a,b; Gobl et al., 2003).

Besides the general use and paralinguistic functions of voice in talker identification, speech recognition, and social interaction, the quality of voice can be used as an important feature in phonemic or phonetic categories. For example, the voice can convey lexical meaning by means of contrastive phonation type (e.g., breathy, creaky vowels in Mazatec, Yi; Esposito and Khan, 2020), or via its association as a secondary sub-phonemic feature of another contrastive category, such as tone (e.g., allophonic creaky voice in Mandarin, White Hmong; Kuang, 2017; Esposito, 2012). It may also be used to mark utterance boundaries (Kreiman, 1982) and signal turn taking (Local et al., 1986). These uses of voice quality are considered to be 'linguistic', where the voice encodes linguistic categories.

The present dissertation aims to investigate a fundamental question regarding human language: how does voice quality affect speech production and perception? Previous studies by Kuang (2013) and Garellek and Esposito (2021), among others, have found that voice quality enriches the representation of the tonal space and expands the dimensions of contrastive tone categories. Mandarin Chinese, a tonal language with contrastive pitch and without contrastive phonation, has predictable creaky voice frequently associated with tones that contain a low-pitch target such as Tones 2, 3, and 4. To probe how voice quality, in particular different kinds of creaky

voice, are used in Mandarin tones, I investigate the commonly-occurring subtypes of creaky voice including period doubling and vocal fry that are widely attested (Yu, 2010).

Thus, I approach this fundamental question through the lens of an under-studied type of voice quality: period doubling. This dissertation focuses on how period doubling (compared to vocal fry), contributes to the production and perception of linguistic tone, using Mandarin Chinese as the target language. The outcomes of this dissertation will address the meaning of voice quality as used in tone categories, which provides direct answers to phonetic and linguistic questions of the role of sub-phonemic units in linguistic contrasts from a suprasegmental and prosodic perspective. A better description of period doubling furthers our understanding of the interaction between pitch and voice in linguistic tones, and enriches socio-indexical and interactional implications of aperiodic voices. Because the study of voice is interdisciplinary in nature, the findings will be relevant not only for theories of the voice in linguistics, but speech-language pathology, and singing.

1.1. How period doubling influences pitch and tone production and perception

Voice quality is integral to both pitch and tone production and perception. Changes in voice quality depend largely on the actions of the vocal folds, which also induce changes in pitch, the primary cue to tone. Typical modal voice possesses a single fundamental frequency (f0) which is the primary acoustic correlate of pitch of the voice. In some subtypes of creaky voice such as vocal fry, pitch is also determinant from a regular, though low, f0 according to the definition in Keating et al. 2015. However, period doubling, another common yet less-studied subtype of creaky voice, poses a challenge to the problem of pitch determination. In this voice, multiple periodicities co-exist; thus, multiple f0s derived from those periods can be identified. Given the presence of multiple f0s, the pitch during period doubling is often indeterminate, and so it is unclear whether tone is identifiable, and how tone is identified, in period doubling. This dissertation thus focuses on capturing defining characteristics of period doubling and probing its role in pitch and tone production and perception. The theoretical framework of how voice quality interacts with pitch

to influence tone perception in general and specifically, period doubling, is schematized in Figure 1.1.

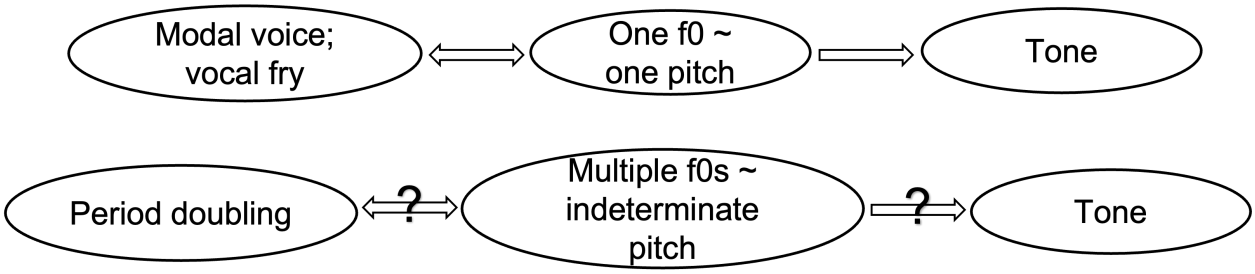


Figure 1.1: Theoretical framework relating voice, pitch, and tone. Second row is specifically for period doubling, the voice quality focused on in this dissertation.

Period doubling is a subtype of creaky voice (Keating et al., 2015) and commonly observed in $\sim 25\%$ of normal speakers' utterances (Klatt and Klatt, 1990). However, it is mostly studied and better known from pathological voice as 'diplophonic voice' (Kelman, 1981; Dejonckere and Lebacq, 1983; Gerratt et al., 1988) and certain styles of singing (such as throat singing styles in Mongolian: Lindestad et al., 2001; Sardinian: Bailly et al., 2010; and Japanese: Yoshinaga and Kong, 2012). Thus, in normal speech period doubling has been understudied and calls for more attention. A thorough analysis of the defining properties from perspectives of articulation, acoustics, and perception that may capture period doubling is needed.

A detailed description of period doubling also contributes to the taxonomy of creaky voice by expanding the finer-grained phonetic categories. Creaky voice, an umbrella term, has been loosely used to refer to different types of irregular voicing, often with vague or ambiguous definitions. For example, 'vocal fry' and 'creak', and 'creaky voice' and 'aperiodic voice', are often used interchangeably, and more accurate descriptions of creaky voice subtypes are lacking or at the preliminary stage. Nonetheless, the most typical characteristics associated with creaky voice are low f0, irregular f0, and constricted glottis (Garellek, 2019).

In summary, probing the defining features of period doubling in typical speech not only explores its use in language, but provides finer phonetic details to the production and perception of creaky voice. This furthers our understanding of f0 and pitch, and crucially, the interaction

of pitch and voice, which contributes to the use of phonation type and tone across languages. Moreover, this study will lay a foundation for further comparisons of period doubling between typical and pathological voice qualities, and similar or related voicing patterns in speech and singing.

Among the many unanswered questions, here I ask four overarching research questions that will lay the foundation for further future research on (creaky) voice, pitch, tone, and prosody:

1) What are the articulatory aspects of period doubling found in typical, non-disordered, speech?

2) What are the acoustic characteristics of period doubling that distinguish it from other voicing types?

3) Where does period doubling occur linguistically? Specifically, does it occur in specific tonal and phrasal environments?

4) How do listeners perceive pitch, voice quality, and tone during period doubling?

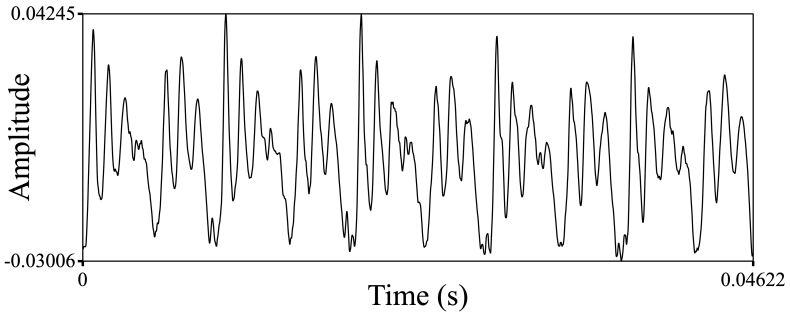
In the following sections, I will start by discussing the definition of period doubling adopted in the dissertation, and related phonation types with their descriptions in Sections 1.2-1.3. Then, I will summarize past studies of period doubling in typical speech (Section 1.4) and other cases, including disordered speech and singing voice (Section 1.5). Finally, Section 1.6 presents the structure of the dissertation and an overview of the chapters that contribute to determining how period doubling is produced and perceived.

1.2. Definition of period doubling adopted in dissertation

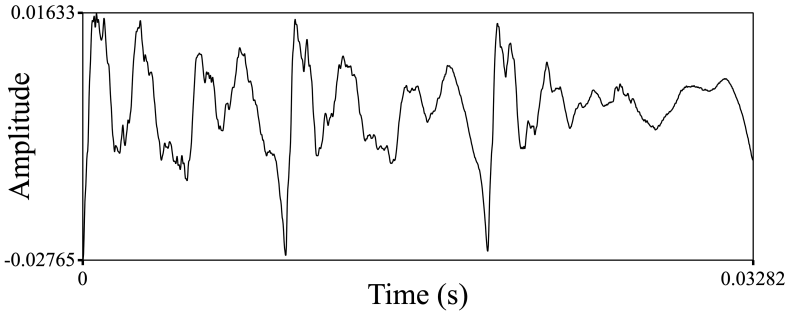
In the framework of subtypes of creaky voice by Keating et al. (2015), 'multiply-pulsed voice' is used to cover a kind of f0 irregularity with alternating pulses and multiple (i.e., two or more) simultaneous periodicities. In this dissertation, I focus on multiply-pulsed voice with *only two* recurring periods or amplitudes, hence period *doubling*. Specifically, period doubling is referred to

the type of voice that consists of alternating glottal pulses with two different periods and/or amplitudes. With the prominent alternation, multiple fundamental periods, or fundamental frequencies (f0s) co-exist. The f0s are recognizable in the acoustic waveform either as different amplitudes or frequencies of the alternating pulses. This description of an alternation has appeared in one of the earlier studies by Titze (1994) as 'bifurcation', which in theory also includes period tripling and period quadrupling, and also work by Gerratt and Kreiman (2001) as one of the 'supraperiodic' (neither periodic nor aperiodic) phonation types.

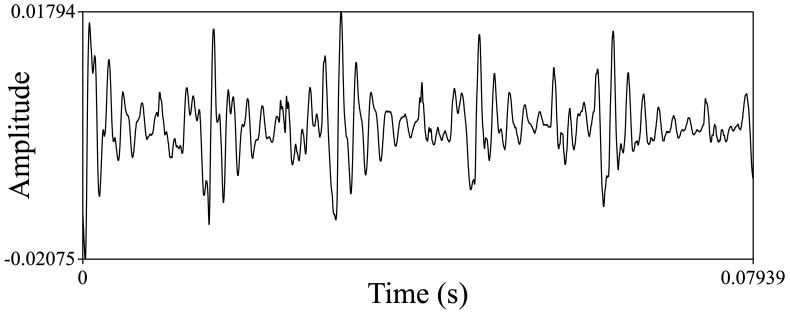
As shown in Figure 1.2, (a-c) are all instances of period doubling, of different types. Type (1.2a) is characterized by amplitude modulation with prominent alternation in pulse amplitudes; Type (1.2b) is characterized by amplitude and frequency modulation with prominent alternation in both pulse amplitudes and periods. Type (1.2c) also contains both amplitude and frequency modulation, though the first pulse is much stronger and longer than the second pulse that would be nearly omitted. Figure (1.2d) is vocal fry.



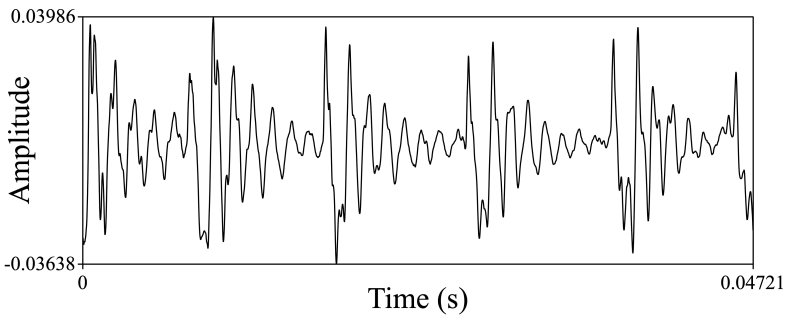
(a) Amplitude-modulated PD



(b) Amplitude- & frequency-modulated PD



(c) Amplitude- & frequency-modulated PD ("Multiply-pulsed vocal fry") (one strong pulse followed by a weaker one, with long periods suggestive of fry)



(d) Vocal fry (no alternation in frequency or amplitude)

Figure 1.2: Types of period doubling (a-c) and vocal fry (d).

One may notice that Type (1.2c) of period doubling has a similar look to 'vocal fry' in Figure (1.2d). In Type (1.2c), pulses during period doubling tend to have a long closure followed by the two successive opening cycles; thus, it has often been conflated with 'vocal fry', termed as 'doubled-pulsed vocal/glottal fry' in previous literature including Moore and Von Leden 1958, Wendahl 1962, or an intermediate pattern between classic vocal fry and period doubling, 'multiply-pulsed vocal fry' in Gerratt and Kreiman 2001.

However, period-doubled pulses in Type (1.2c) have the double length of duration and the modulated weaker pulses alternating with the stronger pulses, contrasted with Figure (1.2d). Here, despite clarifying the conflation, I do not assert that as both subtypes of creaky voice, period doubling and vocal fry, are mutually exclusive; in fact, period doubling can resemble vocal fry in its pulse shape, and they are expected to co-occur in environments that are in favor of creak, as is the case with 'multiply-pulsed vocal fry'.

Typically, the two frequencies in period doubling differ by one octave, which has been noted and termed as 'octave voice breaks' as early as in Wendahl 1962 and Bowler 1964, and perceptually are characterized by noticeable roughness (Keating et al., 2015). Here, period-doubled voice always refers to the presence of alternation in glottal pulses to create a 'doubled original period' with a pair of pulses; 'vocal fry' is defined as constricted glottis with a low and often regular pitch without any alternation. These terms largely conforming to Keating et al.'s (2015) taxonomy, though limited to double pulsing.

Another term, 'diplophonia', has been used in the voice literature to describe period doubling and related voice qualities such as 'bicyclicity' or 'biphasic' (Kreiman et al., 1993; Timcke et al., 1959; Aichinger, 2014). This term has appeared in the literature on pathological voice when it is used to describe a voice disorder, as detailed in Section 1.5.1, which has a slightly different definition from period doubling. However, in normal voice (referring to 'with no pathology', rather than 'modal' or otherwise) and machine classification frameworks, researchers have made direct reference to diplophonia using similar if not equivalent definitions. For example, in the absence of pathology, case studies of volitional diplophonia have been documented (Ward et al.,

1969; Schreibweiss-Merin and Terrio, 1986). In these studies, diplophonia is described as two cooccurring tones that are similar to one another in quality and loudness, but which are perceived as two separate pitches. The classification framework of creak by Batliner et al. (1994) defined diplophonia as regular variation between high and low amplitude. Likewise, Hedelin and Huber (1990) referred to 'diplophonic phonation' as an alternation between strong and weak glottal excitations when identifying aperiodic glottal excitation in Swedish texts. Redi and Shattuck-Hufnagel (2001) also examined diplophonia at English phrase boundaries, defined as alternation in the shape, amplitude, or duration of cycles. Martin (2012) implemented an automatic creak detection algorithm where diplophonic was defined by 'paired pulsing' such that the alternate periods are more similar than consecutive periods, and subharmonics with a very weak intensity in the first component. See Table 1.1 for a summary of past descriptions of period doubling in temporal, spectral, and perceptual domains. Most of the literature converges on the description that period doubling entails alternation of periods/frequencies and amplitudes, presence of subharmonics, and roughness or a 'bitonal' percept.

Table 1.1: Descriptions of period doubling in literature.

Temporal	Spectral	Perceptual	References
Alternating longer and shorter pulses/frequency and/or higher and lower amplitudes; pitch halving in the f0 track	Appearance of subharmonics	Bitonal	Yu (2010), Yu (2019)
Alternating longer and shorter pulses; two simultaneous periodicities	Two sets of subharmonics; low H1–H2	An indeterminate pitch, plus roughness	Keating et al. (2015)
Alternation of large and small amplitudes/periods	Appearance of subharmonics of f0	-	Herzel, Berry, Titze, & Saleh, (1994)
Alternating long and short periods from the original period	Subharmonic frequency at f0/2	-	Švec, Schutte, & Miller (1996)

Table 1.1: Descriptions of period doubling in literature. (cont.)

Temporal	Spectral	Perceptual	References
Alternation in shape, amplitude, or duration of successive pulses	-	Rough voice quality accompanied by the percept of a distinct pitch approximately an octave below the pitch perceived for a nearby modal region	Redi & Shattuck-Hufnagel (2001)
-	Additional peaks in the middle of the harmonics of the original pitch; sudden jumps to sub-harmonic regimes; low-frequency modulation	-	Herzel & Reuter (1996)
-	A subharmonic at one half of the fundamental frequency	-	Mende, Herzel, & Wermke (1990)
A period-two up-down pattern of an arbitrary cyclic parameter	-	-	Aichinger et al. (2017)
Two simultaneous frequencies	-	Two similar tones heard as two separate pitches	Ward, Sanders, Goldman, & Moore (1969), Schreibweiss-Merin & Terrio (1986)
Regular variation between high and low amplitude	-	-	Batliner, Burger, Johne, & Kießling (1994)
Alternation between strong and weak glottal excitations	-	-	Hedelin & Huber (1990)
Paired pulsing	Subharmonics; very weak intensity of the first component	-	Martin (2012)

1.3. Other definitions of related phenomena

This section summarizes various terms that have been used to describe period doubling and its related phonation in the literature.

'Biphonation', defined as having two independent frequencies, has been found in cries of newborn infants with or without congenital disorders (Sirviö and Michelsson, 1976; Herzel and Reuter, 1996), non-cry vocalization samples from children of different ages (9 years old, Herzel and Reuter, 1997; 11-24 months, Robb and Saxman, 1988), normal and forceful phonation from healthy speakers or those with laryngeal nerve paralysis (Tigges et al., 1997), patients in a voice clinic, singing, and excised larynx experiments (Herzel and Reuter, 1996), and dysphonic patients (Herzel et al., 1994). In particular, biphonation has been viewed differently from 'double harmonic break' (Sirviö and Michelsson, 1976) or 'harmonic doubling' (Buhr and Keating, 1977; Keating, 1980) which describes the presence of extra harmonics that are parallel to f0 and its harmonics. The two series of f0s in biphonation are not required to have parallel melodies (Sirviö and Michelsson, 1976; Robb and Saxman, 1988), and have been characterized by nonparallel dark lines on the spectrogram representing independently varying spectral peaks (Robb and Saxman, 1988; Herzel et al., 1994).

Specifically, biphonation was found to occur with high-pitched voice (Herzel and Reuter, 1997; Tigges et al., 1997). An incomplete glottal closure, high-pitched phonation (Herzel and Reuter, 1997; Tigges et al., 1997), and large subglottal pressure (Tigges et al., 1997) were claimed to induce desynchronized vocal folds (vibrating at two different frequencies), which is often the source of biphonation. Herzel and Reuter (1997) found biphonation in high-pitched vocalizations of a 9-year child as the voice transitioned from falsetto ($\sim 950Hz$) to whistle-like vocalizations ($1000-1700Hz$). In contrast, Buhr and Keating (1977) attributed harmonic doubling to the result of vocal instability at extreme ranges of the modal register, and stated that the effect was typically seen in children across a range of $160-250Hz$ (in shifts between modal and fry registers) and $380-800Hz$ (in shifts between modal and high registers).

A closely related term to 'biphonation' is 'bifurcation', brought up by Mende et al. (1990) using newborn infant cries, and Herzel et al. (1994) from dysphonic patients. Titze (1994) described bifurcation as "period and amplitude alternating between two states in asymmetric vocal folds", and can be seen as period doubling, period tripling, period quadrupling, or low-frequency modulation over periods. But Herzel et al. (1994) treated bifurcation differently from period doubling, such that the former refers to modulation from another frequency that is not in a harmonic relationship to f0, whereas the latter consists of a new cycle that is roughly double that of the original period. In addition, Mende et al. (1990) found that bifurcations often signify the presence of noise-like segments in spectrograms.

Kreiman et al. (1993) described a similar phonation which they call 'bicyclicity', involving a pattern of two cycles which differ in period and/or amplitude that shows a 'big-small-big-small' appearance in the acoustic waveform. Timcke et al. (1959) mentioned that 'biphasic' vibrations were seen in normal subjects, which were typically low and harsh in quality, and which included a strong subharmonic at half the fundamental frequency. And Moore and Von Leden (1958) termed a voice quality found in normal larynx as 'dicrotic dysphonia', which was characterized by two glottal pulses occurring in rapid succession, followed by a longer closed period and which can alternate with a single opening pattern of normal phonation. Moreover, the authors hypothesized that the subglottal pressure was significantly reduced during the two open cycles, and so required a closed phase of from one-third to half the longer cycle to restore enough pressure to initiate another vibration.

In the spectral domain, a defining characteristic of period doubling and related aforementioned voice qualities is the presence of subharmonics (Švec et al., 1996; Herzel, 1993; Timcke et al., 1959; Kelman, 1981; Monsen, 1979; Keating et al., 2015). The location of subharmonics is typically at one-half (Kelman, 1981; Timcke et al., 1959; Herzel, 1993; Monsen, 1979) or one-third of the preceding frequency (Herzel, 1993; Monsen, 1979). However, Švec et al. (1996) found that the frequency ratio of the found subharmonic vibratory pattern was 3:2 (rather than

frequency halving 2:1), where the parameters such as a large open quotient and high airflow could distinguish well the subharmonic vibration from vocal fry.

In the dissertation, period doubling is identified and defined mainly from temporal domain characteristics, which appears as two alternating cycles differing in frequency or amplitude in the acoustic waveform. Upon a *post-hoc* examination of the identified tokens of period doubling, in the spectral domain, there are subharmonics that are integer divisors of the fundamental frequency, and the amplitude of the subharmonics gradually increases in higher frequencies and becomes stronger than that of the original harmonics; Chapter 2 will discuss the spectral patterns in more detail.

1.4. Period doubling in typical speech

1.4.1. Production studies

Limited studies have examined the properties of period doubling in typical speech. Studies on articulatory and physiological mechanisms are rare, as most used acoustic analysis to study period doubling. Among the earliest descriptive studies, Schreibweiss-Merin and Terrio (1986) documented volitionally produced period doubling ('diplophonia') cases from a 24-year-old female English speaker where the two f0s were above $300Hz$. They found that listeners perceived a low-frequency buzz that was not present in the signal. In addition, the speaker phonated significantly longer, and had shorter pause times, during diplophonic voicing. Similar to what Moore and Von Leden (1958) had suggested, Schreibweiss-Merin and Terrio proposed that greater air pressure is needed for period doubling; without it, a finer and rapid termination of voicing cannot be realized, which thus results in longer voicing periods.

More recent studies investigated period doubling from perspectives of acoustic attributes, quantification measures, and linguistic distribution. Keating et al. (2015) summarized the acoustic properties of creaky voice and the subtypes of creaky voice which can be characterized by different combinations of acoustic attributes. As a subtype of creaky voice, period doubling is

characterized as having higher noise, increased glottal constriction, and presence of (stronger) subharmonics (Keating et al., 2015). As already noted, period doubling contains two sets of harmonics which can be seen on the spectrum that one set of harmonics is stronger. To quantify the relative strengths of the two sets of harmonics, Subharmonic-to-Harmonic Ratio (SHR) was proposed by Sun (2002) to measure the magnitude of subharmonics with respect to harmonics that reflects the degree of deviation from modal voice. Herbst (2020) evaluated the validity and accuracy of SHR using both naturalistic electroglottographic (EGG) data from an English corpus and synthesized EGG signals with varying degrees of amplitude and frequency modulation, fundamental frequency, periodicity, and signal-to-noise ratio (SNR). They found that the SHR metric was robust in assessing the degree of subharmonics that appeared in voice signals (maximum sensitivity: 87%, specificity: 90%) given an adaptive parameter value setting. Redi and Shattuck-Hufnagel (2001), who examined period doubling ('diplophonia') at English phrase boundaries, found that utterance-final boundaries were associated with higher glottalization rates than utterance-medial boundaries, and full intonation phrases were glottalized more often than intermediate intonational phrases.

Besides English, several studies also investigated the distributions of period doubling in speech and prosody, and in non-tonal and tonal languages. In Swedish read speech, Hedelin and Huber (1990) found that period doubling ('diplophonic phonation') occurred predominantly at utterance-internal positions in word junctures between adjacent vocalic sounds, where it serves as a hiatus. Period doubling often cooccurred with other voice qualities (e.g., breathiness) at intonation offset locations (low f0 contour), or in the transition before a stretch of creaky voice and/or devoicing (Hedelin and Huber, 1990). Kim et al. (2020) used creaky voice detectors in low tones in Beijing Mandarin, Cantonese, and White Hmong. They found irregular voicing, period doubling (regular voicing with strong secondary excitation peaks), and vocal fry (regular without strong secondary excitation peaks). However, no clear separation of these creak types was seen using principal component analysis (PCA). Yu (2010) also investigated creaky voice ('laryngealization') in Cantonese and Mandarin and found that vocal fry (a train of discrete pulses

of extremely low frequency, with damping between pulses) and period doubling were abundant in both females and males. In Cantonese, period doubling was common in females, which according to Yu can be due to the fact that period doubling is not contingent on particular f0s. In Mandarin, female speakers had both period doubling and vocal fry and male speakers had vocal fry as the dominant form of creaky voice (Yu, 2010). Further, Yu found that the creaky voice in Mandarin Tone 3 was typically realized as vocal fry rather than as period doubling. Moreover, for Mandarin, 'laryngealization' (including vocal fry and period doubling) has been found as a cue to signify low tones more than other tones (Belotel-Grenié and Grenié, 1997), and boundary phenomenon such as being a marker of utterance finality (Belotel-Grenié and Grenié, 2004). For other uses, Li et al. (2020) found that multiple pulsing appears frequently in sarcastic speech in Mandarin.

In sum, past studies have clarified some of the articulatory and acoustic properties of period doubling in typical speech, as well as the linguistic environments where it has been found to occur. However, a systematic and consistent description is lacking, as well as the distributional properties of period doubling in relation to linguistic categories. Thus, the current dissertation aims to probe the defining characteristics of period doubling from both articulation, acoustics, and distributional aspects.

1.4.2. Perception studies

The perception of period doubling involves both pitch and voice quality percepts. Though multiple frequencies reside in period doubling, the voice is often heard as rough, which likely results from an indeterminate pitch (Yiu et al., 2002; Keating et al., 2015). A recent study by Davidson (2020) sought to resolve the pitch percept problem during period doubling by asking English listeners to rate the pitch during utterances that at least partially contain creak (with multiple pulsing and prototypical creak) and those that are fully modal. Davidson found that listeners perceived lower pitch during utterances containing multiple pulsing, but only if the speaker's modal voice does not have a low habitual pitch. This suggests that due to the indeterminate or bitonal percept, period doubling likely cues a lower pitch for high-pitched voices but not for low-pitched voices.

Researchers looking into the relationship between roughness and acoustic attributes have found that a low fundamental frequency, noise, and irregularities in the voice induce and reinforce a percept of roughness. Specifically, in synthesized period doubling signals, the perceived pitch becomes lower as the amount of modulation increased, which can be predicted by measures in the frequency domain (e.g., subharmonic-to-harmonic ratio; Sun, 2002); however, perceived pitch differs across fundamental frequencies and modulation types. For example, a lower f0 prompts earlier identification of the subharmonic (the lower frequency) as the true pitch, and the pitch drops more quickly in frequency- than amplitude-modulated tokens (Sun and Xu, 2002; Bergan and Titze, 2001). In addition, Fraj et al. (2012) assessed the perception of simulated disordered voice using additive pulsatile noise, frequency jitter and tremor, and amplitude shimmer and tremor. They found that the degree of roughness is positively associated with pulsatile noise and frequency jitter while negatively related to the vocal frequency.

Researchers also studied the relationship between subharmonics and roughness in pathological voices in connected speech (Omori et al., 1997; Kramer et al., 2013; Murphy, 2000). In particular, Omori et al. (1997) found that the degree of roughness was related to the frequency and power of subharmonics, such that the stronger the amplitude and the lower the frequency, the rougher listeners perceived the voice. In contrast, Kramer et al. (2013) found that the degree of irregularity and the percentage of low f0 values, but not the power of subharmonics, were salient cues to the degree of roughness.

As for the perceptual relationship between period doubling and other subtypes of creaky voice, it was found that period doubling can be easily distinguished from amplitude-modulated and aperiodic voice using samples from speakers with vocal pathology (Gerratt and Kreiman, 2001). Previously, diplophonic voice had been found to be perceptually distinct from bicyclic, noisy, and other clinical rough or breathy voices (Kreiman et al., 1993).

Still, questions remain as to (1) what the relationship is between the perceived pitch(es) during period doubling and the extent of frequency and amplitude modulation in the time domain; (2) what attributes in period doubling differentiate it from other types of creak; (3) what the

relationship is between the perceived pitch(es) and the harmonics and subharmonics (apart by one octave) in the spectral domain; (4) how period doubling and other types of creak are similar and dissimilar from each other linguistically. This dissertation will address questions (1), (2), preliminarily (3), and has implications for (4), and leave the other questions to future studies.

1.5. Other studies on period doubling or related phenomena

1.5.1. Pathological voice

In voice disorders, the phenomenon of period doubling has long been noticed and investigated, as it appears in a closely related pathology, diplophonia, and more general dysphonia.

Here, as a type of dysphonia, diplophonia has been defined as the perception of more than one fundamental frequency component in a voice (Gerratt et al., 1988). Articulatorily, this voice is caused by independent vibration of the two vocal folds at slightly different frequencies, similar to a beating phenomenon (Gerratt et al., 1988; Dejonckere and Lebacq, 1983), or by the vocal folds oscillating abnormally to produce double- or multiple-phased closing/opening cycles (Dejonckere and Lebacq, 1983), or by the vibration of the ventricular folds along with the vocal folds (Ward et al., 1969). Note that the use and definition of diplophonia in these studies is distinct conceptually from the meaning found in linguistic studies that in fact refer to period doubling such as Redi and Shattuck-Hufnagel (2001).

Acoustically, temporal variation of the vibratory pattern of vocal cords is seen in patients with diplophonia; that is, an alternation between vibrations with and without glottal closure (Kiritani et al., 1993; KIRITANI, 1995). In both studies, Kiritani and colleagues used a high-speed digital imaging technique to examine the patients' phonation, which showed a difference in the vibratory frequency between the left and right vocal folds. They also found that glottal closure and the phase of the movements of vocal folds are closely related such that glottal closure is complete and the glottal excitation is strong when movements are in phase; but glottal closure is incomplete and the excitation is weak to result in a quasi-periodic vibration when the movements are out of

phase. Alternatively, diplophonia is also seen from the addition of a second subharmonic series to the original f0 harmonic structure (Dejonckere and Lebacq, 1983; Kelman, 1981).

Using a binary classifier, Bae et al. (2019) compared different measures in discriminating subtypes of diplophonia in clinical diagnosis. They focused on three types of asymmetric vibratory patterns in diplophonia: left and right vocal folds asymmetry (called 'aerodynamic coupling' due to intermodulation via the glottal flow; Aichinger et al., 2017), anterior and posterior cords asymmetry (called 'mechanical coupling' for tissues connecting the distinct oscillators; Aichinger et al., 2017), and vocal and ventricular folds asymmetry, and found that digital kymography had the highest diagnostic accuracy, though the auditory-perceptual evaluation was the easier and fastest method for trained raters.

The relationship between roughness and diplophonia was also discussed in past studies. Dejonckere and Lebacq (1983) found that due to a small closed phase, diplophonia is associated with excessive airflow, which thus results in hoarseness. Imaizumi and Gauffin (1992) asked listeners to rate synthetic voices that are created by cycle-by-cycle perturbation, using the GRBAS (Grade, Roughness, Breathiness, Asthenia, Strain) scale (De Bodt et al., 1997). Similar to Kreiman et al. (1993), they found that listeners can discriminate between roughness and diplophonia, such that roughness was perceived if listeners holistically perceived the perturbations as one coherent quality but diplophonia would be perceived when listeners analytically perceived the effect of perturbations as two or more separate frequency components.

Further, Aichinger et al. (2016) investigated the values of jitter and shimmer in diplophonia. The results showed that jitter and shimmer had higher median values for diplophonic than for non-diplophonic voices when the voices had equal grades of hoarseness; moderately dysphonic voices also had a larger variance of jitter than a corpus without diplophonic samples.

According to Aichinger et al. (2019), extra pulses occur frequently in dysphonic voices and this occurrence may be caused by (slight) desynchronization of the anterior and posterior vocal folds. They noted that this vibration mode can be regarded as an intermediate stage between

modal phonation and biphonation or diplophonia, which contains a more extreme case of extra pulsing known as 'double pulsing'.

1.5.2. Singing registers (period-doubled register)

In singing voice, period doubling has been documented in different types of singing that generally involve two separate frequencies, which is typically induced by both the vocal folds and the ventricular folds. In Mongolian Kargyraa throat singing, Tibetan Dzo-ke chants, Sardinian A Tenore Bassu singing, as well as in Japanese traditional Noh singing, the perceived pitch is about one octave lower than the fundamental frequency, and the ventricular folds oscillate at half the frequency of vocal folds to create an extremely low-pitched glottal phonation characterized by the damping of every other glottal excitation (Lindestad et al., 2001; Fuks et al., 1998; Sakakibara, 2003; Henrich et al., 2006; Bailly et al., 2010; Yoshinaga and Kong, 2012).

Moreover, for Sardinian singing, Henrich et al. (2006) found that the subharmonic vibratory pattern forms quickly in musical sentences starting with sonorous consonants /m/ or /b/, and that very relaxed vocal folds and low subglottal pressure often favor the occurrence of period doubling, which is consistent with singers' laryngeal sensations. Bailly et al. (2010) stated that from simulations of this specific singing style, the aerodynamic interaction between vocal and ventricular folds can predict period doubling patterns of alternating vibration amplitude. For Japanese traditional singing, Yoshinaga and Kong (2012) found that the singing pattern is characterized by low open quotient and high speed quotient; ventricular and aryepiglottic fold vibrations which are half the fundamental frequency add to the low-pitch sounding.

Aside from these styles, the Russian "lament style" register exhibits rich voice quality types such as vocal fry, diplphonia, and simultaneous amplitude and frequency modulation (Mazo, 1995). Mazo (1995) showed that instability of pitch, duration, and voice quality was abundant in this emotional register. Further, there was increasing intensity of the upper frequencies toward the end of the vowel. Diplophonia, defined as split f0, contained two sets of f0 with ratios of 3:4 and 2:3. The simultaneous amplitude and frequency modulation had a modulation around

40 to 50Hz, which was distinct from aperiodic fluctuations such as jitter or shimmer, or vibrato at a slower rate. Specifically, the simultaneous modulation occurred in the highest range of the melodic pattern where the contour changed direction from its previous tone (Mazo, 1995).

1.6. Specifics and structure of dissertation

Given that the majority of past studies on period doubling or related phenomena have been focused on pathological voices, here I investigate period doubling in typical speech. I choose Mandarin Chinese as the target language. Mandarin is a tone language, where both period doubling and vocal fry can occur during the creaky voice associated with certain low pitches. Thus, both creaky voice subtypes serve as sub-phonemic and secondary correlates to tone categories. In the production experiments, I document both period doubling and vocal fry and compare their articulatory and acoustic characteristics and linguistic distributions. In perception experiments, I resynthesized stimuli that build on the findings in the production studies and recruit both groups of English and Mandarin listeners to further investigate the language and tone effect on perceiving period doubling as linguistic tones.

Having summarized across the past findings, the four research questions are restated as follows:

1) What are the articulatory aspects of period doubling found in typical, non-disordered, speech?

In Chapter 2, I address this question by presenting an articulatory study using electroglottography (EGG) recordings of a scripted corpus in Mandarin. I incorporate articulatory measures derived from EGG waveforms and propose temporal and spectral analysis to quantify the physical properties of alternating glottal pulses during period doubling, with reference to vocal fry and modal voice. Part of the earlier version of this chapter is in Huang (2022).

2) What are the acoustic characteristics of period doubling that distinguish it from other voicing types?

3) Where does period doubling occur linguistically? Specifically, does it occur in specific tonal and phrasal environments?

In Chapter 3, I address these two questions using simultaneous audio recordings of the same Mandarin corpus. First, I analyze the acoustic features of period doubling, as compared to vocal fry and modal voice. Then, using machine classification on both acoustic features and the articulatory measures discussed in Chapter 2, I study the importance of these factors in distinguishing different voicing types. Finally, I discuss the results of tonal and phrasal distributions of period doubling (and vocal fry) as identified from the corpus.

4) How do listeners perceive pitch, voice quality, and tone during period doubling?

In Chapter 4, I address this question by studying the perception and imitation of pitch and voice during period doubling. I resynthesized stimuli of period-doubled voice based on the empirical temporal measures in Chapter 2 to determine how listeners perceive and imitate period doubling in an artificial language learning paradigm with two implicit categories of 'high' versus 'low' tones. I investigate whether and when they have a clear preference towards one of the two octaves when identifying linguistic tones, or perceive two cooccurring tones and/or roughness with varying degrees of modulation. I discuss the effects of different types and degrees of modulation to inform theories of pitch and tone perception.

Chapter 5 includes a general discussion and conclusions on the phonetics of period doubling, the implications of this work for voice and linguistic theories, and avenues for future exploration.

Chapter 2

How is period doubling articulated? – An electroglottographic study of Mandarin

2.1. Introduction

Period doubling (PD) refers to a subtype of creaky voice with two simultaneous periodicities, which contributes to an indeterminate pitch with low and rough quality (Keating et al., 2015; Yu, 2010; Schreibweiss-Merin and Terrio, 1986). As a special case of "multiply-pulsed" voice, period doubling is often additionally characterized as having increased noise and glottal constriction (Keating et al., 2015), as well as by the presence of subharmonics (or interharmonics) in the spectrum (Yu, 2010; Omori et al., 1997; Kramer et al., 2013; Herbst, 2020). Though commonly found in spoken language ($\sim$ 25% of normal speakers' utterances; Klatt and Klatt, 1990), little is yet known for the articulation of period doubling – how different are the alternating glottal cycles in the voice source? What is the relationship between the two simultaneous periods/frequencies? Moreover, how do phonatory properties of period doubling differ from those of vocal fry, another common subtype of creaky voice? Investigating the articulatory aspects of period doubling clarifies the distinctions within creaky voice, and lays the basis for studying linguistic functions and perceptual meanings of period doubling.

This chapter focuses on the articulatory aspects of period doubling in spoken language, using Mandarin as a representative language with predictable non-modal phonation along with contrastive pitch contours. As noted in Yu (2010), non-modal phonation in Mandarin typically involves vocal fry and period doubling, with different distributions across genders such that vocal

fry is dominant in male speakers, and both types of voice were observed in female speakers. In the present study, the canonical period-doubled voice samples used are utterances where two periods alternate in frequency and/or amplitude as they appear in the electroglottographic (EGG) waveforms from a read speech corpus in Mandarin. To quantify the physical aspects of articulation of period doubling, first I present characteristics that are unique to the two alternating cycles within period-doubled voice based on the EGG waveforms. Second, I report various glottal constriction measures of period doubling with reference to more established voice categories – vocal fry and modal voice – to better capture the alternating glottal cycles. Then, I discuss representative spectra that are associated with multiple kinds of period doubling found in the corpus. Determining what are the articulatory features of period doubling has implications for its linguistic function and role in speech perception and production, which may be potentially different from other types of creaky voice, and helps clarify its place within a taxonomy of creaky voice subtypes.

2.2. Background

2.2.1. Production studies on period doubling

Production studies that have documented period doubling have often found that the periods and amplitudes alternate between two states (high and low; long and short) in the speech signal (Titze, 1994). See also a review by Gerratt and Kreiman (2001) for other related phonation types, including 'biphonation', 'bifurcation', etc. However, apart from limited acoustic and perceptual descriptions, there lacks a systematic analysis of the phonatory properties of period doubling needed to quantify the regularly alternating pattern in this subtype of creaky voice in natural speech. The defining characteristics of period doubling have important implications for the interaction between pitch and voice quality, and between tone and phonation as linguistic categories. For example, Davidson (2020) found that English listeners perceive lower pitch during utterances that at least partially contain creak with multiple pulsing. In Swedish, Hedelin and Huber (1990)

found that period doubling ('diplophonic phonation') occurs predominantly utterance-medially at word junctures between two vowels. In Cantonese, period doubling was common in females (Yu, 2010). In Mandarin, female speakers had both period doubling and vocal fry, while male speakers had vocal fry as the dominant form of creaky voice (Yu, 2010). Moreover, though non-contrastive, laryngealization (which includes vocal fry and period doubling) has been found to carry phonological function by signifying low tones more than other tones (Belotel-Grenié and Grenié, 1997), and boundary phenomenon such as being a marker of utterance finality (Belotel-Grenié and Grenié, 2004). Also found are correlations between creaky voice, vocal attractiveness (Xu and Lee, 2018), and sarcastic speech (Li et al., 2020). But whether period doubling and vocal fry differ in their phonological or social functions remains unclear. To approach the question regarding the role of subtypes of creaky voice in speech production and perception, here, I study how period doubling is articulated differently from vocal fry and modal voice, using EGG waveforms.

2.2.2. Articulatory measures using EGG

Electroglottography (EGG) is a non-invasive method of measuring vocal fold contact during vibration (Fourcin, 1986; Baken and Orlikoff, 2000). This device generates a low-amplitude, high frequency current passing between the two electrodes that are typically placed around the thyroid prominence. The current flow thus tracks the vocal fold activity by varying admittance of the glottis: more vocal fold contact leads to greater relative measures of contact via EGG (Howard, 1995). The admittance is proportional to the contact area between the two vocal folds, so that the EGG signal is deemed as a decent representation of the time-varying contact area of the vocal folds. EGG has been widely used to study and document phonation types and voice qualities across languages (e.g., Maa: Guion et al., 2004; Vietnamese: Michaud, 2004; Tamang: Mazaudon and Michaud, 2008; Takhian Thong Chong: DiCanio, 2009; Santa Ana Del Valle Zapotec: Esposito, 2010; Gujarati, White Hmong: Esposito and Khan, 2012; English: Garellek, 2014; Yi

languages: Garellek et al., 2016). See also D'Amario and Daffern (2017) for a systematic review on the use of EGG in singing voice. I employ established physiological measures derived from EGG waveforms to characterize the articulation of period-doubled pulses. An EGG pulse consists of a closing, or contacting phase, as well as an opening, or de-contacting phase. Three families of time-domain measures quantifying the duration and slope of the contacting and de-contacting phases within a glottal cycle during vocal fold vibration are contact quotient, speed quotient, and peak increase in contact.

The most commonly-used EGG measure, contact quotient (CQ), is defined as the proportion of a cycle during which there is contact (Rothenberg and Mahshie, 1988; Baken and Orlikoff, 2000). Studies have shown that EGG well correlates with vocal fold contact, measured by direct imaging or airflow (Childers et al., 1990; Holmberg et al., 1995a; Granqvist et al., 2003; Herbst et al., 2017). Thus, CQ is assumed to be a representative measure of vocal fold contact. Creaky voice (laryngealization, tense voice) usually shows a greater value of CQ because of longer glottal constriction whereas breathy voice usually shows a smaller CQ. Past studies have used CQ extensively including voicing and phonation types in different languages (Guion et al., 2004; Michaud, 2004; Mazaudon and Michaud, 2008; DiCanio, 2009; Esposito, 2010; Esposito and Khan, 2012; Garellek, 2014; Kuang and Keating, 2014; Garellek et al., 2016).

Apart from linguistic studies, CQ has been found to signal emotional states (van Mersbergen et al., 2017), and has been used in singing and voice studies. For example, EGG-derived CQ was used to assess voice quality in different styles of singing from a soprano (Bateman, 2003); open quotient (OQ, defined as the complement of CQ within a glottal cycle) was used to differentiate between modal phonation and falsetto phonation with the consideration of simultaneous f0 changes (Yokonishi et al., 2016; Echternach et al., 2010; Henrich et al., 2005).

In addition, CQ has been used in studies of voice disorders. For example, Guzmán et al. (2016) combined CQ with pressure measures to investigate vocal fold adduction in subjects with normal voice, trained voice, muscle tension dysphonia, and unilateral vocal fold paralysis. Relat-

edly, OQ quantifying the ratio of the opening phase was effective in diagnosing hypofunctional dysphonia (Szkiełkowska et al., 2018) and describing variability in the phases of vocal-fold vibration (Sapienza et al., 1991). CQ was also useful in probing effects of age and gender on vocal fold vibratory behaviors during vowel prolongation and connected speech (Ma and Love, 2010).

However, despite the advantages of CQ, one caution of use is the determination of the glottal opening and closing instants. Depending on the *a priori* chosen moments or thresholds to pinpoint where the glottal contact starts and ends, the contact and opening phases can vary to a great extent. Commonly-seen methods are derivatives of EGG (locating the peak velocity of vocal fold changes), while threshold (using a preassigned percent of the cycle amplitude) and hybrid methods (combining the derivative and threshold) are adopted as shown in later studies that EGG contact calculation at a threshold of 20% or 25% was found to best correlate with vocal fold contact measure (closed quotients) obtained from direct videokymographic imaging (Herbst, 2004) and effective in distinguishing linguistic voice quality (Garellek et al., 2016). Also see a recent review by Herbst (2020) on the limitations and interpretations of CQ given different landmarks. In this chapter, I will discuss and compare the results derived from all available methods of CQ calculation based on EGGWorks: derivative, threshold, hybrid, and Henry Tehrani method which uses the derivative and incorporates the DC component of the EGG signal into the calculation (Tehrani, 2009).

The second measure, speed quotient (SQ), is defined as the ratio of the duration of the contacting phase to the duration of the de-contacting phase of the EGG waveform (Sapienza et al., 1998; Slavit and McCaffrey, 1995). Thus, SQ provides an estimation of the symmetry of the glottal cycle. Creaky voice tends to have a lower value of SQ (more constricted and abruptly closing) whereas breathy voice tends to have the highest SQ. Although used less often than CQ, SQ has been adopted in studies on segmental and suprasegmental contrasts, voice qualities, and pathologies. For example, in German, Marasek (1996) used SQ to study lexical stress and tense and lax vowels and found that the increase of steepness in closing and opening slopes is one of the

main correlates of stress. In a follow-up study by Mooshammer (2010), SQ was found to signal focus, but not stress or vocal effort. SQ can also be influenced by vowel types, gender, and age (Marasek, 1996; Ma and Love, 2010). In addition, SQ was an important predictor besides CQ in distinguishing tense and lax phonation contrasts in Yi languages (Kuang and Keating, 2014); it well correlated with tenseness of the tones in Hanoi Vietnamese (Shimizu and Dantsuji, 2000); it was also used to examine the tonal quality and laryngeal features in Jianchuan Bai (Wang, 2015).

Besides linguistic studies, Esling et al. (1992) used SQ to classify different voice categories at different pitches including modal, creaky, falsetto, breathy, whispery, harsh, and ventricular voices and found that SQ is independent from f0 in creaky, modal, and falsetto voices. Chen et al. (2002) found that SQ was lower in vocal fry than modal for both female and male speakers, indicating a longer duration of the opening phase per glottal cycle. Women also demonstrated a greater decrease in SQ during vocal fry than men, indicating greater asymmetry of the glottal pulse. In voice disorders, similar to CQ, SQ is also used to describe variability in the phases of vocal-fold vibration (Sapienza et al., 1991), and differentiate vibratory patterns among patients with different types of disorders (Hanson et al., 1983).

The third measure, peak increase in contact (PIC), originated from the Derivative-EGG closure peak amplitude (DECPA) measure in Michaud (2004) to study focus prosody (also see a related non-derivative-based measure of average contacting "speed" in Baken and Orlikoff 2000). PIC is defined as the amplitude of the positive peak in the derivative EGG signal, referring to the maximum rate of increase of vocal fold contact. It measures the speed of glottal closure. Breathy voice with a spread glottis, produced with faster and shorter glottal closure, have greater PIC values than modal or constricted voice qualities produced with slower glottal closure. Studies on contrastive phonation categories showed differences in PIC, including breathy versus modal (Esposito, 2012; Esposito and Khan, 2012; Keating et al., 2010; Garellek et al., 2016), and tense versus lax voices (Kuang and Keating, 2014).

In the current articulatory study, I expect that vocal fry will have high CQ, low SQ, and low PIC values. Period doubling, according to Keating et al. (2015), behaves similar to vocal fry in terms of glottal constriction; thus, I would expect it to also exhibit high CQ, low SQ and low PIC values.

2.3. Methods

In this section, I describe the methodology and present examples of period-doubled phonation located from speech production data in Mandarin. I will discuss how period doubling is identified in the voice source, as measured by EGG signals, and introduce measures for quantifying the alternating pulse shapes on EGG waveforms to determine the defining properties of period doubling.

2.3.1. Materials

The speech materials are from a Mandarin read speech corpus collected to document a full range of tonal contextual variation (Huang, ur). The corpus was created to investigate the relationship between pitch and voice quality during coarticulated Mandarin tones. The stimuli consist of a fixed carrier sentence with varying trisyllabic compound words: *wo3 tɕau1 niˀi3 WORD tsən3-mɤ0 ʂʷo1* 'I teach you WORD how to say.' In the compound words, each of the four Mandarin tones was flanked by varying Mandarin Tones 1–4, for a full range of $4 \times 4 \times 4 = 64$ combinations. The compounds are frequently-used words and familiar to native Mandarin speakers. Some words have left or right branching syntactically whereas others are fixed or metaphorical expressions. Three sets of 64 sentences with two repetitions (384 sentences in total) were elicited per recording. The 'neutral tone' was not investigated for the current purpose. Period-doubled phonation and vocal fry were identified anywhere in the phrase – during the target words or the carrier sentence.

2.3.2. Participants

Twenty native Mandarin speakers ($10F,\ mean\ age = 20.1,\ range = 18 - 22$) participated in the production experiment. Fifteen were from northern provinces and five from southern provinces in China. One speaker spoke a Sichuan dialect, and three speakers also spoke a Wu dialect (Shanghainese or Yangzhounese). All speakers had moved to the United States before or for college. At the time of the experiment, the average time they have spent in the US was 3.7 years across all participants (ranging from $0 - 8$ years). The speakers were recruited from the UCSD Psychology Subject Pool and received undergraduate course credit for their participation. No language or speech disorders were reported.

2.3.3. Procedure

Participants were recorded in a sound-attenuated booth at the Phonetics Lab at the University of California San Diego. The experiment was implemented and presented in PsychoPy (Peirce, 2007). Speakers were instructed to produce the sentences as if they were in a natural conversation to ensure the naturalistic quality of speech. To avoid any strategy or fatigue in producing the sentences, picture fillers were used every four sentences. Participants needed to briefly describe the object that the picture showed. The experiment lasted for 40-50 minutes, and participants could take breaks during the experiment.

Audio and electroglottography (EGG) recordings were obtained simultaneously. Participants wore a Shure SM-10 head-mounted microphone and an electroglottographic collar attached with two electrodes right below their thyroid prominence. EGG was used to record the degree of contact between the vibrating vocal folds directly from the larynx. The EGG signal was recorded using a Glottal Enterprises EG2-PCS. Both audio and EGG signals were pre-amplified through a Focusrite Scarlett 8i6 preamplifier and digitized with the computer's sound card using Audacity at a $44,100Hz$ sampling rate and 32-bit float rate to maximize precision.

2.3.4. Analysis

Criteria for identifying period doubling The phrases and words in the audio recordings were segmented in PRAAT (Boersma and Weenink, 2022) and the EGG recordings were used to locate source pulses with period-doubled voice and vocal fry. I used the EGG signal instead of the audio to identify period doubling because the EGG has no formant structure, which avoids possible formant-induced interferences with the voicing signal. Based on Kreiman et al. (1993), canonical period doubling is often characterized by sequences of two different cycles which differ in amplitude of the pulses, or length of the periods or frequency; in other words, by amplitude modulation or frequency modulation. For example, amplitude-modulated period doubling manifests "high-low-high" alternating amplitudes whereas frequency-modulated period doubling manifests "long-short-long" alternating periods (but also with amplitude modulation) on the waveform. Figures 2.1 and 2.2 show sample EGG waveforms and spectra of typical-looking period doubling from two tokens in the corpus (from subjects F33 and F30). There was also an asymmetry regarding the distribution of amplitude (5068) versus amplitude- and frequency-modulated (1480) tokens. The degrees of differences between sub-cycles of period doubling were on a continuum, which sometimes can be subtle. To be as inclusive as possible, I labeled any tokens that resemble Figures 2.1 or 2.2 in the following subsection, or when the voicing caused pitch halving and/or doubling in PRAAT's f_0 algorithm.

The EGG recordings were pre-processed using a built-in band-pass filter between $40Hz$ and $22050Hz$ with smoothing at $50Hz$ in PRAAT to remove the low-frequency DC component of EGG below $40Hz$ and higher frequency noise. Note that the instances of frequency-modulated period doubling tend to cooccur with amplitude modulation, similar to what Mazo (1995) reported in a Russian 'lament style of singing'. Also, though period-doubled phonation in pathological speech and certain singing styles involves the vibrations of ventricular folds, Bailly et al. (2010) on Sardinian singing found that EGG exclusively encodes information about the two consecutive

glottal cycles, so it is unlikely the period-doubled phonation observed here is due to ventricular fold vibrations.

Frequency and amplitude ratios To quantify the differences in the two alternating periods in period doubling, I obtained the duration and amplitude of each alternating strong and weak adjacent cycles during samples of period doubling containing at least one pair of cycles. In any pair of pulses, there are three periods defined. Following Bailly et al. (2010), who investigated period doubling phonation in singing, here I define T_0 as the longest period, from strong-to-strong pulses, which is also the presumptive fundamental cycle. For the two shorter periods, the strong-weak cycle is defined as T_1 and the other weak-strong cycle is defined as T_2. As equation 2.1 shows, the duration of the two shorter cycles always adds up to the longer duration: $T_1 + T_2 = T_0$. The frequency-ratio R_T is defined as the ratio of the duration of the two alternating glottal cycles: the longer strong-to-weak period T_1 divided by the shorter weak-to-strong one T_2, or the higher frequency f_2 over the lower frequency f_1. The formula in 2.2 shows the calculation. Illustrations are shown on the waveforms in Figures 2.1 and 2.2. The frequency ratio R_T thus indicates the relative relationship between any of the two short periods, which together contribute to the fundamental cycle and/or frequencies.

Similarly, the amplitude ratio R_A, defined as A_1, the amplitude of the stronger pulse divided by A_2, that of the weaker pulse, is shown in equation 2.3. The top panels of Figures 2.1-2.2 also illustrate the calculation of the respective amplitude values. Here, A_1 and A_2 are defined as the total displacement from the minimum to the maximum of a glottal cycle, to better capture the strength of each excitation. (A considerable number of weak pulses did not exceed the zero-crossing floor, which is an arbitrary threshold.)

The time points for different periods at each rising zero-crossing, and the amplitude values at the extremes of the stronger and weaker pulses were extracted using a custom PRAAT script and the ratios of the periods/amplitudes were further calculated in R (R Core Team, 2022). After

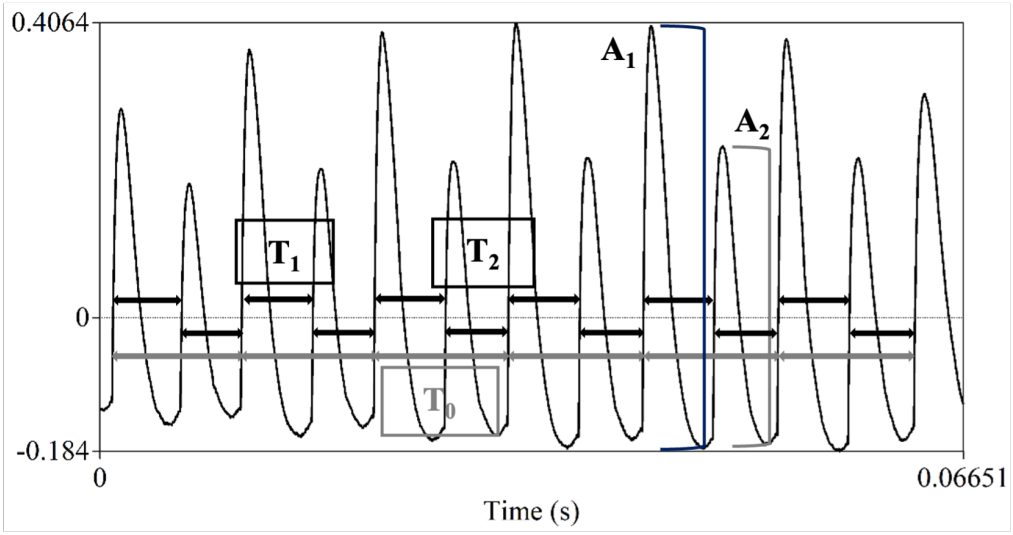
the calculation, I removed outliers if after log transformation, they had a z-score larger than 2.5 standard deviations away from the mean within each subject.

$$T_1 + T_2 = T_0 \tag{2.1}$$

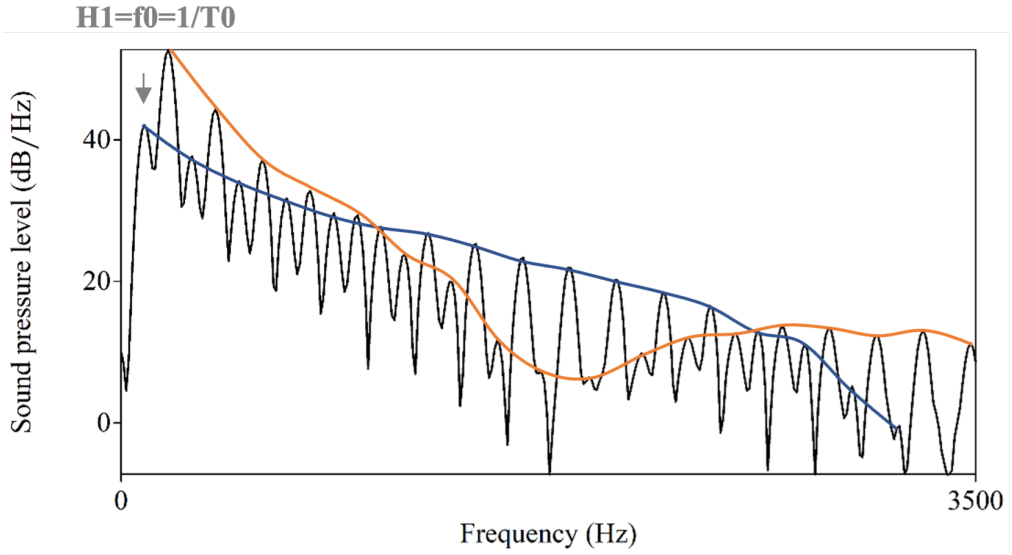
$$R_T = \frac{T_1}{T_2} = \frac{f_2}{f_1} \tag{2.2}$$

$$R_A = \frac{A_1}{A_2} \tag{2.3}$$

The EGG signal for period-doubled voice is analyzed in both time and frequency domains. In the following sections, I first review the prevalence of period doubling (PD), and differences of the two adjacent strong and weak pulses in terms of their periods and amplitudes in the time domain, and then investigate the glottal characteristics of PD including the abovementioned parameters obtained from EGG: contact quotient (CQ), speed quotient (SQ), as well as peak increase in contact (PIC). In the frequency domain, I present and discuss the spectral differences of samples of representing different types of period doubling.

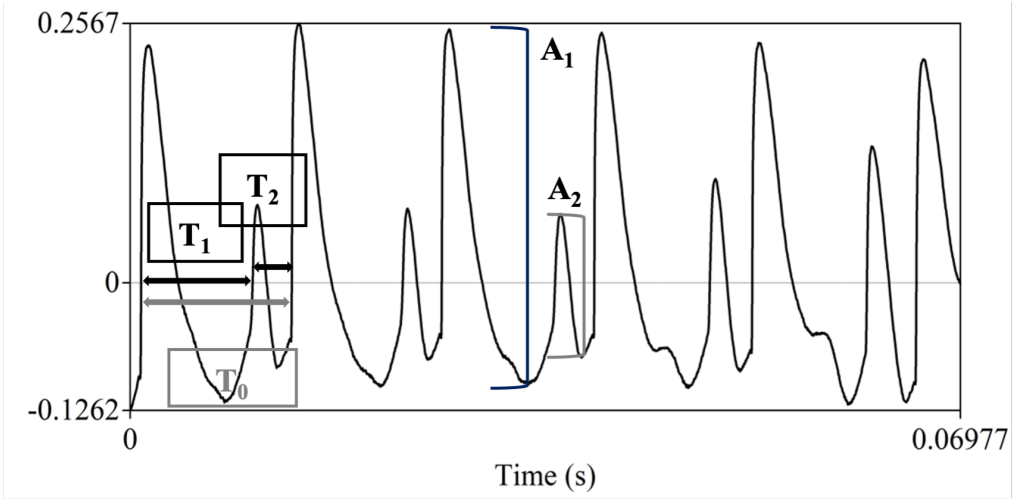


(a)

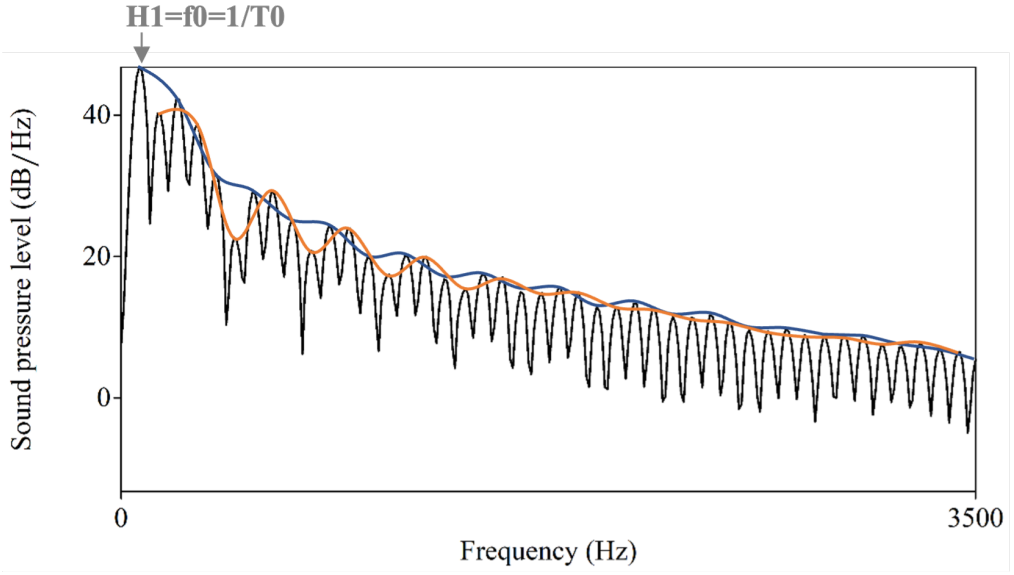


(b)

Figure 2.1: EGG Waveform (a) and spectrum (b) of amplitude-modulated period doubling. In the waveform, short black lines and arrows represent the smaller periods T_1 and T_2 and longer grey lines and arrows represent the fundamental period T_0. The black and grey brackets show the overall displacement of the larger pulse A_1 and the smaller pulse A_2 respectively. The spectrum shows some alternation between odd and even harmonics in terms of their amplitudes, delineated by the blue (odd) and orange (even) lines: e.g., $H1 < H2$, $H3 < H4$, but $H2 > H3$. The f_0 is $97Hz$, corresponding to the T_0 on the waveform; the mean $f_1 = 1/T_1 = 183Hz$, the mean $f_2 = 1/T_2 = 207Hz$. The mean frequency ratio $R_T = f_2/f_1 = 1.13$. The mean amplitude ratio $A_1/A_2 = 1.4$.



(a)



(b)

Figure 2.2: EGG waveform (a) and spectrum (b) of amplitude- and frequency-modulated period doubling. Note that there is a larger difference between the duration of the two shorter periods T_1 and T_2 (represented by the black lines) than the example in Figure 2.1. The black and grey brackets show the overall displacement of the larger pulse A_1 and the smaller pulse A_2 respectively. The source spectrum shows some amplitude dips for some harmonics, odd ones delineated by the blue line and even by the orange one: (e.g., $H2 < H1$, $H3$; $H6 < H5$, $H7$), instead of a logarithmic power spectrum, likely influenced by the two subcycles T_1 and T_2 on the waveform. The f0 is $77Hz$, corresponding to the T0 on the waveform; the mean $f_1 = 1/T_1 = 105Hz$, the mean $f_2 = 1/T_2 = 287Hz$. The mean frequency ratio $R_T = f_2/f_1 = 2.73$. The mean amplitude ratio $A_1/A_2 = 2$.

2.4. Results: Distribution and defining characteristics of PD

2.4.1. Distribution of period doubling in the corpus

First, I summarize the distribution of period-doubled cycles across speakers in Table 2.1. Here the unit of 'cycle' is defined as a cycle-of-cycles; that is, a 'meta-cycle' including a pair of strong and weak pulses, corresponding to the longest period T0 shown on Figures 2.1-2.2. Recall that period doubling was identified anywhere in 384 carrier phrases of 8-9 syllables and 96 filler sentences (every 4 test stimuli) of 4-5 syllables; together, there were 480 sentences with an estimate of ∼3500 syllables. Here, a token that is associated with period doubling found in the EGG signal could be any segments within a syllable.

Table 2.1: The distribution of period-doubled tokens across speakers.

ID	Tokens of PD identified	Mean duration (ms)	Mean # cycles	Min # cycles	Max # cycles
F04	229	29.54	3.09	1	16
F09	731	21.98	2.41	1	12
F12	255	29.61	2.88	1	21
F22	838	37.79	3.02	1	21
F30	812	31.46	3.16	1	20
F31	278	30.04	3.16	1	16
F33	617	31.32	2.77	1	19
F37	817	28.50	2.70	1	18
F47	172	27.76	2.78	1	10
F48	360	23.72	2.28	1	15
M05	60	34.75	1.65	1	6
M15	163	42.27	1.76	1	6
M17	31	42.02	1.61	1	8
M35	417	37.15	2.01	1	11
M36	221	36.86	1.82	1	7
M38	26	32.74	1.73	1	4
M39	23	54.44	2.30	1	5
M40	263	56.87	3.21	1	11
M42	145	35.08	1.61	1	5
M46	36	32.71	1.47	1	3

A t-test comparing the number of period-doubled utterances found in women and men shows that women significantly had more occurrences of period doubling than men with a mean of 510.9 tokens versus 138.5 tokens ($t = 3.85, p < .01$), consistent with findings by Hedelin and Huber (1990) for Swedish and Yu (2010) for Mandarin and Cantonese. Though, individual variation was seen such that women with the lowest number of period-doubled tokens (F47: 172, F04: 229) had fewer tokens than men with the highest number (M40: 263, M35: 417). It may be explained by the higher vocal frequency of women than men, so that there is a wider f0 range to allow for more room to realize multiple different frequencies during period doubling, given that the lowest frequency derived from the longest period would be half of the original f0. In addition, as Yu (2010) mentioned that compared to vocal fry, period doubling is not contingent on any particular f0, the higher vocal ranges of women may facilitate the production of such voice quality. Women also had longer sustained voice samples of period doubling than men for an average of 2.82 vs. 1.92 cycles per utterance ($t = 4.82, p < .001$), and a maximum of 21 cycles. Within an utterance of a certain duration, there are fewer cycles if the frequency is lower; thus, this finding is expected, given that male voices have lower fundamental frequencies and/or longer periods.

2.4.2. Defining characteristics of period doubling: frequency and amplitude ratios

Having quantified the relationship between the relative time courses and amplitudes of the two alternating cycles using the formulas defined in Section 2.3.4, here I report results of the frequency and amplitude ratios. I start with the frequency ratio R_T between the higher f_2 and lower f_1 glottal frequencies, given by T_1/T_2, the longer period over the shorter one. Note that I chose to start the analysis with the stronger and longer cycle T_1, which is an arbitrary choice. To preclude potentially incorrectly identified period doubling from normal jitter found in modal voice, I calculated the pulse-to-pulse variation in duration based on the adjacent samples of modal voice and removed the period-doubled cycles whose values fall within the jitter range for modal ($0.997 \sim 1.02$), which

corresponded to 0.5% of the data. Data from the 20 speakers showed that the mean R_T ranged from 1.26 to 2.05, such that the weaker pulses occur later in the course of articulation, after the midpoint of the longest period T0, shown in Table 2.2. I also included the mean ratio of T_1/T_0, to directly compare the proportion of the first glottal cycle relative to the entire meta-cycle, which is bound at 1.

Table 2.2: Mean frequencies of the fundamental cycles and two glottal sub-cycles in period doubling, the frequency ratio R_T, and the mean (SD) ratio between the first glottal period and the fundamental period in each subject.

ID	Mean f0 (Hz)	Mean f_1 (Hz)	Mean f_2 (Hz)	Mean R_T (SD)	Mean T_1/T_0 (SD)
F04	104.46	173.86	261.68	1.56 (0.41)	0.6 (0.06)
F09	109.62	194.81	250.67	1.3 (0.21)	0.56 (0.04)
F12	97.92	163.19	244.8	1.54 (0.45)	0.6 (0.06)
F22	78.81	118.81	234.11	2.05 (0.84)	0.65 (0.08)
F30	100.9	161.65	268.46	1.71 (0.52)	0.62 (0.06)
F31	106.33	184.28	251.38	1.39 (0.27)	0.58 (0.04)
F33	89.48	161.19	201.14	1.26 (0.25)	0.55 (0.04)
F37	95.49	151.41	258.55	1.76 (0.7)	0.62 (0.08)
F47	100.68	166.4	254.92	1.6 (0.42)	0.61 (0.06)
F48	98.36	170.28	232.91	1.39 (0.38)	0.57 (0.05)
M05	46.9	82.68	108.36	1.33 (0.22)	0.57 (0.04)
M15	41.68	63.17	122.52	1.97 (0.43)	0.66 (0.05)
M17	38.54	64.31	96.19	1.51 (0.3)	0.6 (0.04)
M35	54.43	92.08	133.14	1.47 (0.3)	0.59 (0.04)
M36	49.11	81.64	123.25	1.53 (0.35)	0.6 (0.05)
M38	52.85	91.73	124.67	1.39 (0.24)	0.58 (0.04)
M39	42.33	68.59	110.55	1.63 (0.22)	0.62 (0.03)
M40	56.21	98.25	131.38	1.36 (0.21)	0.57 (0.04)
M42	45.84	75.95	115.59	1.53 (0.21)	0.6 (0.03)
M46	43.61	76.87	100.79	1.33 (0.29)	0.57 (0.05)

In this dataset the lowest frequency during period doubling, or f0, for women was approximately twice that for men. The smaller glottal frequency f_1 did not exceed $200Hz$ and the larger glottal frequency f_2 did not exceed $300Hz$. This is different from the findings by Schreibweiss-Merin and Terrio (1986), who found for volitional diplophonic voice from a female English speaker that the higher glottal frequency was always above $300Hz$. It is possible that

with a tone language, the underlying tone could restrict the multiple frequencies brought about by period doubling; tonal distributions are discussed in Chapter 3. A t-test indicates that women had a higher frequency ratio in period-doubled voice than men with a mean of 1.61 versus 1.49 ($t = 14.82, p < .001$).

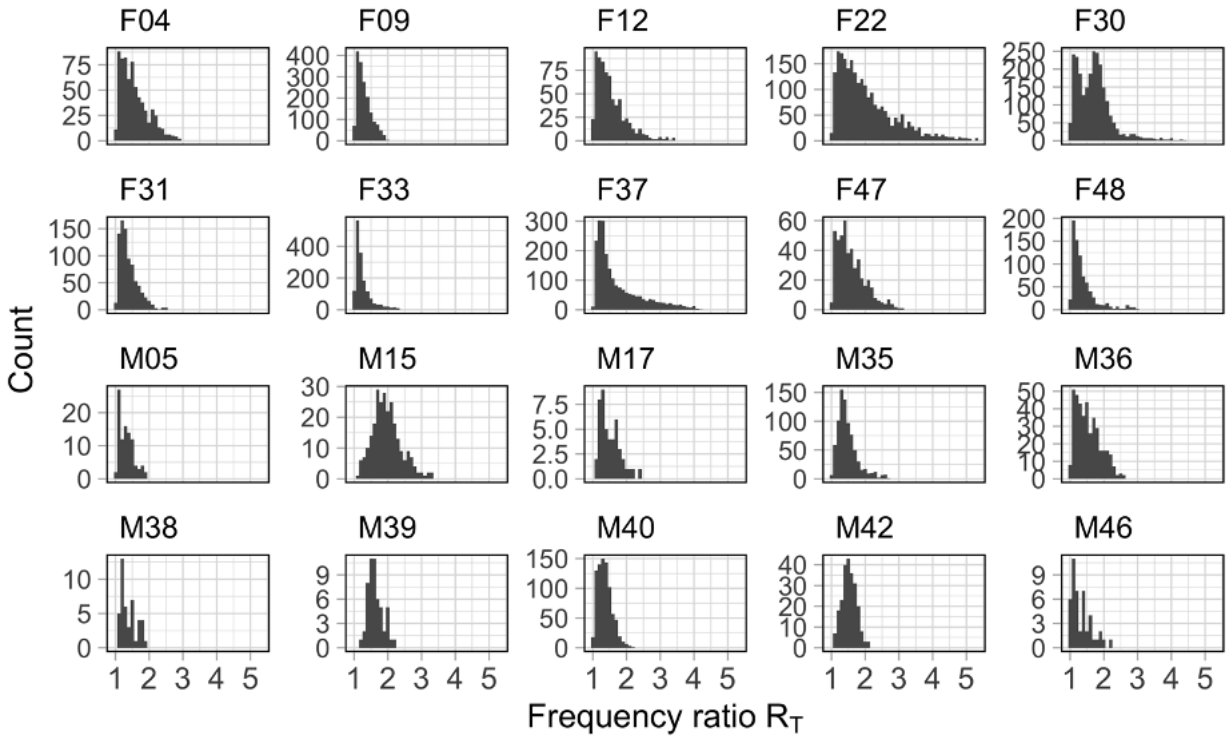


Figure 2.3: Distribution of frequency ratio R_T by speaker.

Figure 2.3 further shows the distribution of the frequency ratio R_T in period-doubled pulses in different subjects. Clearly women exhibited more instances of period doubling than men. Despite individual differences, women tended to have a right-tail distribution of period doubling which spans across higher values, whereas men tended to limit the frequency ratios up to 3. In general, as Table 2.2 shows, most subjects had a ratio of T_1/T_0 centered around 0.6, meaning that the duration of the two alternating periods has a relationship of 6:4 or 3:2. This is consistent with past findings on an $f0/2$ subharmonic vibratory pattern with 3:2 entrainment instead of the typical 1:1 ratio as in modal voice (Švec et al., 1996).

I then look further into the frequency ratio in the two different types of period doubling, amplitude modulation and combined amplitude plus frequency modulation. In period-doubled voices like Figure 2.1, the average frequency ratio R_T was 1.46 ($\sim 3/2$), meaning that the sub-pulse occurred closer to the midpoint of the fundamental period (meta-cycle); in samples like Figure 2.2, the average R_T was 1.92 ($\sim 2/1$), meaning the sub-pulse was closer to 2/3 of the fundamental period. The two alternating glottal cycles thus exhibit a ratio approximating 3:2 or 2:1, depending on the types of period doubling.

How does this frequency ratio R_T change with regards to its fundamental frequency? That is, when fundamental frequency lowers, does R_T vary accordingly? Figure 2.4 plots the relationship between the frequency ratio R_T and its fundamental frequency.

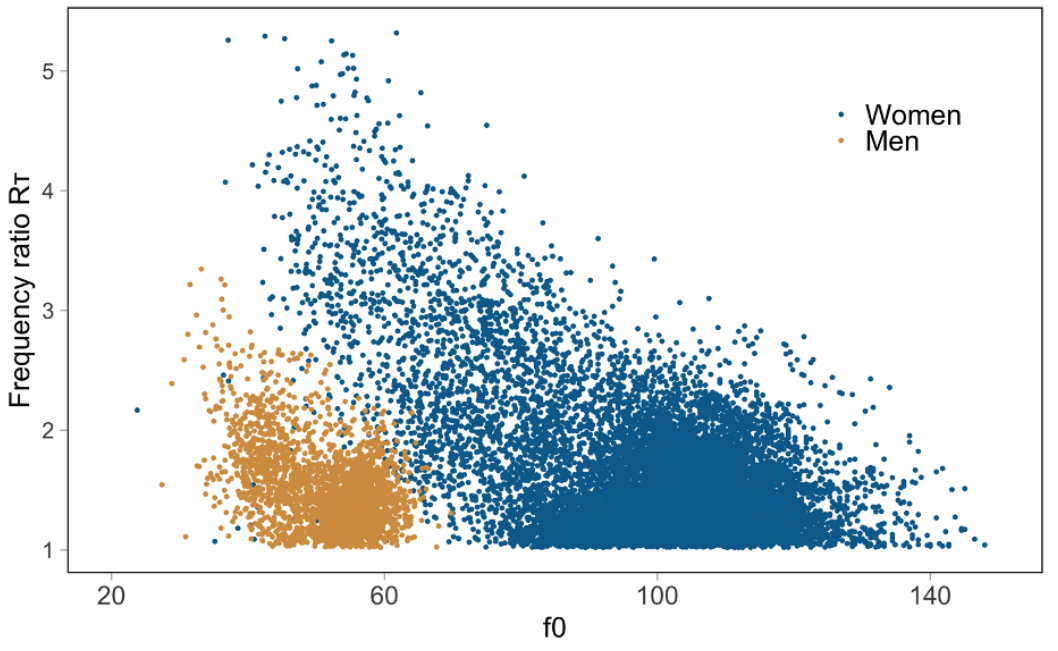


Figure 2.4: Scatter plot of the frequency ratio R_T (higher frequency/lower frequency) as a function of the fundamental frequency (determined by the longest period T_0) in period-doubled samples.

As f0 increases, the frequency ratio tends to decrease (Pearson's $r = -0.51$ for women, and -0.45 for men). Recall that $R_T = f_2/f_1$, meaning that as f0 increases, the lower glottal frequency f_1 also increases and at a more rapid rate than the increase of the higher frequency f_2.

It follows that, when period doubling occurs in voices with very low f0 (long periods), the first glottal sub-cycle tends to also be longer, and the second sub-cycle would occur even later during the articulation of the main cycle, resulting in a long-short-long alternation in the waveform, similar to the amplitude and frequency-modulated waveform in Figure 2.2. In fact, if we compare the two samples in Figures 2.1-2.2, the frequency ratio R_T in Figure 2.1 is smaller than that in Figure 2.2: its f0 is higher – corresponding to the negative correlation between the frequency ratio and the f0 during period doubling.

Then, I compare the mean f0 and the two glottal frequencies f_1 and f_2 in period doubling with the f0s in modal voice and vocal fry. Modal voice was sampled from the adjacent glottal cycles of period-doubled tokens. I extracted modal voice samples of ~ 50 *ms* from speech, which occurs in the same phrase as the instances of period-doubled voice for comparison. Vocal fry was identified by its defining characteristics such as having low f0, glottal constriction, and high damping (Keating et al., 2015). See Figure 2.5 for an example of vocal fry from subject F30 in the EGG waveform.

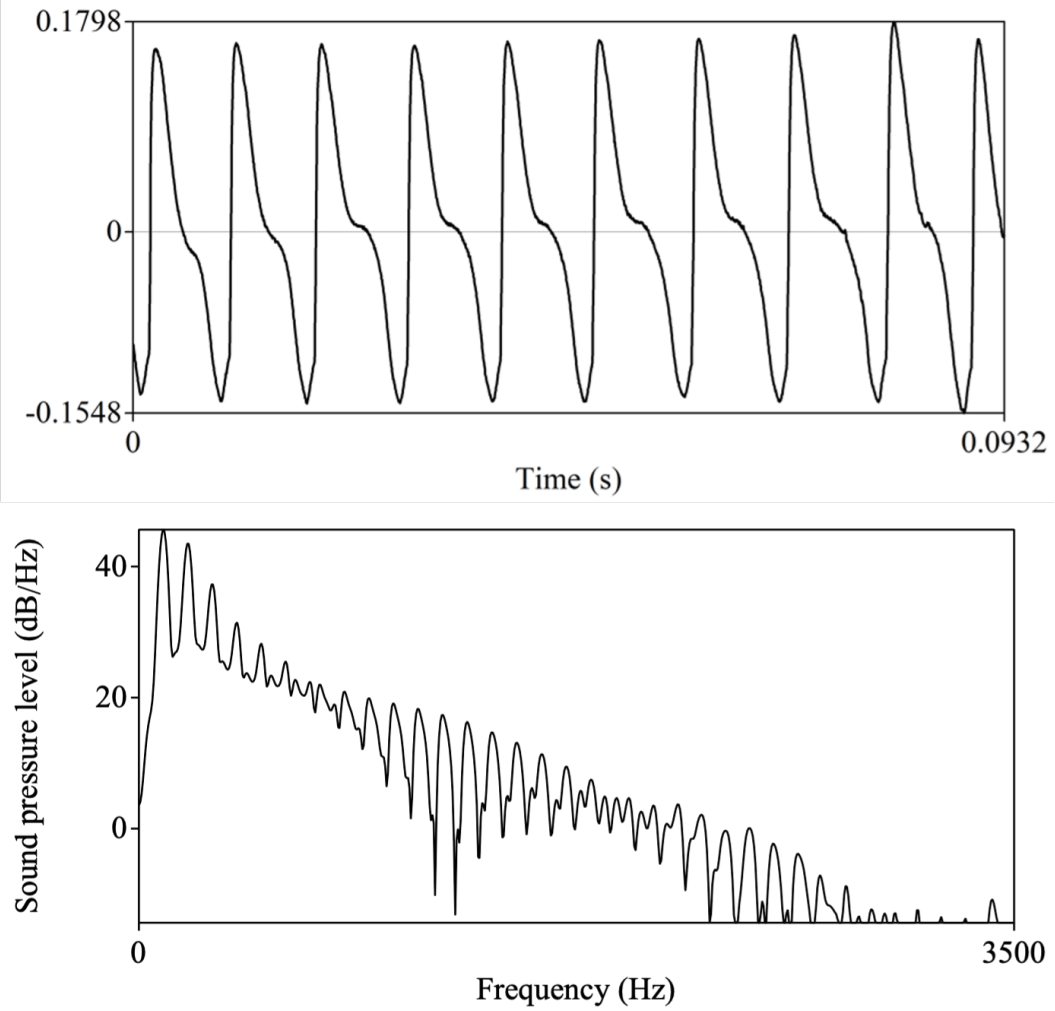
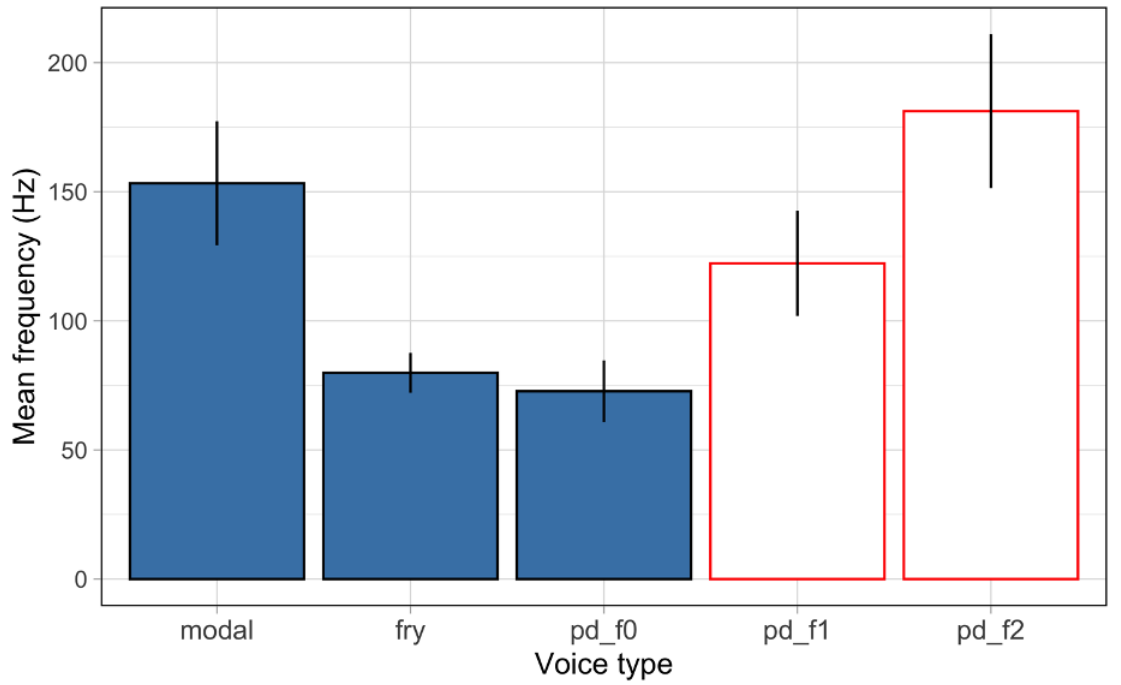


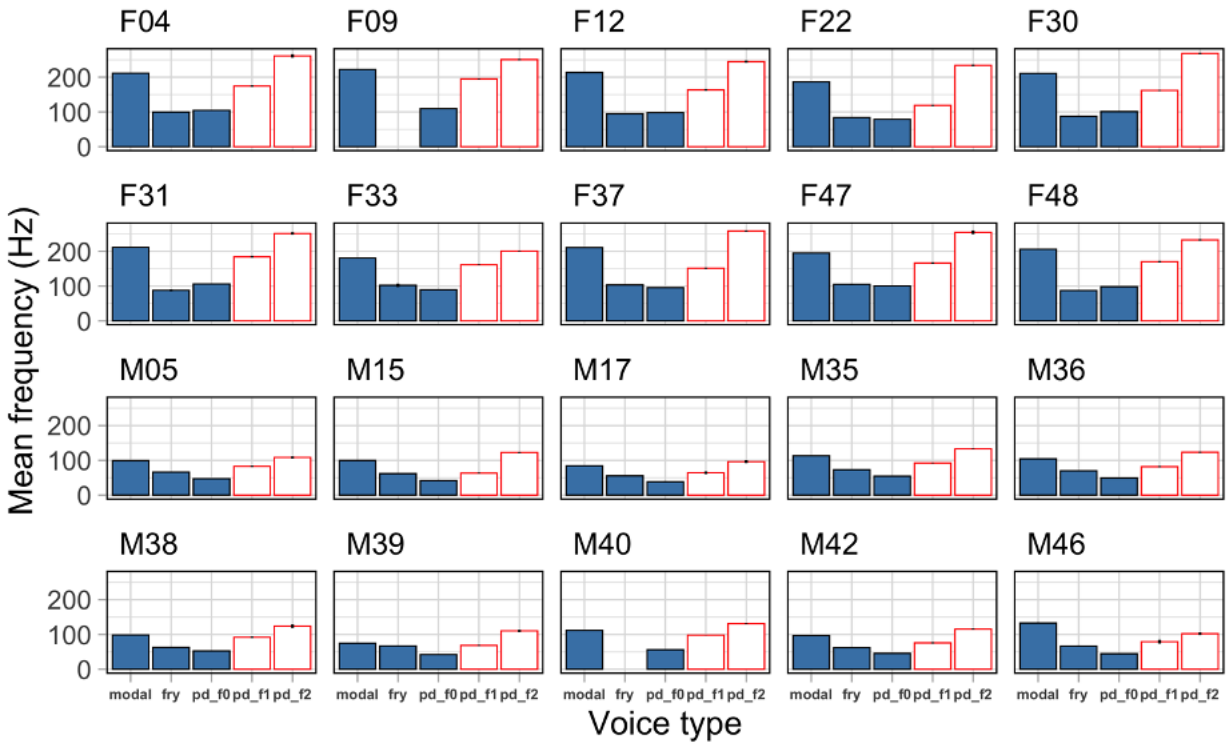
Figure 2.5: EGG waveform and spectrum of vocal fry, with mean CQ (hybrid) = 0.77. For each pulse, the contact phase is approximately equal to the portion of the cycle above the zero crossing.

The five frequencies are listed in Figure 2.6 by each speaker. The results revealed the same pattern across speakers. The presumptive f0 derived from the longest period T0 during period doubling was comparable to that of vocal fry; in period doubling, the f0 and the first glottal frequency f_1 $(= 1/T_1)$ were lower than the f0 during modal voice, which was in turn lower than the second glottal frequency f_2 $(= 1/T_2)$: $f0$ $(period\ doubling) \approx f0$ $(vocal\ fry) <$ f_1 $(period\ doubling) < f0$ $(modal) < f_2$ $(period\ doubling)$. The f0 in period doubling was almost half that of the f0 in modal voice, and the f0 in modal voice was nearly the mean of the two higher glottal frequencies.



(a)

Figure 2.6: Comparison of the f0 in vocal fry, modal voice and the three frequencies in period doubling. Line ranges on top of the bars show a 95% confidence interval around the mean. (F09 and M40 do not have tokens of vocal fry.) The overall pattern averaged across all speakers is shown in (a) and individual patterns are shown in (b) (the CI was sometimes invisible). The left three columns (in blue) stand for the three f0s in modal voice, vocal fry, and period doubling. The rightmost two columns (in white) stand for the two glottal frequencies during period doubling. The pattern agrees on $f0\ (pd) \approx f0\ (fry) < f_1 < f0\ (modal) < f_2$.



(b)

Figure 2.6: Comparison of the f0 in vocal fry, modal voice, and the three frequencies in period doubling. Line ranges on top of the bars show a 95% confidence interval around the mean. (F09 and M40 do not have tokens of vocal fry.) The overall pattern averaged across all speakers is shown in (a) and individual patterns are shown in (b) (the CI was sometimes invisible). The left three columns (in blue) stand for the three f0s in modal voice, vocal fry, and period doubling. The rightmost two columns (in white) stand for the two glottal frequencies during period doubling. The pattern agrees on $f0\ (pd) \approx f0\ (fry) < f_1 < f0\ (modal) < f_2$. (cont.)

Next, I discuss the results of amplitude ratios. I also calculated shimmer from the normal pulse-to-pulse variation in amplitude in modal voice, and excluded the period-doubled cycles whose amplitude ratios were within the modal shimmer range $(0.98 \sim 1.06)$, which were 4.4% of the data. The resulting average of the amplitude ratio was 2.12 (± 0.69). This implies that the amplitude of the weak pulse tended to be (less than) half of the amplitude of the stronger pulse. Table 2.3 shows the respective amplitudes and mean amplitude ratios within each subject. Substantial amount of variation was seen in the amplitude data, compared to the frequency data. It

is expected because very weak pulses were observed in the corpus and the unit (Pa) of the values are smaller than that of frequency (Hertz).

In addition, the signals from women were lower in amplitude, which is expected because EGG signals tend to be stronger in men due to their narrower angles formed by the sides of the thyroid cartilage. However, it remains unclear whether both genders behave similarly when realizing the two alternating cycles using different amplitudes. A t-test indicates that women had a significantly higher amplitude ratio than men in period-doubled voice with a mean of 2.38 versus 1.77 ($t = 21.36, p < .001$).

Table 2.3: Mean amplitudes of the two glottal cycles and their ratio R_A of each subject.

ID	Mean A_1 (Pa)	Mean A_2 (Pa)	Mean R_A	SD R_A
F04	0.04	0.08	2.62	2.21
F09	0.12	0.21	2.24	1.64
F12	0.19	0.28	1.63	0.68
F22	0.24	0.35	1.61	0.79
F30	0.13	0.30	4.33	3.63
F31	0.09	0.14	2.33	1.95
F33	0.50	0.73	1.58	0.56
F37	0.34	0.57	1.92	1.03
F47	0.26	0.43	2.36	1.80
F48	0.16	0.28	2.06	1.00
M05	0.58	0.98	3.16	3.63
M15	0.35	0.47	1.45	0.50
M17	0.49	0.79	2.22	1.40
M35	0.90	1.16	1.34	0.28
M36	0.30	0.47	1.96	1.04
M38	0.29	0.50	2.36	1.48
M39	0.19	0.33	2.20	1.17
M40	0.22	0.33	1.90	0.92
M42	0.37	0.59	1.84	0.66
M46	0.35	0.45	1.35	0.28

Figure 2.7 shows the distribution of amplitude ratios within each speaker. Though individual distributions differ in their range of values such that women spanned a larger range than men, they did not differ much in the right-tailed shape.

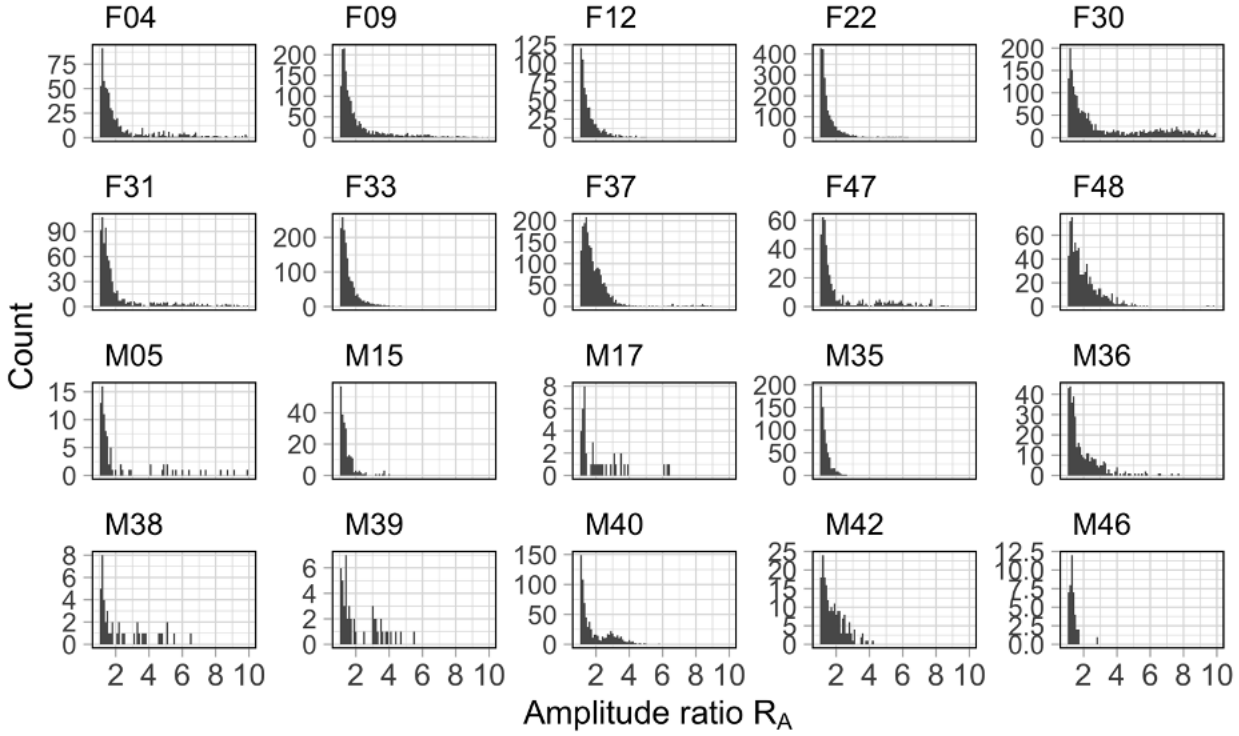


Figure 2.7: Distribution of amplitude ratio R_A by speaker.

The amplitude ratio R_A also varied according to the type of period doubling, depending on whether it was amplitude modulated or amplitude and frequency modulated: results indicate that the difference in amplitude was larger in amplitude modulated tokens. Table 2.4 summarizes the descriptive statistics of frequency and amplitude ratios. In general, R_A exhibits more variability by having a higher SD, especially in amplitude-modulated tokens and in women. The differences for both ratios between two types of period doubling, and between women and men were significant, as shown by two sample t-tests.

Table 2.4: Mean (SD) of frequency and amplitude ratio pooled by speakers. 'AM' stands for amplitude-modulated and 'FM' stands for amplitude and frequency-modulated period-doubled tokens.

	Overall mean	AM	FM	Women	Men
R_T	1.53 (0.21)	1.46 (0.15)	1.92 (0.35)	1.56 (0.24)	1.50 (0.19)
R_A	2.12 (0.69)	2.17 (0.76)	1.70 (0.33)	2.27 (0.81)	1.98 (0.55)

2.5. Results: EGG waveform analysis of PD, vocal fry, and modal voice

EGG signal is well-known for approximating and demonstrating vocal fold contact and spreading phases within each vibratory cycle. In this section, I analyze the articulatory properties of period-doubled pulses as shown in EGG waveforms using glottal constriction measures: contact quotient (CQ), speed quotient (SQ), and peak increase in contact (PIC). An illustration of different methods of CQ calculation along with SQ and PIC is shown in Figure 2.8. Period doubling is compared with two other voice categories that are more established: vocal fry and modal voice. All the EGG measures were obtained from EGGWorks (Tehrani, 2009). There are 6548 tokens of period doubling, 1164 tokens of vocal fry, and 2024 tokens of modal voice used for comparison in the following sections.

2.5.1. Contact quotient measures comparison: PD, vocal fry, modal voice

This section focuses on different measures of contact quotient (CQ), as provided in EGGWorks (Tehrani, 2009). Contact quotient is defined as the ratio of the contacting duration to the length of the vibratory cycle. Higher values of contact quotient indicate more constriction during a pulse.

Creaky voice is often characterized by increased vocal fold constriction, as in canonical creak and vocal fry (Keating et al., 2015). Thus, we might expect period doubling, a subtype of creaky voice ('multiply-pulsed voice'), to be constricted as well, meaning that it should have higher CQ values than modal voice. To investigate the degree of constriction of period doubling, I compared the CQ of period doubling and that of vocal fry and modal voice found in the corpus.

With EGGWorks, contact quotient can be measured using four different algorithms, the mechanisms of these algorithms are also shown in Figure 2.8:

1) The derivative approach (CQd), which uses the peak velocity to determine the onset of contact and the minimum velocity for the offset of contact.

2) The threshold approach (CQt), which uses a preassigned amplitude percentage to determine the portion of the contact phase (Kania et al., 2004).

3) The hybrid method (CQh), which uses the derivative EGG to determine the onset of contact and the percentage threshold for the onset of opening (Howard, 1995).

4) The Henry Tehrani method (CQht), which uses the peak velocity to locate the onset of contact and the crossing of the DC component along the EGG contour for the end of the contact cycle (Tehrani, 2009).

I use all four algorithms to determine which might be most effective at distinguishing period doubling from modal voice and vocal fry. Figure 2.8 illustrates the CQ algorithms and other measures including speed quotient and peak increase in contact which will be discussed in the following subsections.

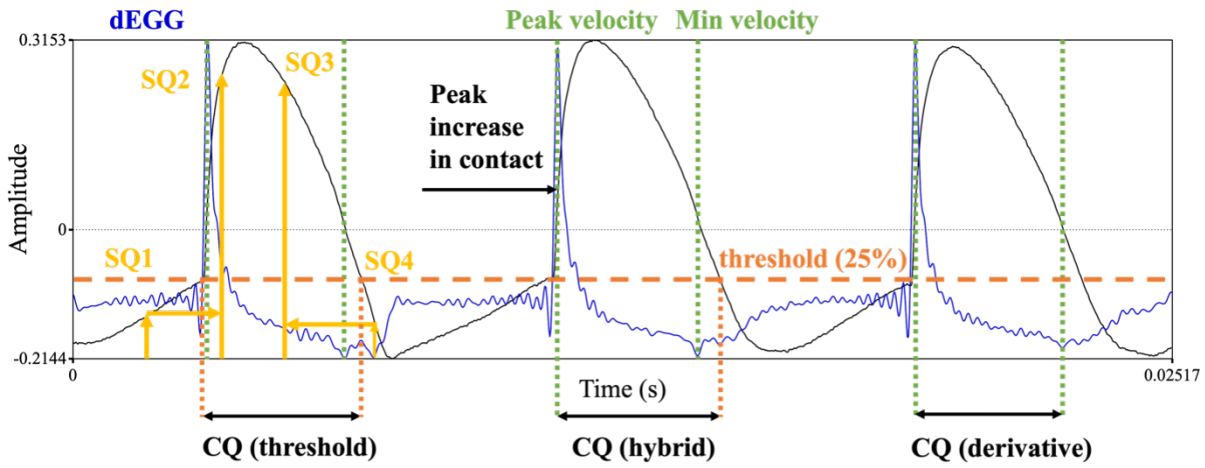


Figure 2.8: Diagram of EGG (black) overlaid by dEGG (blue line) illustrating CQ measures: speed quotient (SQ), and peak increase in contact (PIC), based on EGGWorks (Tehrani, 2009). CQd = duration between peak velocity and min velocity (green dotted line); CQt = duration between EGG crossing threshold line (e.g., 25%; orange dashed line); CQh = duration between peak velocity in the closing slope and EGG crossing threshold line in the opening slope; SQ = $(SQ4 - SQ3)/(SQ2 - SQ1)$ (SQ1: 10% above the min amplitude of closing slope; SQ2: 90% above the min amplitude of closing slope; SQ3: 90% above the min amplitude of opening slope; SQ4: 10% above the min amplitude of opening slope; yellow solid arrows); PIC = value of EGG at the peak velocity. CQht not shown, which uses the peak velocity in the closing slope as the beginning, and the crossing of the DC component along the EGG as the end in the opening slope.

Table 2.5 summarizes mean CQ during period-doubled voice for all four algorithms, and Figure 2.9 visualizes the correlation matrix among these different measures. I also excluded values that fall outside of 2.5 standard deviations from the mean.

Table 2.5: Summary of mean (SD) of different CQ measures during period doubling calculated based on four algorithms within each subject.

ID	CQ (derivative)	CQ (threshold)	CQ (hybrid)	CQ (Henry Tehrani)
F04	0.39 (0.16)	0.58 (0.09)	0.55 (0.12)	0.44 (0.29)
F09	0.34 (0.08)	0.54 (0.07)	0.52 (0.09)	0.56 (0.17)
F12	0.47 (0.13)	0.57 (0.09)	0.55 (0.11)	0.29 (0.32)
F22	0.31 (0.20)	0.62 (0.11)	0.61 (0.12)	0.28 (0.33)
F30	0.28 (0.12)	0.51 (0.13)	0.47 (0.13)	0.39 (0.21)
F31	0.35 (0.10)	0.55 (0.06)	0.53 (0.09)	0.53 (0.22)
F33	0.40 (0.12)	0.51 (0.07)	0.52 (0.08)	0.50 (0.22)
F37	0.35 (0.16)	0.60 (0.10)	0.58 (0.11)	0.45 (0.27)
F47	0.30 (0.11)	0.49 (0.10)	0.48 (0.11)	0.44 (0.25)
F48	0.29 (0.11)	0.50 (0.09)	0.51 (0.08)	0.49 (0.20)
M05	0.26 (0.14)	0.49 (0.13)	0.38 (0.10)	0.35 (0.18)
M15	0.44 (0.13)	0.47 (0.07)	0.42 (0.09)	0.07 (0.18)
M17	0.36 (0.12)	0.52 (0.12)	0.40 (0.10)	0.30 (0.20)
M35	0.29 (0.12)	0.48 (0.08)	0.47 (0.08)	0.34 (0.28)
M36	0.31 (0.13)	0.49 (0.11)	0.44 (0.10)	0.30 (0.24)
M38	0.34 (0.15)	0.55 (0.10)	0.44 (0.11)	0.34 (0.15)
M39	0.36 (0.18)	0.53 (0.11)	0.42 (0.11)	0.22 (0.21)
M40	0.37 (0.10)	0.51 (0.09)	0.46 (0.08)	0.36 (0.19)
M42	0.34 (0.16)	0.49 (0.15)	0.43 (0.14)	0.27 (0.22)
M46	0.39 (0.18)	0.55 (0.12)	0.47 (0.13)	0.32 (0.28)

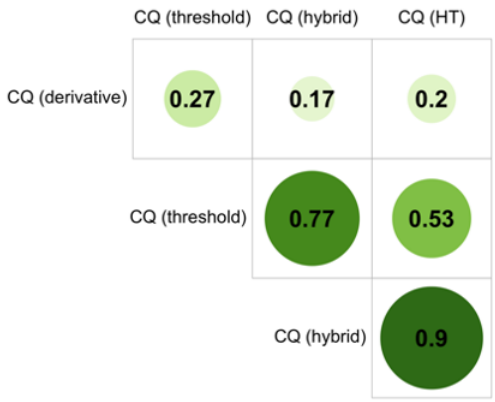


Figure 2.9: Pearson correlation coefficients among CQ measures during period doubling.

The hybrid method has stronger correlations with the threshold and Henry Tehrani methods while the derivative method shows larger difference from the rest.

I then review the CQ results in three types of voice in detail: period doubling, vocal fry, and modal voice. I start by looking at the derivative approach, abbreviated as CQd. Here, the distribution of CQ values during period doubling largely resembled that with modal voice, only with a wider and less peaked spread around the mean values. The mean CQd of period doubling was 0.34 ($\pm$0.15) and that of the modal voice was 0.41 ($\pm$0.11). Vocal fry had a bimodal distribution including modes of low ($\sim$0.125) and high CQ values ($\sim$0.79). The low-CQ pattern during fry typically resulted from pulses where the most negative peak in the derivative occurred early during the contact phase, instead of during the de-contacting phase; these CQ values are therefore spurious.

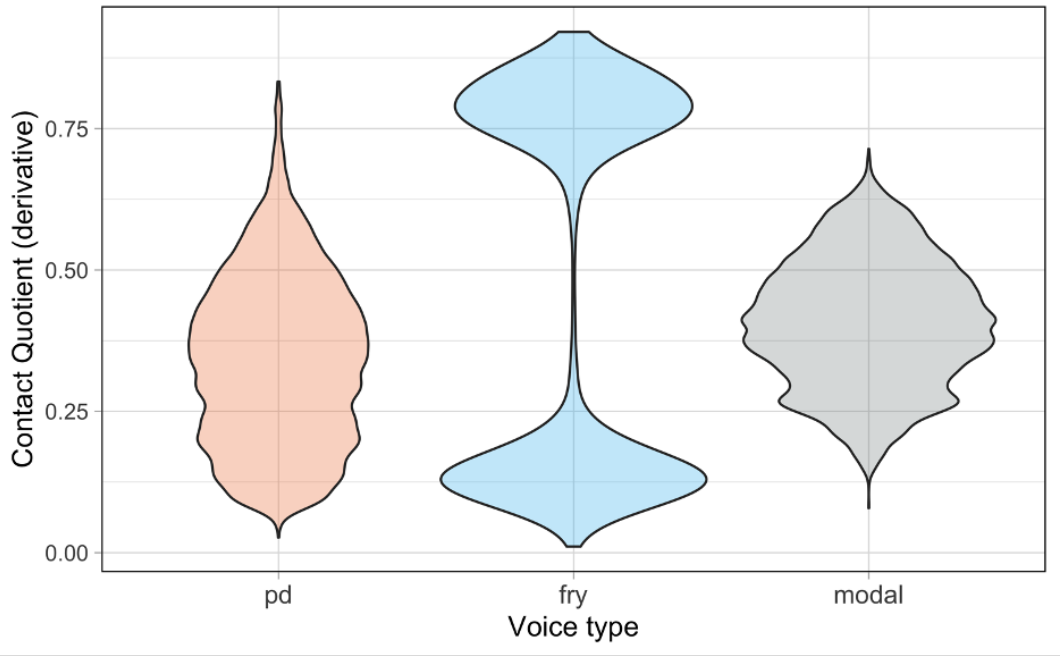


Figure 2.10: Distributions of Contact Quotient (derivative) in period doubling, vocal fry, and modal voice.

Next, I review the CQ results using the threshold approach at 25% percent, abbreviated as CQt. Here, vocal fry had a unimodal distribution centering around 0.74, suggesting a high degree of contact and that (for fry) this algorithm overlapped with the mode of the higher CQ values derived from the derivative algorithm to capture its constriction. The CQt distribution of period doubling was similar to modal voice, only with a larger spread. In fact, the means of CQt of period doubling and modal voice were 0.53 and 0.54, respectively.

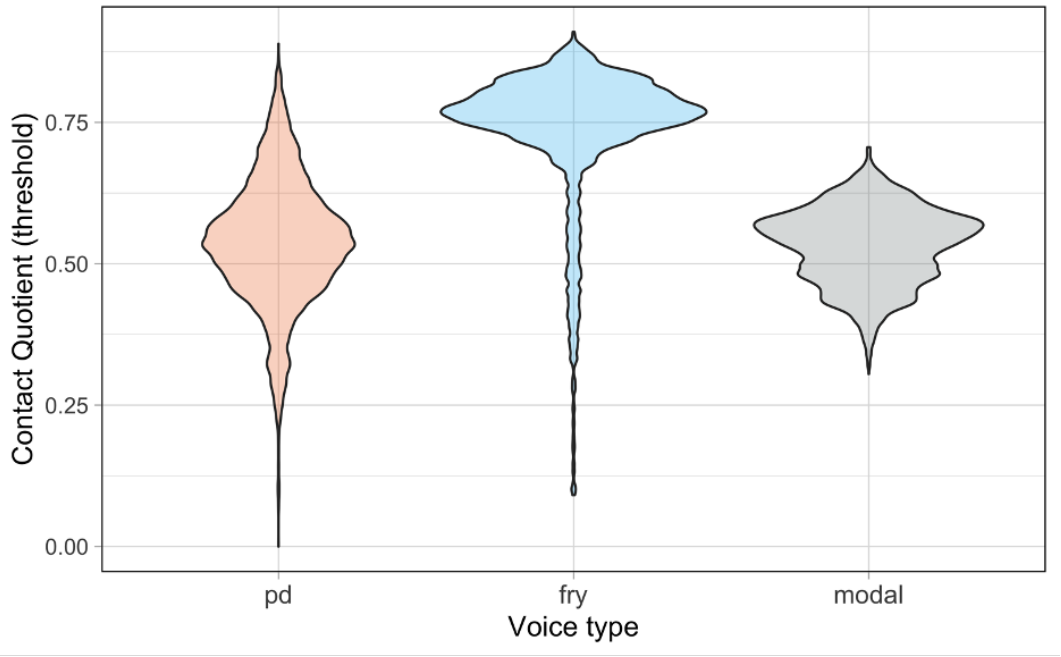


Figure 2.11: Distributions of Contact Quotient (threshold) in period doubling, vocal fry, and modal voice.

Next, I review the CQ results using the hybrid method combining the derivative and a 25% threshold, abbreviated as CQh. The patterns were highly similar to the results of CQt. Again, period doubling and modal voice had a similar distribution and vocal fry had a much higher CQ value. The bimodal distribution of modal voice was largely due to individual variation.

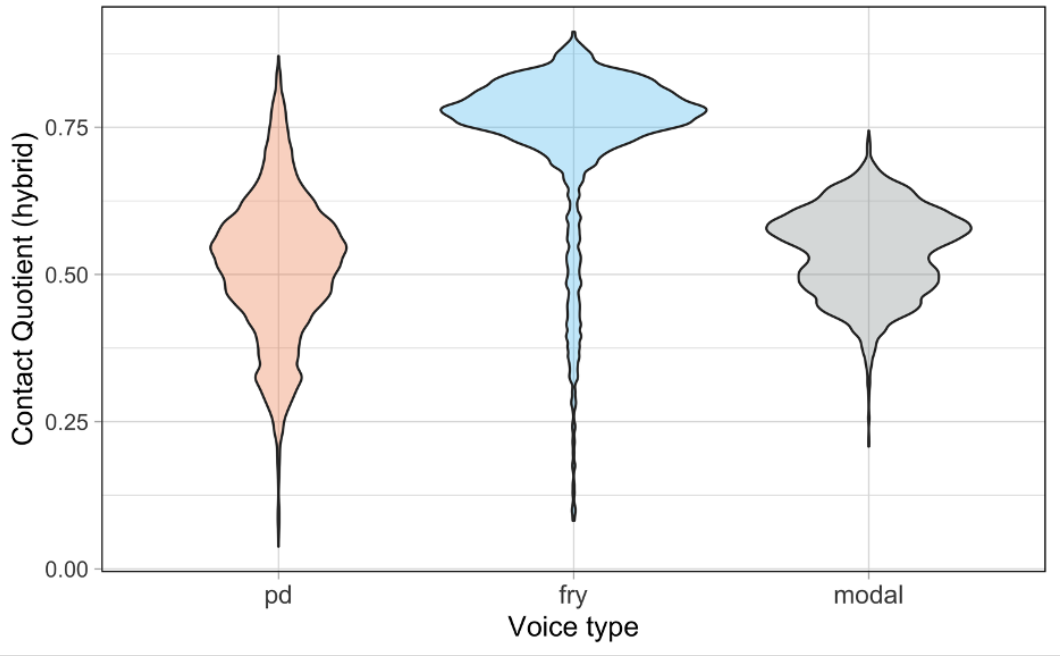


Figure 2.12: Distributions of Contact Quotient (hybrid) in period doubling, vocal fry, and modal voice.

Note that in CQ methods involving threshold determination, there were differences between the two genders (Figure 2.13), such that overall CQ values regardless of voice types were higher in women. Also, the pattern in women's data agreed with the overall pattern such that vocal fry had the highest CQ while period doubling and modal voice had a similar distribution, whereas in men's data, the mean CQ in three voice types were comparable. However, the reasons of the CQ gender differences are not clear.

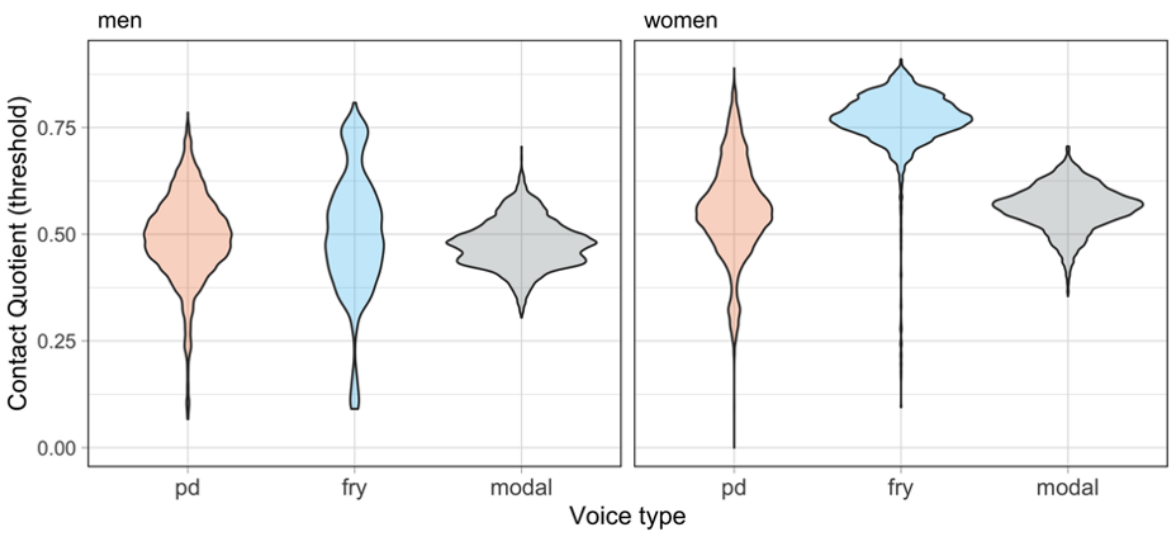


Figure 2.13: Distributions of Contact Quotient (threshold) in period doubling, vocal fry, and modal voice, faceted by gender.

Last, I review the CQ results using the Henry Tehrani method, abbreviated as CQht. Since CQht incorporated the peak increase in velocity to determine the onset of the contact cycle, it produced some lower CQ values in vocal fry around 0.25, similar to the case using a derivative approach in Figure 2.10. Here, period-doubled voice centered around the mean CQht of 0.54 ($\pm$0.14), lower than for modal voice at 0.63 ($\pm$0.11). Note that with this method, many zero values were produced in a few cycles within certain tokens. These data points were excluded, which corresponded to 21.1% of the data.

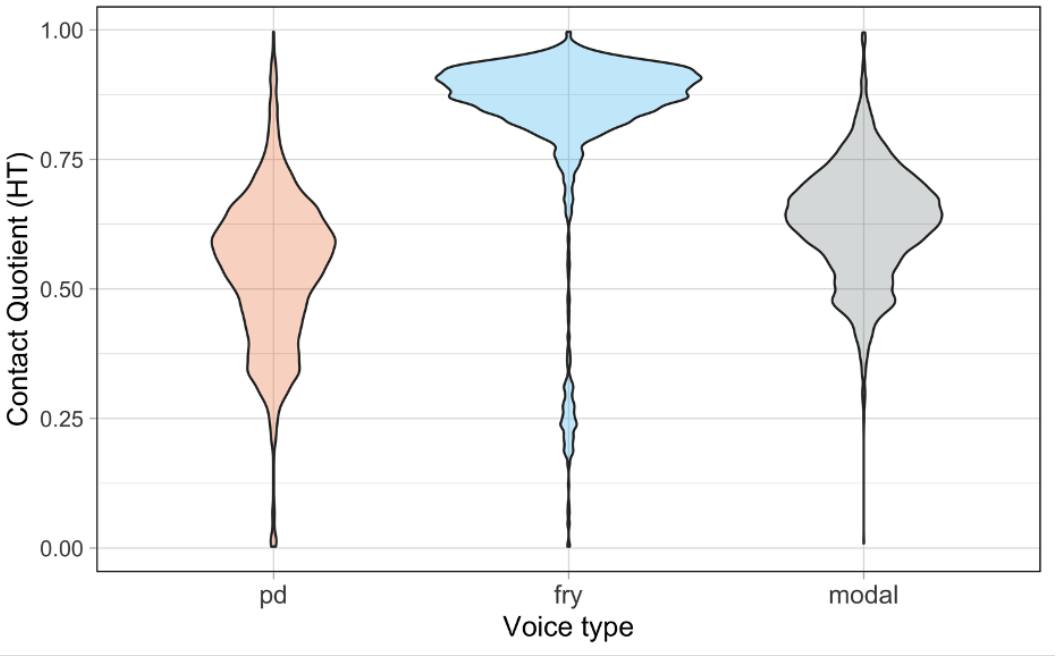


Figure 2.14: Distributions of Contact Quotient (Henry Tehrani method) in period doubling, vocal fry, and modal voice.

Table 2.6 summarizes the respective means (SD) of different CQ measures in these voice types. Linear mixed-effects model comparisons for each of the CQ measures were performed with and without the category of voice (period doubling vs. vocal fry vs. modal voice), with random intercept and slope by subject. For CQd, vocal fry did not differ from modal, but period doubling had a lower value than modal ($\beta = -0.10, p < .001$). For CQt, vocal fry had a higher value than modal ($\beta = 0.10, p < .01$), and period doubling was not different from modal. For CQh, vocal fry was higher than modal ($\beta = 0.08, p < .05$) whereas period doubling had a lower value ($\beta = -0.03, p < .01$). For CQht vocal fry had a higher value than modal ($\beta = 0.10, p < .05$), whereas period doubling had a lower value ($\beta = -0.08, p < .001$).

Table 2.6: Mean (SD) of CQ measures in PD, vocal fry, and modal voice.

Type	Period doubling	Vocal fry	Modal voice
CQ (derivative)	0.34 (0.15)	0.48 (0.33)	0.41 (0.11)
CQ (threshold)	0.54 (0.11)	0.74 (0.12)	0.53 (0.07)
CQ (hybrid)	0.51 (0.12)	0.74 (0.13)	0.54 (0.07)
CQ (HT)	0.54 (0.14)	0.84 (0.16)	0.63 (0.11)

In sum, vocal fry is distinct from both modal voice and period doubling in having a higher CQ across the four methods of CQ estimation. Period doubling and modal voice have similar CQ values, and period doubling has an even lower value in measures that incorporate the derivative in determining contact and open phases. This still holds even if we divide period-doubled tokens into two subtypes based on their typical characteristics, as in Figures 2.1 and 2.2. For example, the mean CQ (hybrid) of in amplitude-modulated tokens of period doubling was 0.49 ($\pm$0.1), showing more balanced contact and open phases; in amplitude and frequency-modulated tokens, the mean CQ (hybrid) of period doubling was 0.59 ($\pm$0.14), showing that the duration of vocal fold contact is slightly longer. This implies that in period doubling, the open and contact phases are more temporally balanced, and that the voice quality is not particularly constricted. In addition, Švec et al. (1996) found a subharmonic vibratory pattern, similar to the period-doubled voice in the current study, that the original glottal cycle splits into two smaller cycles where the open phase is as large as 0.8 (meaning a CQ $\sim$ 0.2), suggesting period doubling does not necessarily entail a high degree of vocal fold contact, but allows for more airflow during vibration. The fact that glottal constriction is not one of the defining characteristics of period doubling suggests that its features are distinct from what would typically be expected from other creak subtypes such as vocal fry and prototypical creaky voice, which are characteristically constricted based on the creaky voice subtypes framework (Keating et al., 2015).

2.5.2. Alternation in articulatory properties during PD: CQ

In Section 2.3.4, I showed some sample cases of period doubling. A typical characteristics of period doubling is that the glottal-pulse lengths and amplitudes are alternating. In Figure 2.2, for example, the alternation of glottal cycles is clearly seen in the time domain, and especially for amplitude and frequency-modulated tokens, not only the frequency and amplitude of every other cycle are different, but sometimes the pulse shapes are distinct. So far, I have been taking a holistic approach towards analyzing period doubling – quantifying the articulatory properties of the meta-cycle (the cycle of a pair of cycles) as the base unit. However, we ought to keep in mind that alternatively period doubling can also be analyzed by separating the two simultaneous periodicities which in particular contribute to the bitonal and rough percept. Along with the alternating periods, I further probe whether alternation is seen in other articulatory properties.

Here I start by investigating the contact quotient. It is predicted that CQ during period doubling would alternate in addition to the alternating amplitudes and frequencies, as the glottal sub-pulses alternate as different entities. The following approaches are taken, in comparison with vocal fry and modal voice:

1) plotting CQ values over a time course of selected samples of period-doubled voice which have a larger number of sustained pulses (at least 10 cycles) and larger shape differences in alternating pulses;

2) fast Fourier transform analysis of the cycle-by-cycle CQ values;

3) plotting separate density distributions of CQ values of every other cycle.

In the dataset of period doubling for example, about 30% of the data contain at least 10 cycles, including both amplitude and combined modulation. Here I particularly selected twenty period-doubled tokens with 10 cycles and a combined frequency and amplitude modulation shown as long-short-long alternation in periods, resembling the waveform in Figure 2.2, to illustrate the

alternating patterns. I also reviewed the results of other cases of period doubling with more prominent amplitude modulation, which showed that the alternations were generalizable.

Since the Henry Tehrani method produced many zeros in the calculation of CQ within a certain token, I only included CQ values obtained from derivative, threshold, and hybrid methods in this approach. Comparing Figures 2.15-2.17, most cases showed an interesting alternation in CQs during period doubling using all three measures, and CQd using the derivative method displayed more cases with alternations than the other two measures. This suggests that the two adjacent glottal cycles have distinct patterns of vocal fold contact and oscillate between two modes.

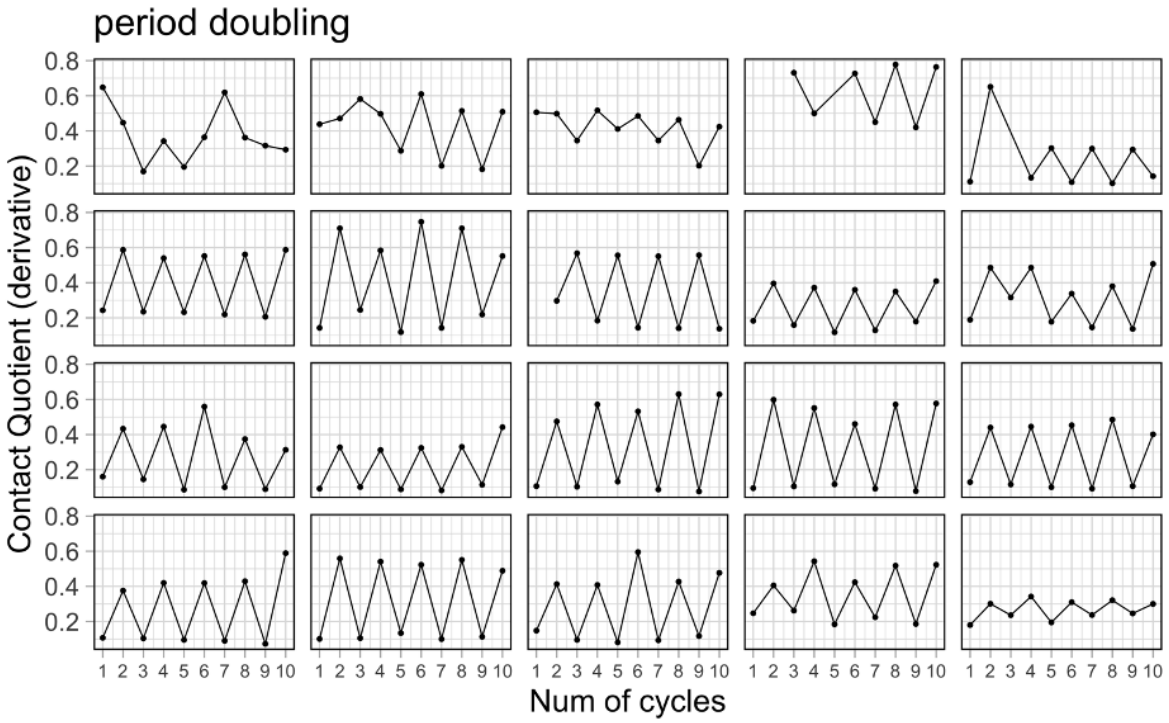


Figure 2.15: Alternation seen in CQ (derivative) during tokens of period doubling containing 10 cycles with larger differences in two glottal periods.

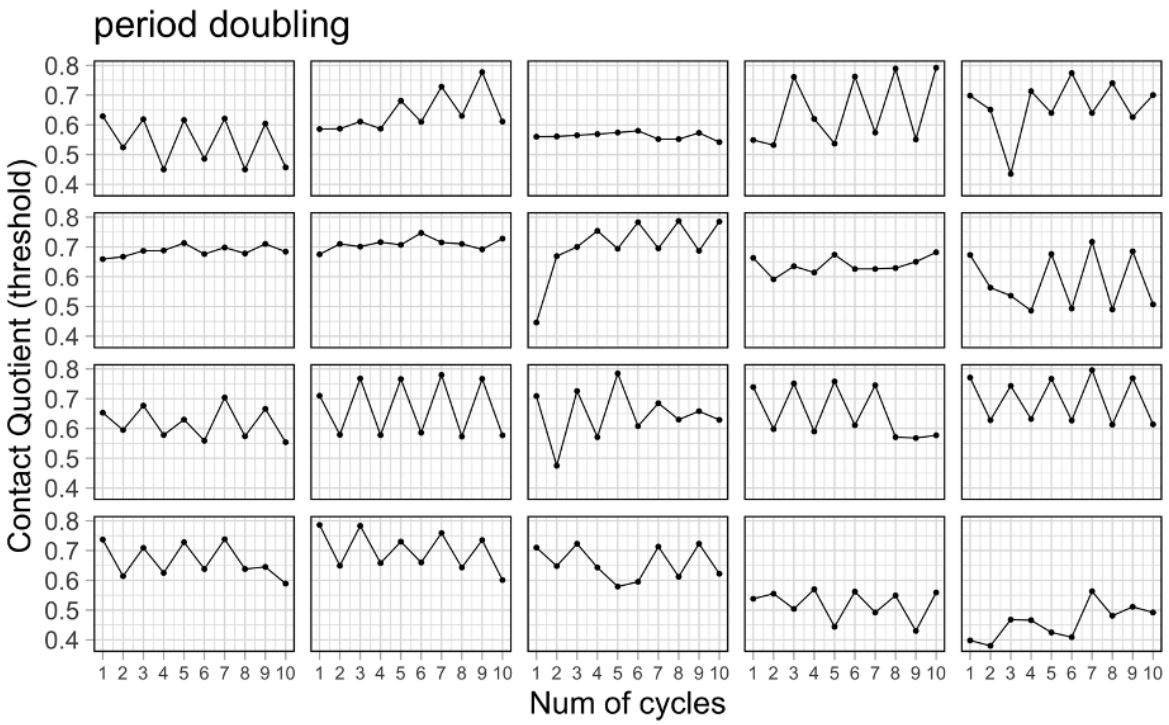


Figure 2.16: Alternation seen in Contact Quotient (threshold) during tokens of period doubling containing 10 cycles with larger differences in two glottal periods.

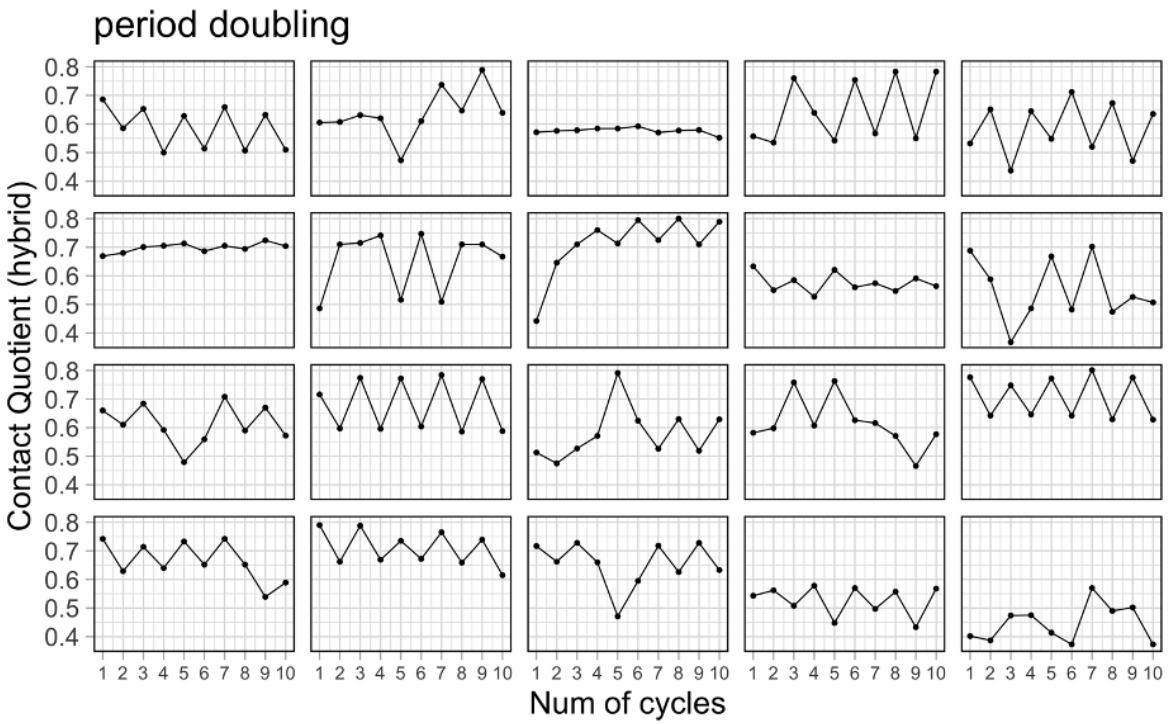


Figure 2.17: Alternation seen in Contact Quotient (hybrid) during tokens of period doubling containing 10 cycles with larger differences in two glottal periods.

Does modal voice or vocal fry have such or similar alternation? Figures 2.18 and 2.19 plot the CQ (derivative) values of modal voice and vocal fry, respectively, using samples consisting of sustained cycles.

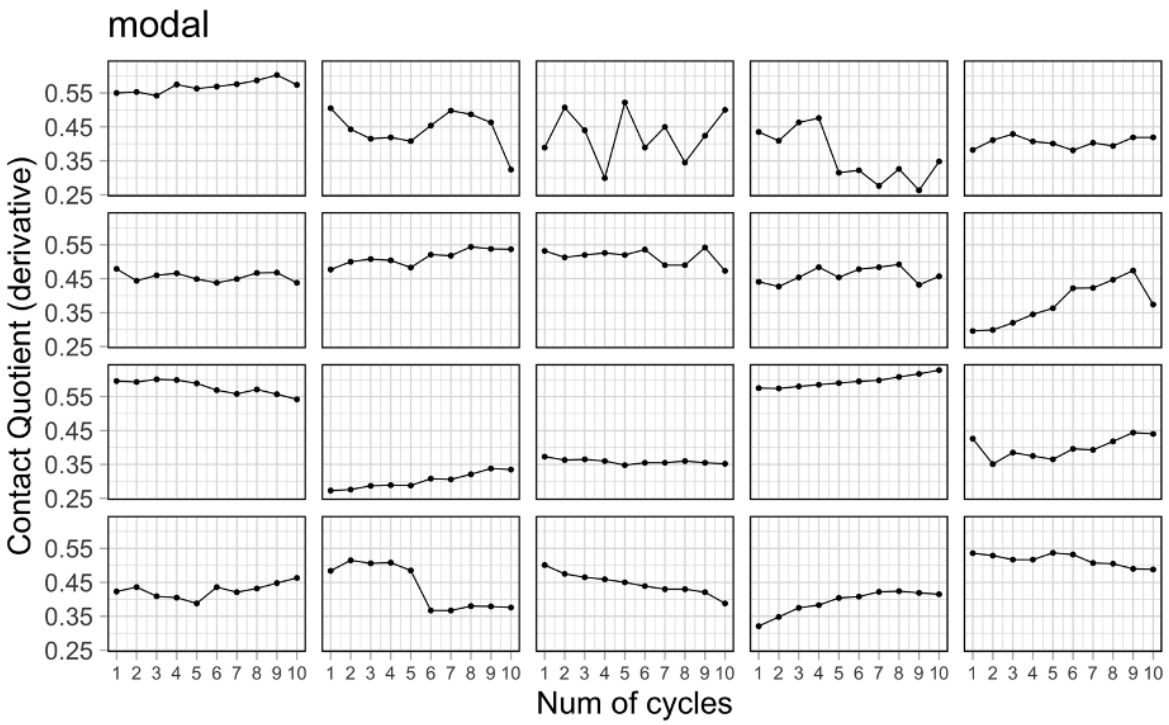


Figure 2.18: Contact Quotient (derivative) of each cycle during tokens of modal voice (10 cycles are plotted).

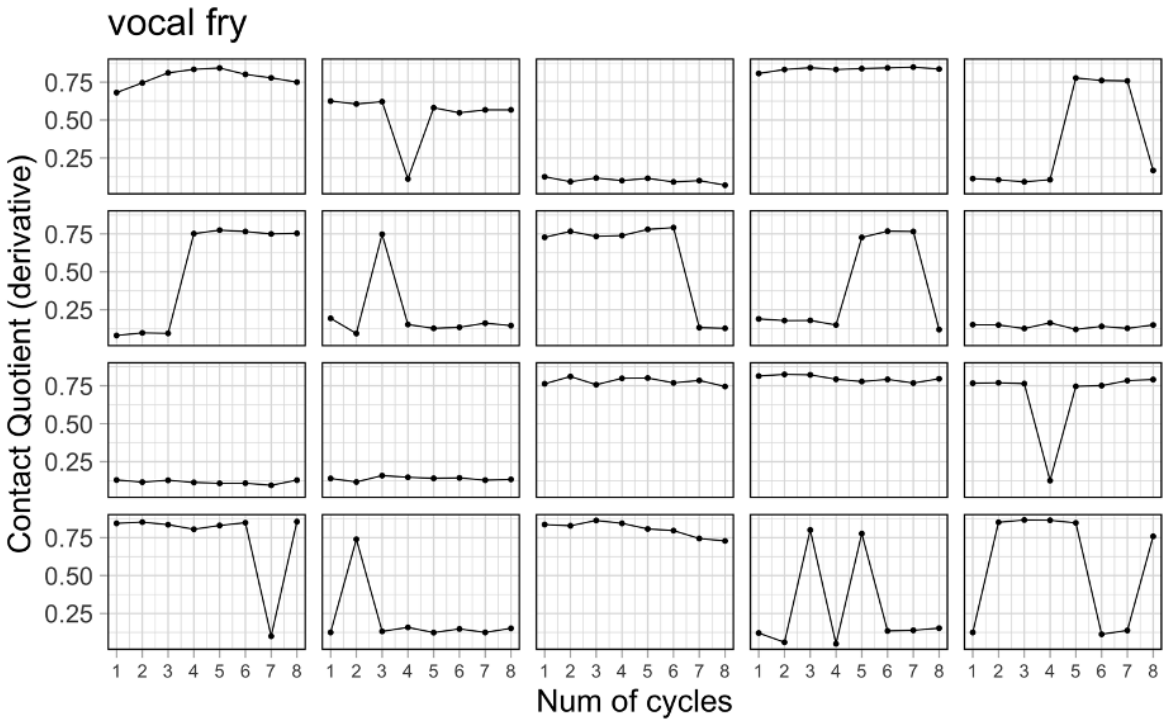


Figure 2.19: Contact Quotient (derivative) of each cycle during tokens of vocal fry (8 cycles are plotted; only a few samples have vocal fry of 10 cycles).

The alternation between CQ values in samples of modal voice and vocal fry is rather limited and incomplete as observed in only a few tokens out of 20 randomly chosen samples, compared to period doubling. Therefore, period doubling seems to be characterized by the alternation of contact portion in adjacent glottal cycles.

To further quantify the regularity of the alternations of CQ in period doubling, as opposed to the relatively irregular pattern in vocal fry and modal voice, I used fast Fourier transform (FFT) on the values of each cycle in each token. Specifically, I compared the median Fourier coefficients of components after FFT: if coefficient values are larger than zero, they indicate the presence of cyclicity in the measures. When the median Fourier coefficients are zero, no prominent spectral frequency component is found, suggesting that the original signal – in this case, the measure per cycle, is not periodic. I selected tokens with at least eight cycles to ensure the validity of cycle-by-cycle regular alternation. Prior to the FFT analysis, all the CQ values from a particular

calculation method were cleaned by removing outliers larger than 2.5 standard deviations from the mean; especially for the HT method, I also excluded zero values. Figure 2.20 plots the distributions of the median coefficients of the spectra of CQ values using the different methods. In all but derivative methods (CQd and CQht which uses the derivative to determine the onset of contact), period doubling had more non-zero median coefficients than vocal fry and modal voice.

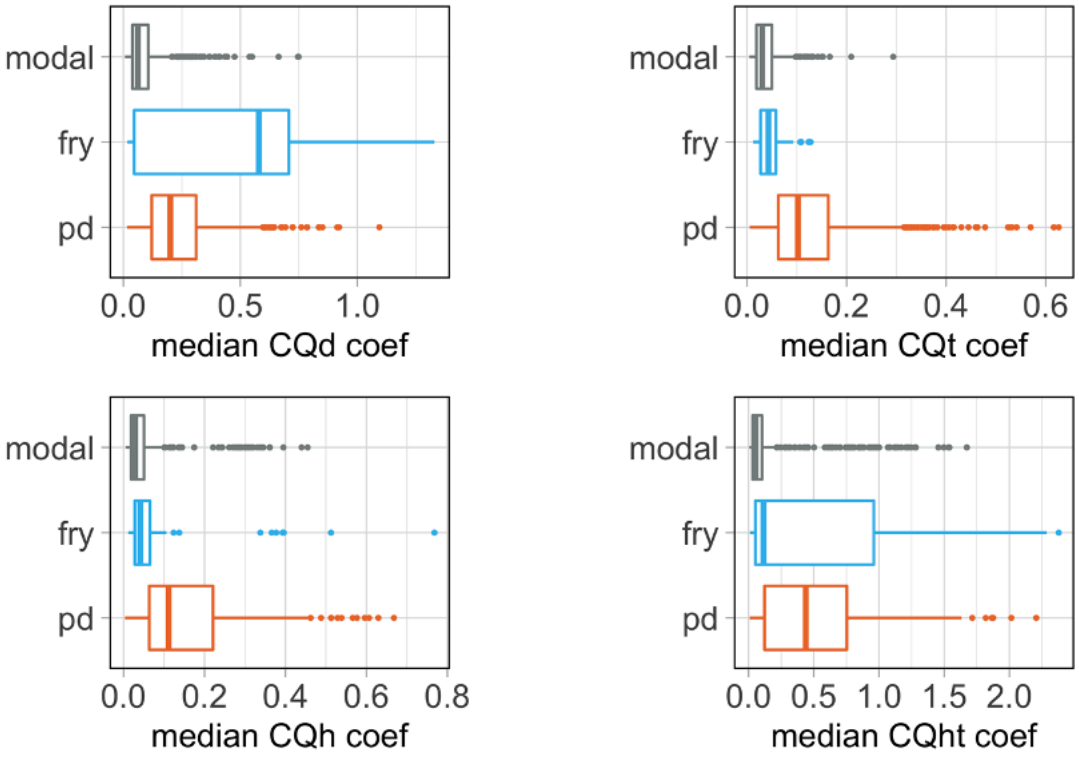


Figure 2.20: Distributions of median coefficients of the FFT of Contact Quotient (derivative, threshold, hybrid, and HT) in modal voice, vocal fry, and period doubling.

Recall that the CQ values using the derivative measure showed more cases of alternation than other CQ measures in period doubling. The pattern shown in Figure 2.20 that the median coefficients of spectral component of CQd (and CQht) in vocal fry exhibit substantial variation, suggesting that the measures using derivatives in vocal fry may be "noisy" and bimodal in that they have both low and high frequency components in units of cycles, and the distinction between period doubling and vocal fry could become blurred.

Lastly, as motivated at the beginning of this section, to investigate the alternating CQs between sub-cycles, I can plot the separate distributions of every other cycle accordingly. Rather than plotting the distribution of CQ in a meta-cycle (Section 2.5.1), I selected samples of three types of voice which contain at least four cycles, and divided every other cycle across all tokens into two distributions. Figure 2.21 plots the separate density distributions of the four CQ measures in period doubling, vocal fry, and modal voice.

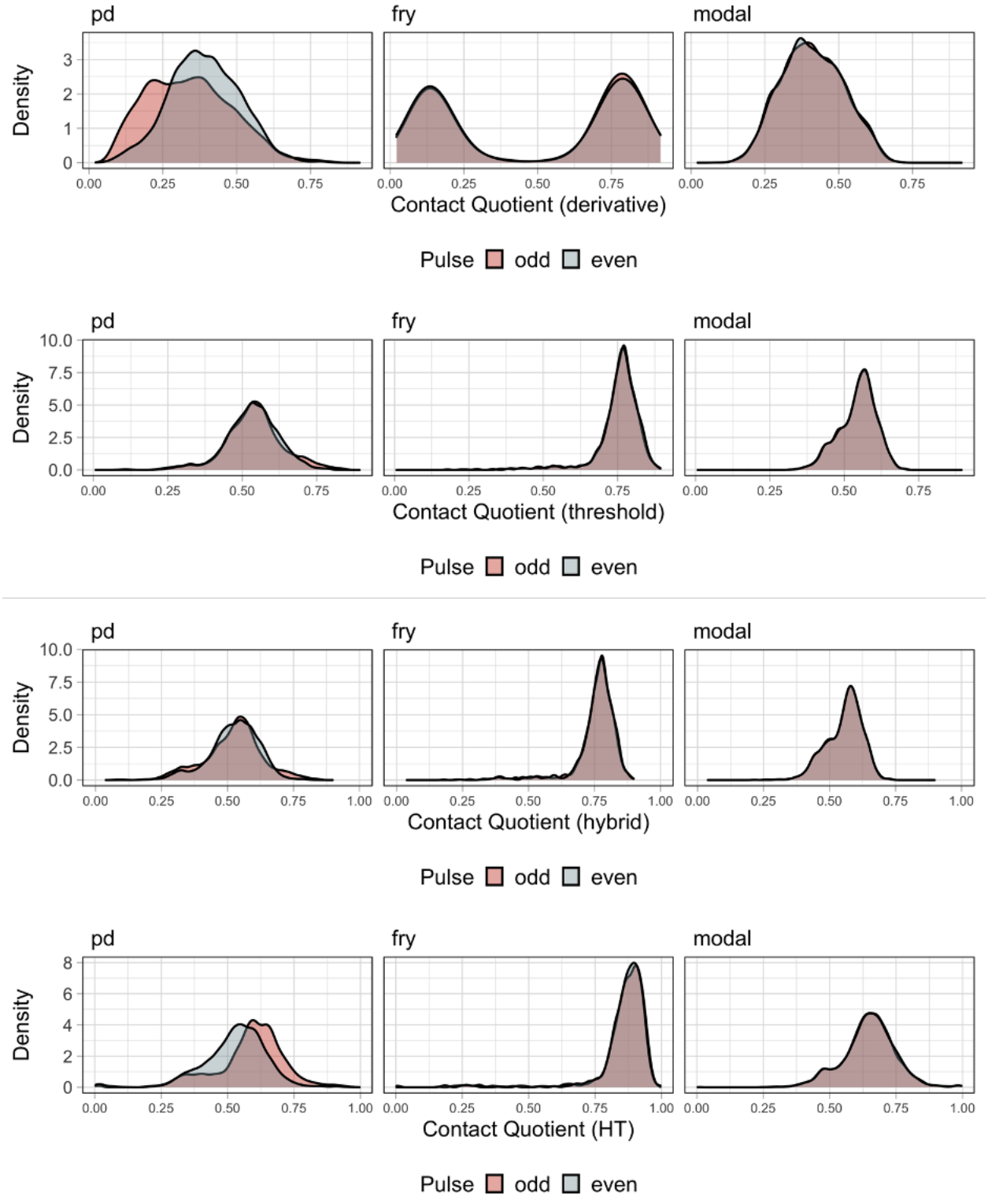


Figure 2.21: Density distributions of CQ measures of every other pulse during period doubling, vocal fry, and modal voice.

The figures further confirm that the alternation in CQs is a prominent feature during period doubling whereas it is rarely observed in vocal fry and modal voice. However, note that with threshold derived approaches (CQt and CQh), the two distributions of every other pulse had substantial overlaps during period doubling, though not a complete overlap as in vocal fry and modal voice. The differences may lie in the particular CQ calculation method, which complements the findings based on the median FFT coefficients. In general, the adjacent glottal sub-pulses in period-doubled voice can alternate between two vibration modes with different degrees of glottal constriction: more versus less constricted.

2.5.3. Speed quotient comparison: PD, vocal fry, modal voice

This section focuses on speed quotient (SQ). Whereas contact quotient indicates the proportion of the contact phase in vocal fold vibration, speed quotient (SQ) indicates the potential differences in the velocity of the contacting and the de-contacting phases. Speed quotient is defined as the fraction of the closing slope over that of the opening slope, and therefore is used to describe the symmetry of an EGG pulse shape. A lower value represents that the contacting phase is shorter than the de-contacting phase, which is common during increased vocal fold constriction. Because speed quotient followed a log-normal distribution with right-skewed long tails, the data were log-transformed for visualization and statistical tests.

Figure 2.22 shows the distributions of log-transformed SQ in period doubling, vocal fry, and modal voice, and Table 2.7 shows the respective mean (SD) SQ (in original units) in these voice types for better interpretation.

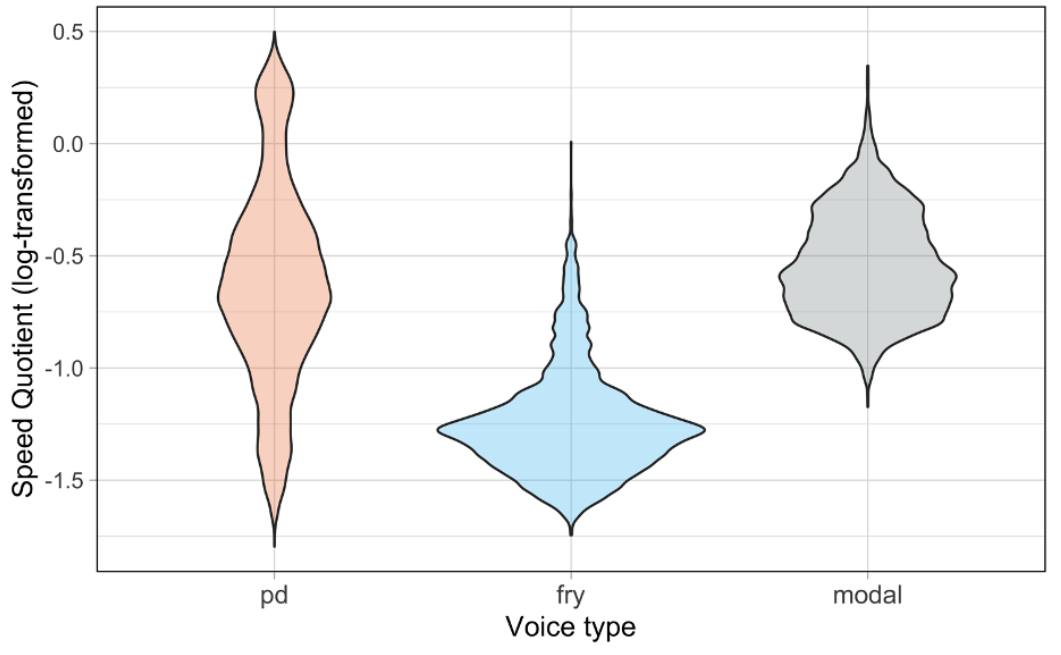


Figure 2.22: Distributions of Speed Quotient (log-transformed) values in period doubling, vocal fry, and modal voice.

Similar to the patterns of CQ, vocal fry stood out for having a low SQ, given that its contacting phase is rapid and its de-contacting or opening phase is slower, resulting in an asymmetry. The distribution of SQ during period doubling resembled that of modal voice, but with a much wider spread, shown by its larger standard deviation.

Table 2.7: Mean (SD) SQ in PD, vocal fry, and modal voice.

Type	Period doubling	Vocal fry	Modal voice
Mean (SD) SQ	0.41 (0.48)	0.07 (0.05)	0.34 (0.21)

The mean SQ of period doubling was slightly higher than that of the modal voice. Linear mixed-effects model comparisons for log-transformed SQ with and without the category of voice (period doubling vs. vocal fry vs. modal), with random intercept and slope by subject, show that vocal fry had a lower SQ than modal ($\beta = -0.20, p < .001$) whereas period doubling had a higher SQ than modal ($\beta = 0.13, p < .01$). This suggests that period doubling does not behave

like vocal fry in terms of its pulse symmetry – it is more temporally balanced, even more so than modal voice.

2.5.4. Alternation in articulatory properties during PD: SQ

In Section 2.5.2, we have found that glottal pulses regularly alternate in period doubling with different contact proportions besides periods and amplitudes. Again, it is expected that alternation also extends to speed quotient, given the alternating pulse shapes and further supported by the widespread distribution of SQ during period doubling in Figure 2.22. Here, I use the same three approaches as used for CQ to investigate the speed quotient. First, I plot the SQ of adjacent cycles in period doubling over the time course of the same set of selected samples with 10 cycles that are more sustained, with reference to the patterns in modal voice and vocal fry, as shown in Figures 2.23-2.25.

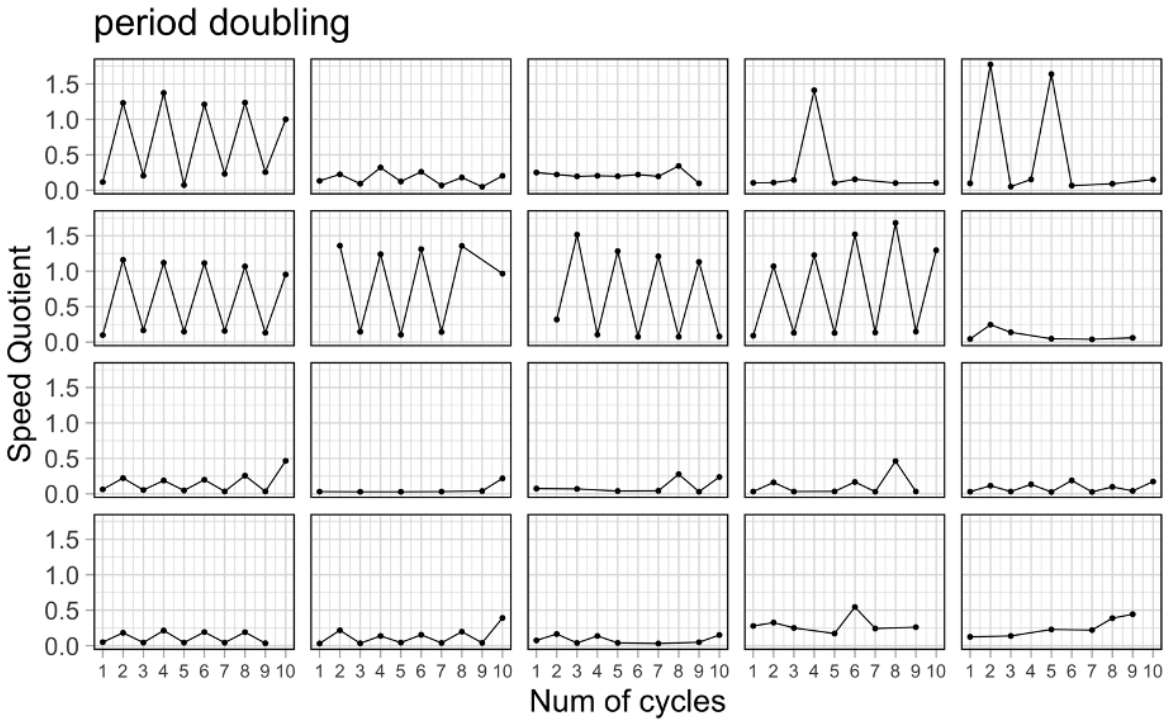


Figure 2.23: Alternation seen in Speed Quotient of each cycle during tokens of period doubling (10 cycles are plotted).

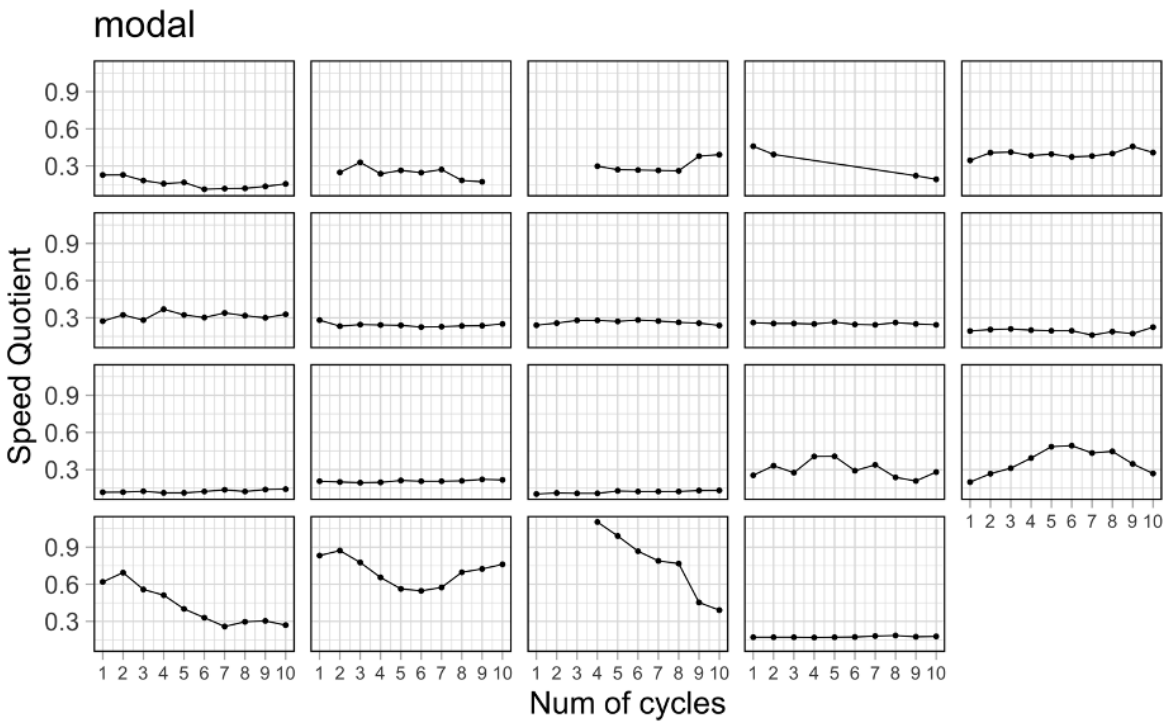


Figure 2.24: Speed Quotient of each cycle during tokens of modal voice (10 cycles are plotted).

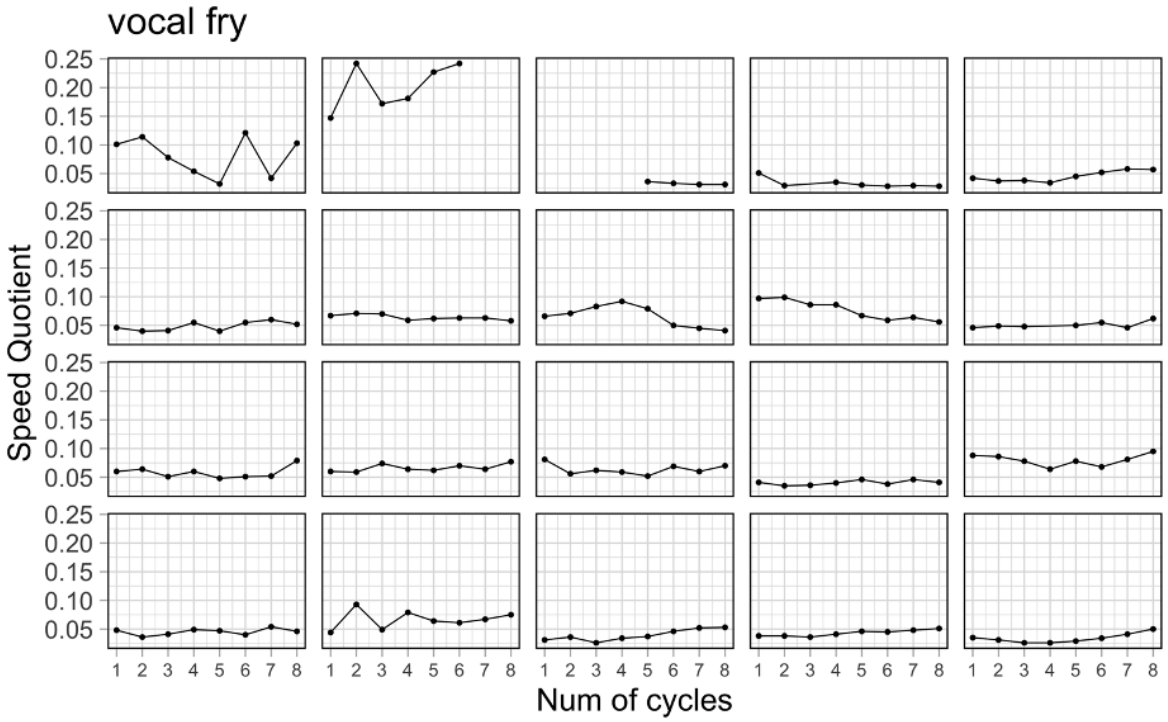


Figure 2.25: Speed Quotient of each cycle during tokens of vocal fry (8 cycles are plotted).

Comparing period doubling to modal voice and vocal fry, many more cases of period doubling showed alternations of the speed quotient, meaning that the pulse shapes of the adjacent pulses alternate in terms of their symmetry between closing and opening phases during articulation. Such a pattern is generally not found in sustained samples of modal voice and vocal fry.

Next, I plot the median FFT coefficients of SQ in the three voice types, as shown in Figure 2.26. The details of this approach are described in Section 2.5.2. Here, the majority of coefficients of SQ during period doubling was concentrated on non-zero values, compared to vocal fry and modal voice, meaning that more cyclicity of SQ values was found during period doubling.

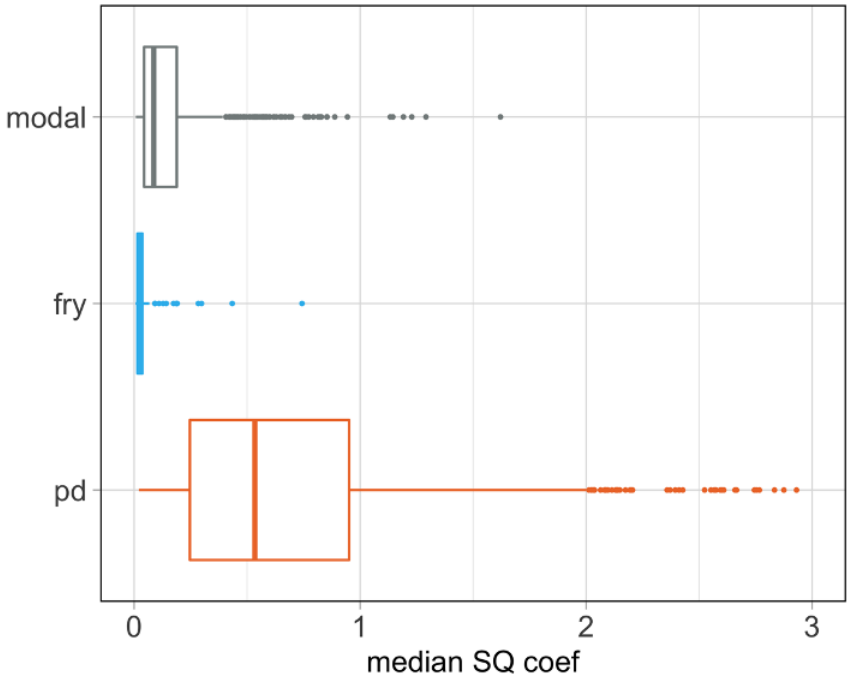


Figure 2.26: Distributions of median FFT coefficients of Speed Quotient in period doubling, vocal fry, and modal voice.

Lastly, I plot the separate distributions of every other glottal pulse from samples of three voice types which contain at least four cycles, as shown in Figure 2.27.

Figure 2.27: Density distributions of Speed Quotient (log-transformed) of every other pulse during period doubling, vocal fry, and modal voice.

The non-overlapping distributions of SQ values in period doubling are distinct from those in vocal fry and modal voice. This confirms that SQ also alternates in period doubling, to add to the pattern of periods, amplitudes, and CQs. Thus, period doubling not only has alternating CQs, but its pulse symmetry also alternates in terms of the relative durations of contacting and de-contacting phases.

2.5.5. Peak increase in contact comparison: PD, vocal fry, modal voice

This section focuses on the third measure derived from the EGG waveform: peak increase in contact (PIC), which indicates the maximum velocity during contact in each glottal cycle. Figure 2.28 shows the different distributions of PIC in three voice types. The majority tokens of vocal fry had a higher PIC than modal voice and period doubling, which is expected given that the vocal folds often close more abruptly during productions of vocal fry. In this regard, the distributions

of period doubling and modal voice were similar; their ranges of variation also spanned widely and resulted in substantial standard deviation.

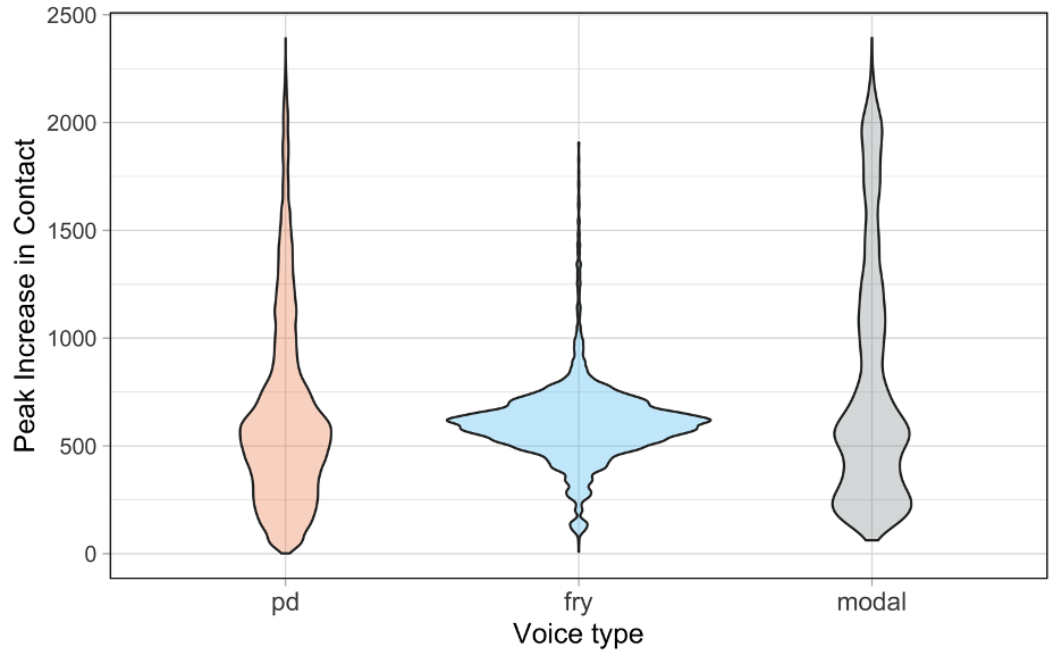


Figure 2.28: Distributions of Peak Increase in Contact in period doubling, vocal fry, and modal voice.

Table 2.8 shows the respective mean (SD) PIC values in these voice types. Linear mixed-effects model comparisons for PIC with and without the category of voice (period doubling vs. vocal fry vs. modal), with random intercept and slope by subject, show that both vocal fry ($\beta = -241.94, p < .01$) and period doubling ($\beta = -213.17, p < .001$) had a lower PIC than modal, and period doubling and vocal fry were not different in mean PIC values. This suggests that period doubling is also produced with abrupt vocal folds closing, but with a larger extent of variation similar to modal voice.

Table 2.8: Mean (SD) PIC in period doubling, vocal fry, and modal voice.

Type	Period doubling	Vocal fry	Modal voice
Mean (SD) PIC	676.15 (439.93)	604.26 (171.72)	798.98 (563.45)

2.5.6. Alternation in articulatory properties: PIC

Following up on the previous results on contact quotient and speech quotient that the two subcycles in period doubling differ impressionistically in terms of their shape and articulatory characteristics, I expect that PIC values also exhibit an alternating pattern. Figures 2.29-2.31 compare PIC for period doubling, vocal fry, and modal voice over a time course in selected tokens with 10 cycles. As with CQ and SQ, the peak increase in contact showed some cyclicity in values-per-cycle for period-doubled voice, suggesting alternations in cycle peak velocity that were absent in vocal fry and modal voice.

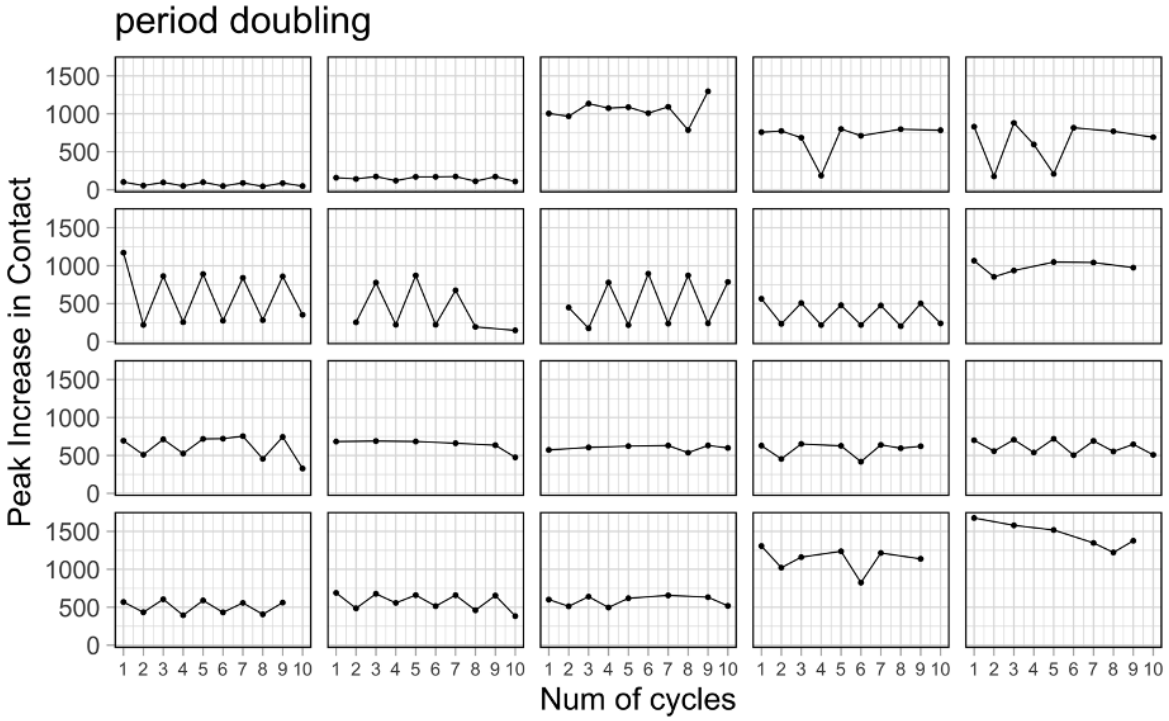


Figure 2.29: Peak increase in contact of each cycle in sustained period doubling tokens (10 cycles are plotted).

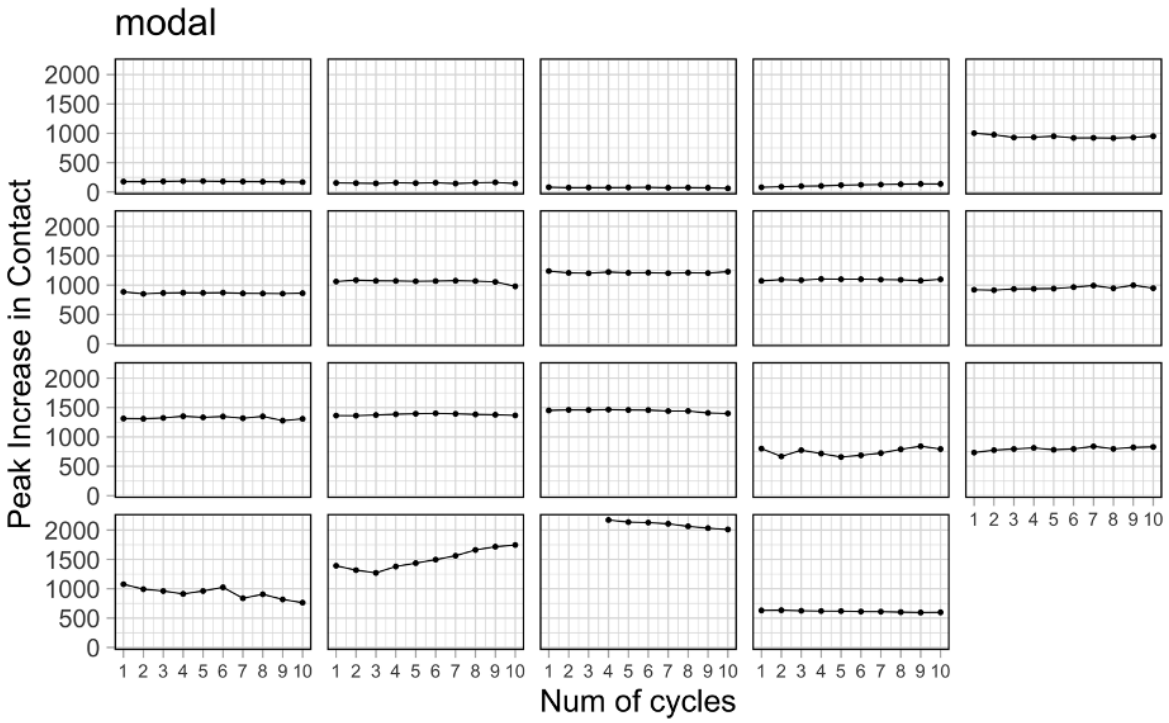


Figure 2.30: Peak increase in contact of each cycle in modal tokens (10 cycles are plotted).

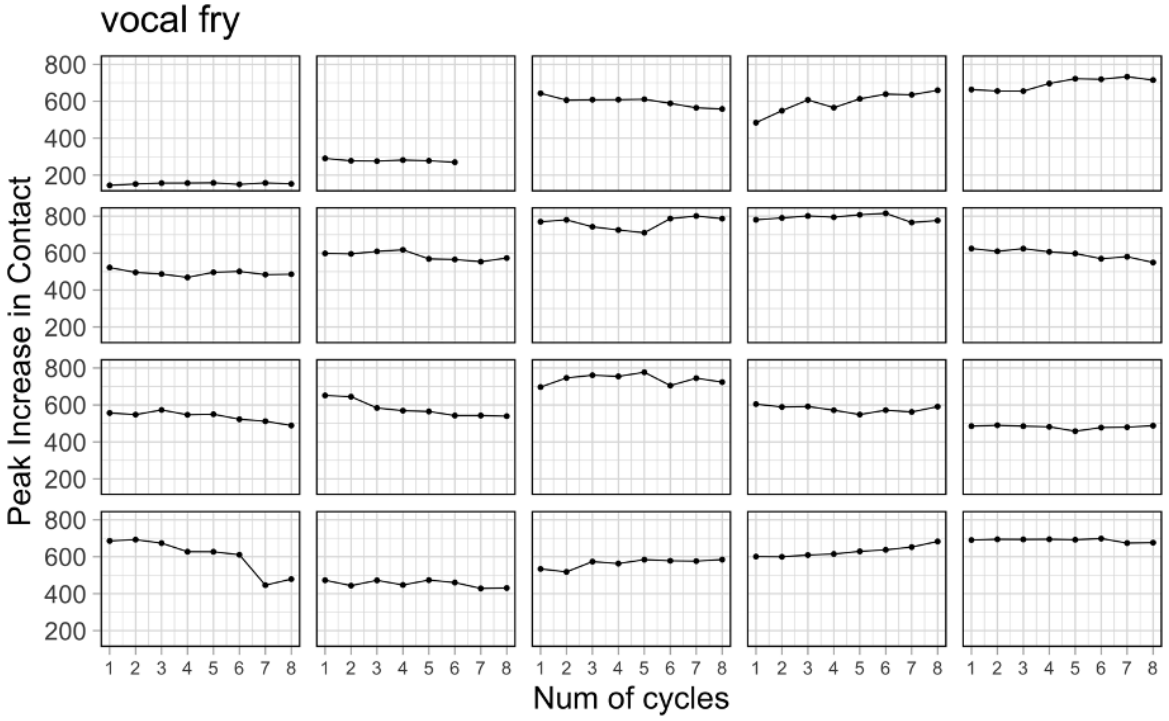


Figure 2.31: Peak increase in contact of each cycle in sustained vocal fry tokens (8 cycles are plotted).

Next, Figure 2.32 shows the distributions of median FFT coefficients of PIC in period doubling, vocal fry, and modal voice. Again, there were more non-zero coefficients in period doubling than the other two types of voice, supporting the observations in the previous graphs that there are more regular alternations in PIC during period doubling.

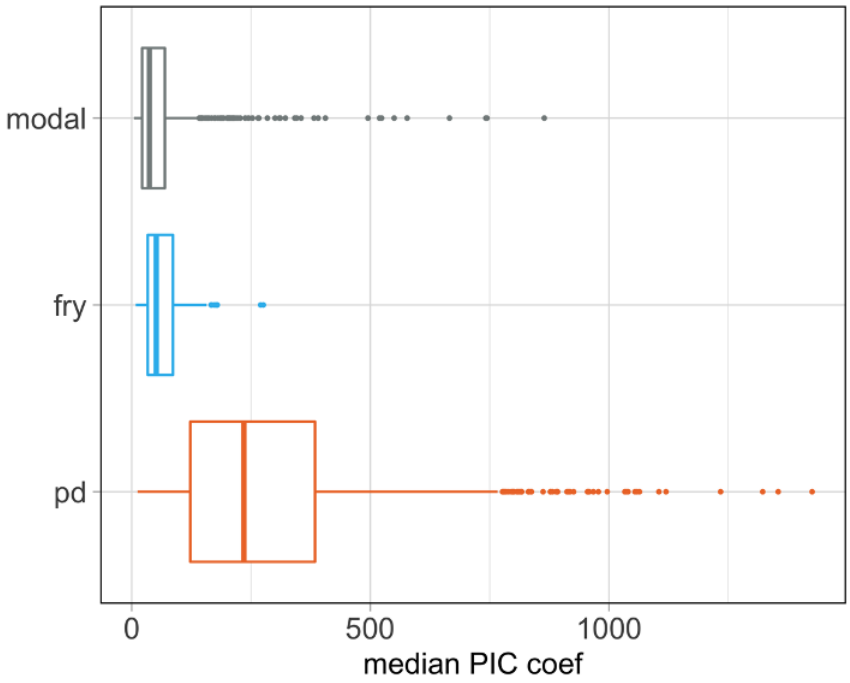


Figure 2.32: Distributions of median FFT coefficients of Peak Increase in Contact in modal voice, vocal try, and period doubling.

Lastly, I plot the density distributions of every other glottal cycle during these voice types, as shown in Figure 2.33. Period doubling apparently had non-overlapping distributions of PIC in every other pulse over the time course, which was lacking in vocal fry and modal voice. Thus, we confirm the pattern that besides the varying lengths, amplitudes, contacting duration, and duration of vocal fold contacting and de-contacting, the speed of contacting changes from pulse to pulse in period doubling.

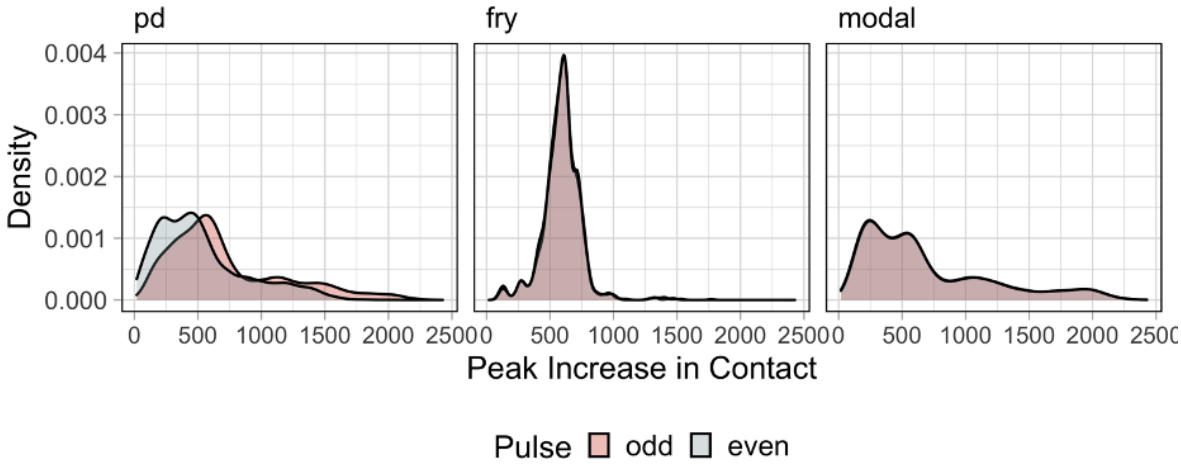


Figure 2.33: Density distributions of Peak Increase in Contact of every other pulse during period doubling, vocal fry, and modal voice.

In sum, we see that producing period doubling involves phonation of two distinct types of glottal pulses which not only regularly alternate between their periods and amplitudes: strong vs. weak and/or short vs. long, but also between pulses with more versus less constriction, more vs. less symmetrical pulses, and more rapid vs. slower contacting. Importantly, these articulatory properties of period doubling essentially speak to the co-existence of two distinct types of glottal pulses articulated with different voice qualities and pitches, which makes period doubling a distinct voice category from both vocal fry and modal voice. The concurrent voice qualities likely contribute to the indeterminate pitch – as well as quality – that one perceives during period doubling.

2.6. Results: EGG spectrum analysis of different types of period doubling

In this section, I discuss the findings in the frequency domain of EGG signals: spectral differences of period doubling within its subtypes and those compared to vocal fry and modal voice. I will analyze spectra using representative samples found in the corpus with a qualitative and impressionistic examination, while formal calculations and models to quantify the frequency components of period doubling are saved for future studies. I extracted and analyzed spectra of all the period-doubled tokens located in the EGG waveform from nine speakers (4F, 5M: F04, F30, F37, M05, M15, M38, M39, M40, M46). The spectral slices were obtained using a *Kaiser2* window function with the same width as the original EGG waveforms and fast Fourier transform in PRAAT. Due to the relatively small window length of period-doubled tokens (~ 30 *ms* on average across tokens), considerable spectral noise was present in a large number of the spectra, making it challenging to discern generalizable patterns or regularity in those spectra. And because male speakers had longer periods and fewer cycles within a period-doubled sample, the spectra were often noisier than data from female speakers. Thus, I focus on the spectral characteristics from tokens with longer, more sustained, period-doubled cycles. Notably, alternating patterns on the spectrum were more commonly found in women speakers, perhaps suggesting a stronger presence of sub-harmonics along with the harmonic structure in female voice due to their higher f0 ranges than men.

2.6.1. Noticeable patterns in EGG spectra

I first explore the spectral differences as shown in Figures 2.1 and 2.2 for the two distinct subtypes of period doubling: amplitude-modulated (2.1b) and amplitude and frequency-modulated tokens (2.2b). Figure 2.34 (same as 2.1b, pasted below) shows that, at the lower frequencies, H2 is stronger than H1, and the two sets of even and odd harmonics alternate in magnitude in the lower frequency range up to H8 ($\sim 773Hz$) such that the even harmonics are stronger than the odd ones.

After a plateau of H9 and H10, it starts alternating again at H11 – this time, the odd harmonics are stronger than the even ones.

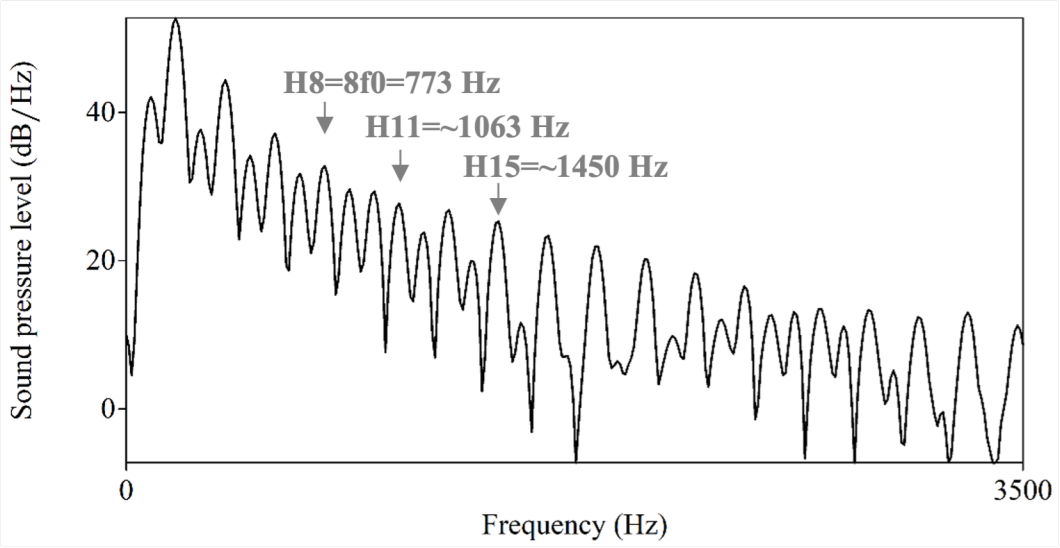


Figure 2.34: EGG spectrum of amplitude-modulated period doubling.

At the higher frequencies, starting at H15, the even harmonics become weaker and damped, nearly seen as interharmonics or subharmonics, resulting in a larger spacing between the harmonics which only consists of the odd set of harmonics (H15 – H21, and later higher harmonics $\sim 3500Hz$). This was commonly found across the amplitude-modulated tokens ($\sim 40\%$), though the specific harmonics that alternate may vary according to the specific sample. The following example of another amplitude modulated token, as shown in Figure 2.35, illustrates that the alternation starts from H1 through H6 ($\sim 655Hz$) and flips and resumes at H11 ($\sim 1200Hz$) throughout the mid to high frequency ranges.

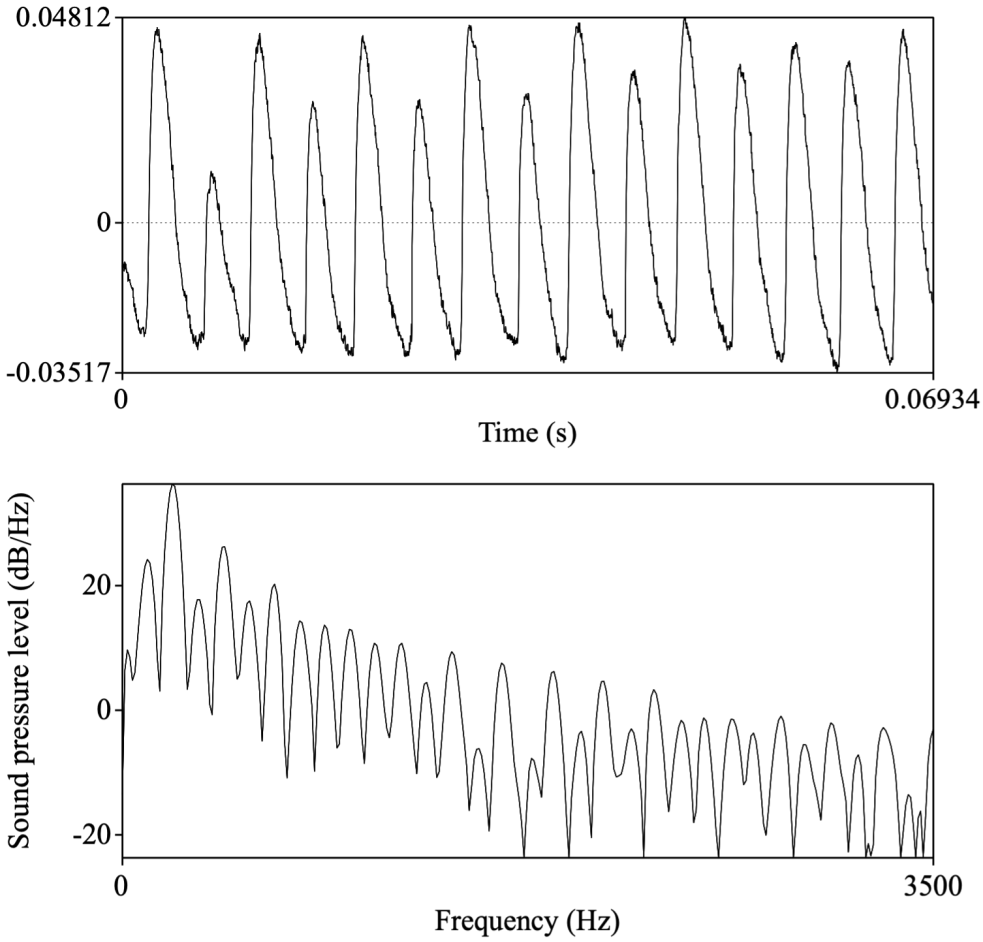


Figure 2.35: Waveform and spectrum of amplitude modulated period doubling (from speaker F04).

Figure 2.36 (same as 2.2b, pasted below) of an amplitude and frequency modulated example shows a different spectral pattern: in the lower frequency range, H1 is stronger than H2, and regular amplitude dips are seen for certain harmonics such that H2 < H3, H6 < H7 and H8, H10 < H11 and H12, etc, instead of a logarithmic power spectrum. This profile of groupings of harmonics may have implications for sounding of period doubling. In the higher frequency range, the bandwidths of harmonics are evenly spaced, compared to the amplitude-modulated spectra, which show skewed bandwidths between the alternating harmonics and the weaker harmonics are narrower and even damped.

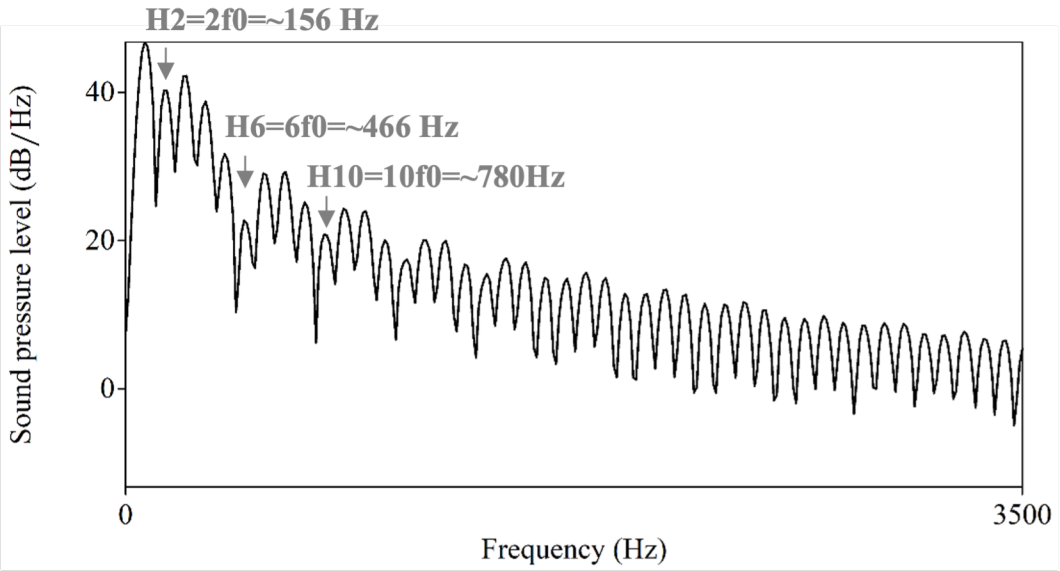


Figure 2.36: EGG spectrum of amplitude and frequency-modulated period doubling.

Figure 2.37 shows the waveform and spectrum of another amplitude and frequency modulated token, of such amplitude dips and groupings of harmonics, even with a slightly different spectral pattern.

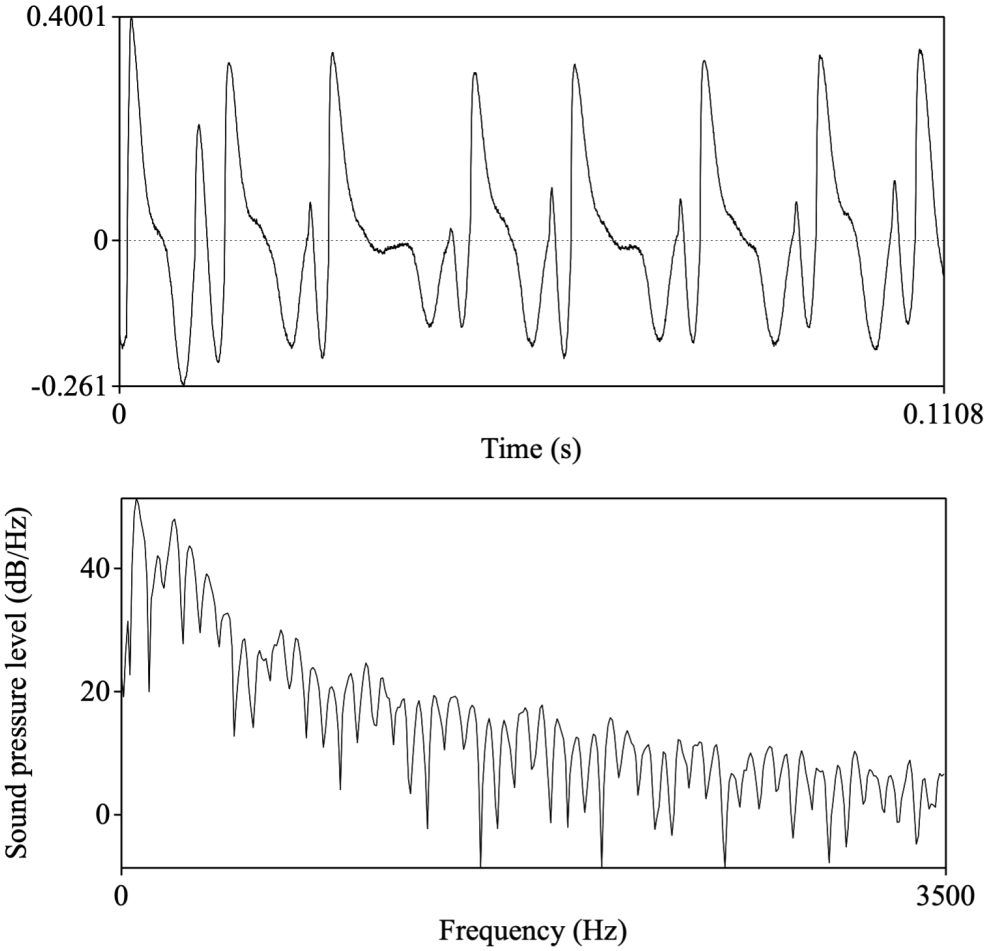


Figure 2.37: Waveform and spectrum of amplitude and frequency modulated period doubling (from speaker F37).

Besides the two representative patterns, other noticeable configurations were identified and generalized among the various spectra across different tokens. Next, I discuss two other frequently-observed configurations in the two kinds of period doubling.

Figure 2.38 depicts another example of the amplitude alternation of harmonics. The spectrum of this token is slightly different from the above-mentioned alternation in amplitude modulated tokens by having H3 $>$ H4.

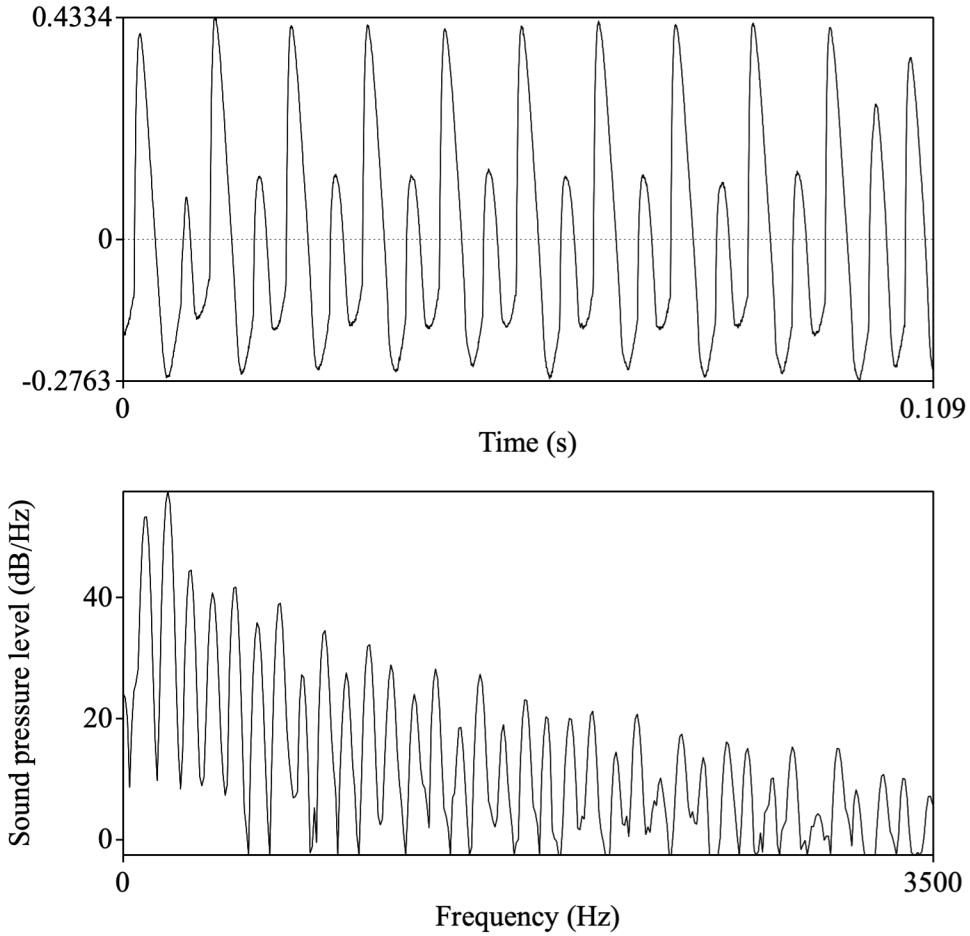


Figure 2.38: Waveform and spectrum of amplitude modulated period doubling (from speaker F37).

Figure 2.39 depicts an amplitude and frequency modulated token where H1 and H2 appear to be equal in magnitude, and H4 dips while H5 appears to be equal to H6. Further, H7, H10, H13, H16 all dip in amplitude.

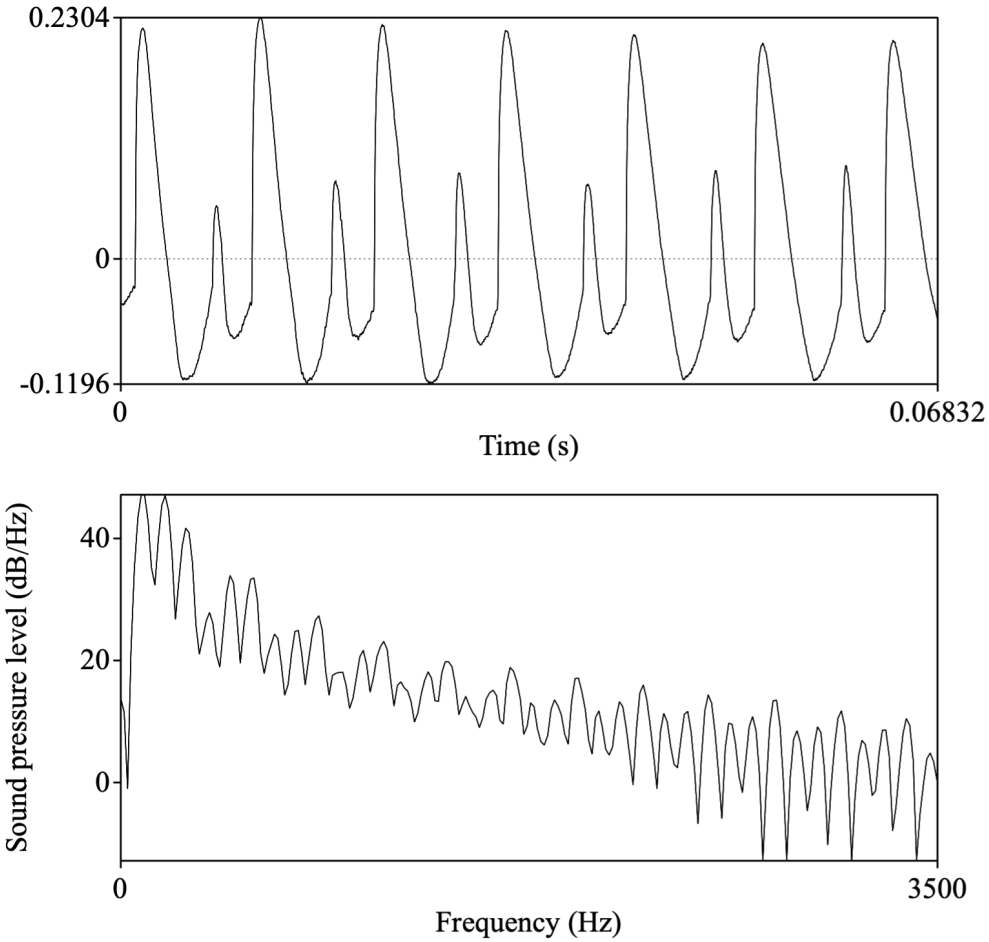


Figure 2.39: Waveform and spectrum of amplitude and frequency modulated period doubling (from speaker F30).

I also calculated the proportion of tokens that show a stronger H1 or a stronger H2 on the spectrum, and the numbers are comparable: 40.7% of the tokens had a stronger H1 and 40.2% of the tokens had a stronger H2, whereas 11% of the tokens had an equal H1 and H2. The other 8% of the samples were ambiguous because of substantial spectral noise as to whether H1 or H2 was the stronger component.

2.6.2. Relative amplitudes of subharmonics and harmonics

Next, I discuss the relationship between subharmonics and harmonics, as the presence of subharmonics is one of the key characteristics of period doubling and likely contributes to its bitonal and rough percept. When the amplitude of the weaker pulse becomes stronger and comparable to the stronger pulse, I observe stronger subharmonics, which could become harmonics and alternate with the original set of harmonics. Figures 2.40-2.41 include two examples showing how such weak interharmonics change over the time course of period-doubled tokens.

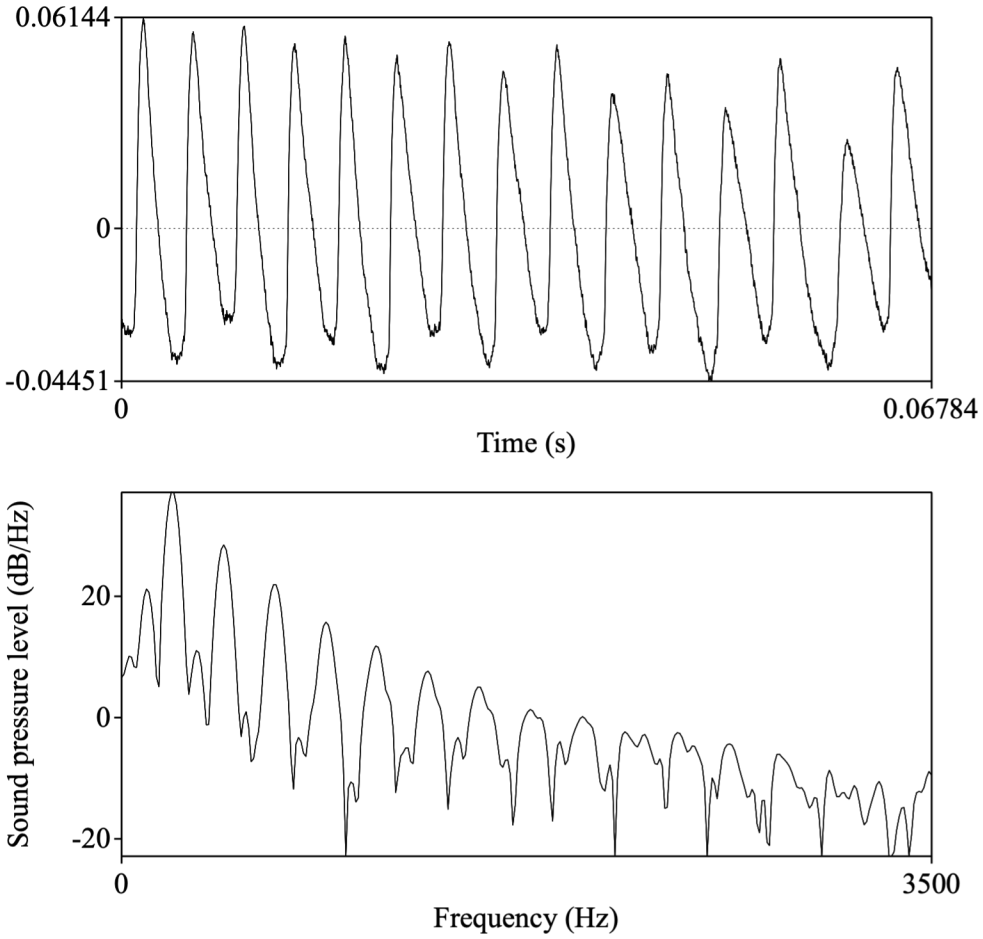


Figure 2.40: Waveform and spectrum of amplitude modulated period doubling showing interharmonics (from speaker F04).

In Figure 2.40, based on the waveform, because the amplitude of the second glottal pulse is decreasing while the cycles proceed, it is more and more likely that it creates a percept of lower fundamental rather than the original higher one because the cycle is elongated to include two pulses. As evidenced by the spectrum, its H1 is lower in both amplitude and frequency. However, as frequency increases, the contribution of the subharmonics becomes weaker and weaker such that when reaching higher frequencies, the subharmonics nearly becomes undetectable, with the presence of spectral noise. By having a lower H1 and wider spacings between harmonics, it is expected to have a competing tonal percept.

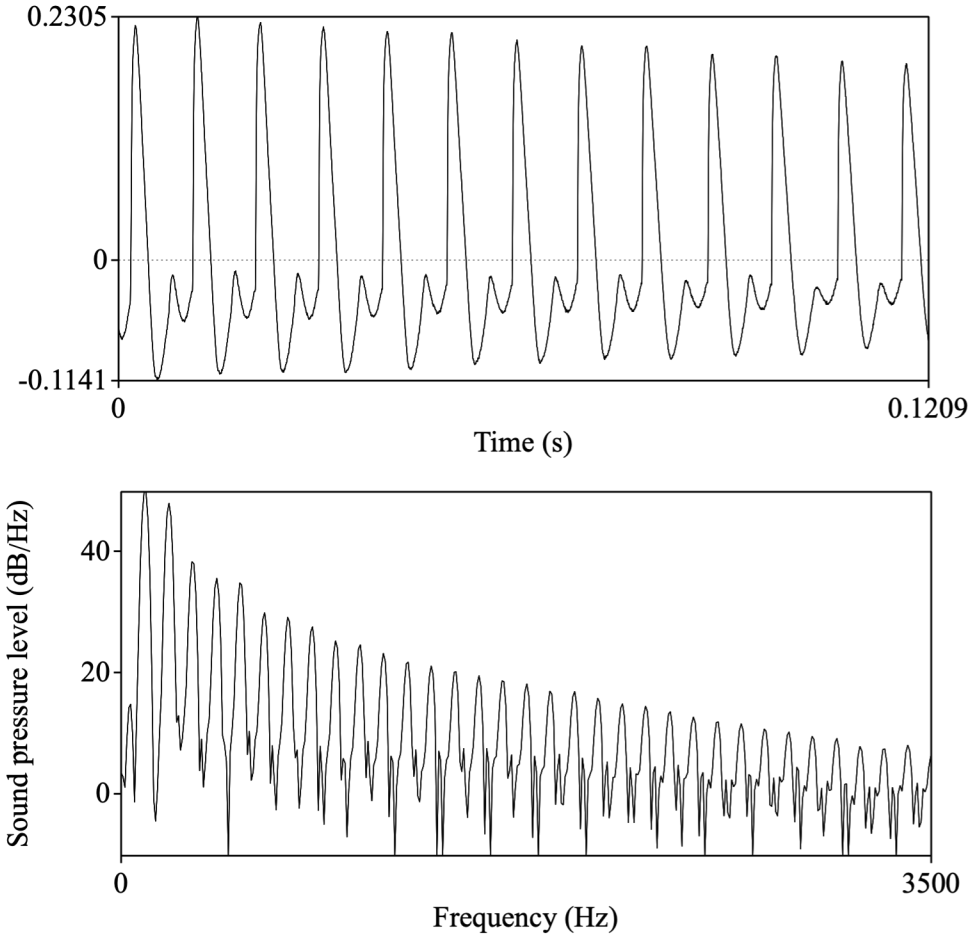


Figure 2.41: Waveform and spectrum of amplitude modulated period doubling showing interharmonics in higher frequencies (from speaker F30).

In the example of Figure 2.41, damping of every other pulse was seen on the waveform. The spectrum shows a robust H1 at a lower frequency, even though H2 is almost equal to H1. It is less obvious whether two sets of harmonics actually exist because of the narrow even spacing throughout frequency ranges, which corresponds to the strong-to-strong cycle in the time domain. Weak interharmonics most likely are treated as spectral noise. This token would sound less ambiguous than the previous example, with a lower pitch, and possibly roughness stemming from the weaker pulses. To summarize, even though the two tokens in Figures 2.40-2.41 both have a fundamental frequency of $\sim 100Hz$, their spacings between harmonics as shown in the spectrum are different.

In Figures 2.42-2.43, I plot the narrowband spectrograms of amplitude modulated and amplitude-and-frequency modulated period doubling to show the split of f0 and the presence of subharmonics, which is consistent with Martin (2012).

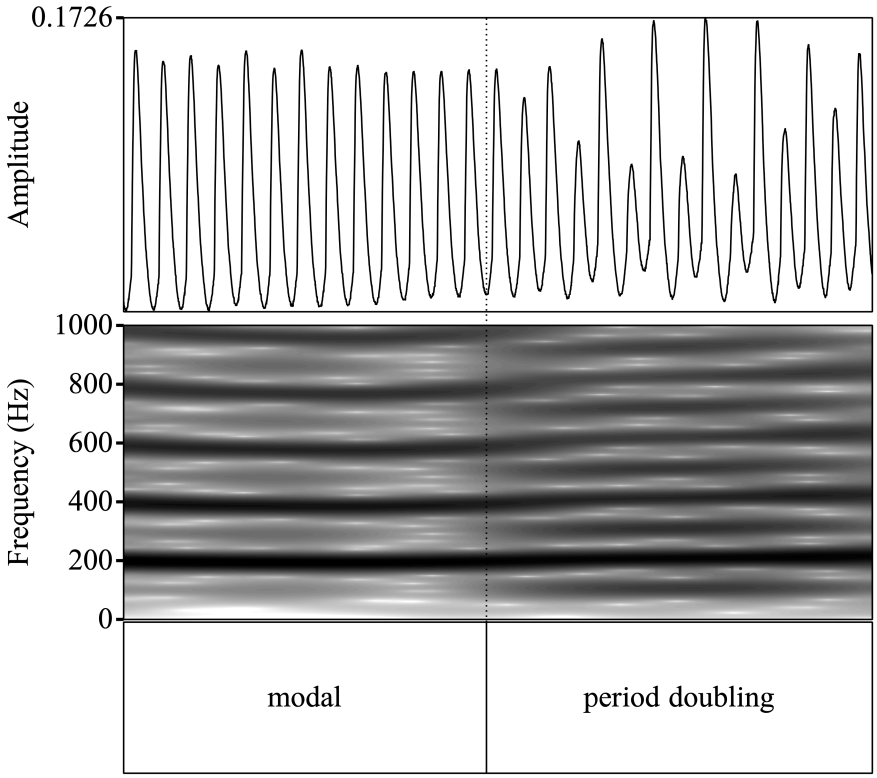


Figure 2.42: Waveform and spectrogram of amplitude modulated period doubling showing bifurcation of f0 (from speaker F30).

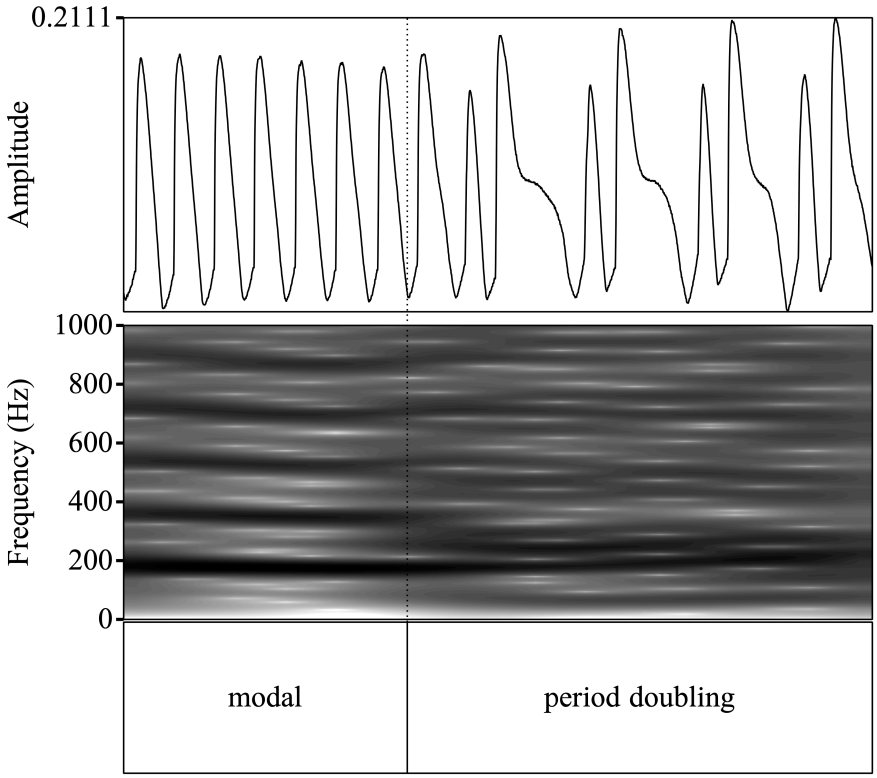


Figure 2.43: Waveform and spectrogram of amplitude and frequency modulated period doubling showing bifurcation of f0 (from speaker F30).

The f0 bifurcation is clearly seen in Figure 2.42 when compared to the modal portion; however, with both amplitude and frequency modulation in Figure 2.43, the bifurcation of f0 is less well demarcated and appears to be non-parallel as opposed to the amplitude modulated token.

Lastly, to compare the spectral characteristics of period to that of vocal fry and modal voice, I plot the waveform and spectrum of sample vocal fry and modal voice, as shown in Figure 2.44.

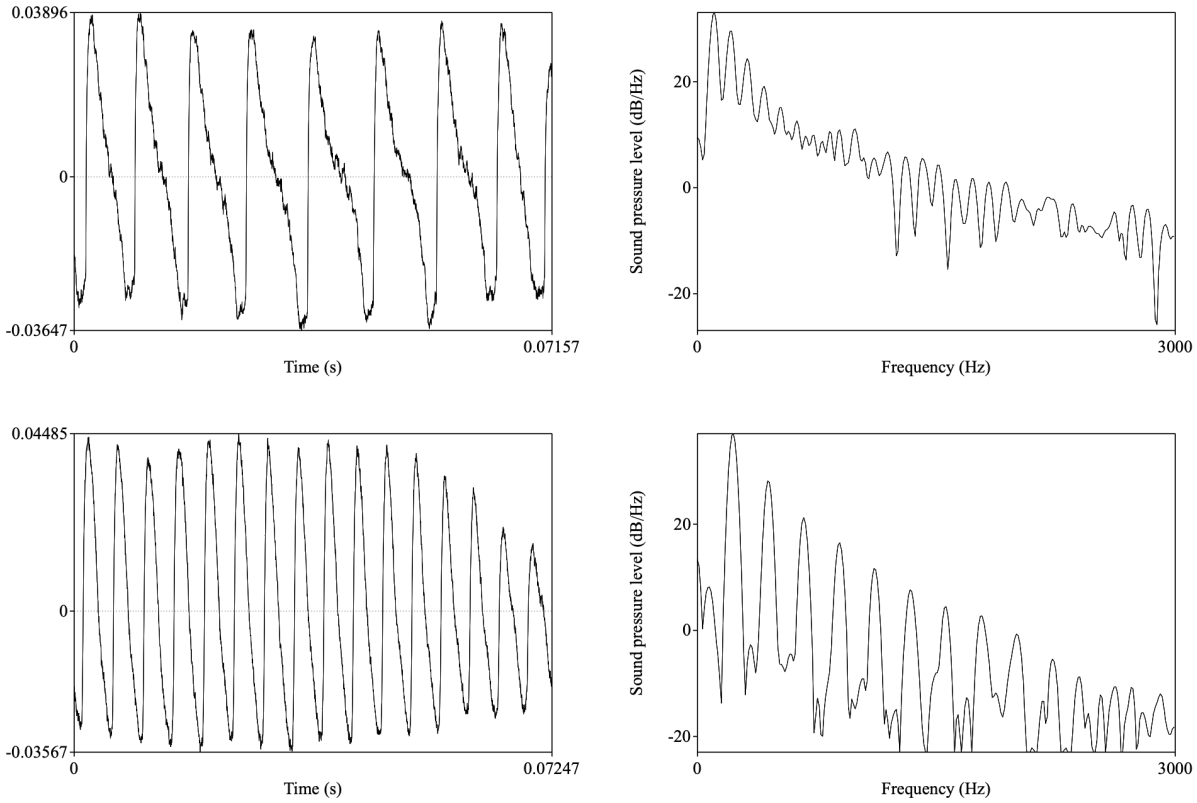


Figure 2.44: Waveforms and spectra of vocal fry (first row) and modal voice (second row) (from speaker F04).

Compared to vocal fry, period doubling has a substantially stronger presence of subharmonics along with the original harmonic structure, and alternation between sets of (sub)harmonics are prominent. In modal voice, subharmonics are sometimes seen as spectral noise, and the higher frequency components are cleaner. In general, regular alternation and amplitude dipping among specific harmonics are likely only a defining trait of period doubling.

2.7. Discussion

One of the most striking characteristics of period doubling is the presence of multiple periods; thus, multiple f0s. By extracting the three periods during this voice: the longest and the two alternating shorter ones, we can compare these frequencies among themselves and to the original f0

of modal voice of the same speaker. The overall average frequency ratio exhibited a relationship of 3:2, which agrees with the subharmonic vibratory pattern found by Švec et al. (1996). The stronger pulse was twice in amplitude than the weaker pulse on average. Both frequency and amplitude ratios differed among genders and subtypes of period doubling: amplitude modulated or concurrent amplitude and frequency modulated. Based on the descriptive measures that quantify the alternating cycles in period doubling, the ranges of frequency and amplitude alternations lay the foundation for further investigating how the different configurations of modulation influence pitch perception and production during period doubling. Chapter 4 includes the study of pitch perception and shadowing during period doubling that directly follows up on exploring the relationship between the varying degrees of frequency/amplitude modulation and the result of pitch percept.

I also confirmed that the period-doubled f0 is about half that of modal f0, which is expected and evidenced by both waveform and narrow spectrogram showing that the f0 splits into half with its "doubled" periods. Regarding the relationship among the f0 in the modal voice and the two glottal frequencies, it was observed that the modal f0 is at times closer to the first glottal frequency, or the second one, varying by speakers, tokens, and the degrees of modulation. However, how the original f0 bifurcates into two higher frequencies which then converge back, and how the original glottal pulse splits into two successive smaller pulses are yet to be determined. Future studies may investigate the bifurcation pattern using high-speed imaging techniques such as laryngoscopy to capture the variations during modal and period-doubled vocal folds vibration. In particular, since articulation and acoustics are in a many-to-many mapping, and the focus of the present study is on the product rather than underlying mechanism of such vibration, it would be useful to probe the multiple vibratory configurations towards realizing period doubling in more detail to understand why period doubling is formed and used frequently. In addition, as Hanson et al. (1983) noted, sometimes a few abnormal conditions can make interpretation of the EGG waveform difficult. For example, the presence of a tissue or mucus between the vocal folds or the ventricular folds may affect the impedance between the two electrodes to be deviant from the standard. To be

able to further compare the findings of period doubling in natural speech to pathological voices, simultaneous recording of glottal measures with fiberoptic laryngoscopy would be necessary.

A novel and significant finding is that pulses in period-doubled voice not only alternate between periods and amplitudes, but also between degrees of glottal constriction, proportions of contacting and opening phases of the vibratory cycle, and speeds of constriction. These have far-reaching consequences in explaining how the multi-faceted human vocal folds can vibrate between different modes to realize the various linguistic meaning and serve communicative functions. Different from 'diplophonia' in pathological voice where two vocal folds are asynchronously vibrating at different frequencies (Gerratt et al., 1988; Dejonckere and Lebacq, 1983), it is probable that period doubling in natural speech has both vocal folds oscillating between two 'settings'. Previous studies often characterize period doubling as having an indeterminate pitch with a low and rough quality (Keating et al., 2015; Yu, 2010; Schreibweiss-Merin and Terrio, 1986), mainly because of the presence of simultaneous periodicities. The present findings show that roughness may come from its indeterminate quality in addition to, or instead of, its indeterminate pitch. Concurrent voice qualities can be derived from the alternation in voice qualities inherent to period doubling, on the top of distinct pitches. In addition, alternating more and less constricted qualities may interfere with other voice qualities such as breathiness. This was attested in Swedish (Hedelin and Huber, 1990) such that period doubling coocurred with breathiness. It may also explain the covariation between harshness and breathiness in languages such as White Hmong (Keating et al., 2010), !Xóõ (Garellek, 2020), and Northern Vietnamese (Brunelle et al., 2010). Perhaps the voicing during breathiness also contains period doubling (especially when produced at a low f0), which could lead to the development of breathy to harsh voice. This raises more questions about the interaction between voice quality and pitch, for example, how do we perceive multiple voice qualities interacting with different pitches simultaneously, and why do we perceive roughness as a result of indeterminate quality and pitch?

The case of period doubling will be of interest to models of pitch perception (de Cheveigné, 2005) and tone perception (Gandour, 1978). For example, to what extent do the alternating glottal frequencies matter in addition to the presumptive f0 derived from the fundamental meta-cycle? How do the alternating frequencies contribute to the bitonal and competing percept? Do we perceive period doubling based on cues in the time domain or the frequency domain? What is the relationship between the frequency/amplitude ratios in the waveform and the presence and strength of subharmonics in the spectrum? How do subphonemic differences in voice and the indeterminate pitch contribute to lexical tonal contrasts? These ought to be addressed in future studies.

As a subtype of creaky voice, period doubling is different from vocal fry or predictions of creaky voice, because it does not necessarily involve glottal constriction. Variation across pulses was attested, but even in the separate distributions of odd and even pulses, the contact quotient of period doubling was closer to those of modal voice, if not lower (even less constricted). This suggests that period doubling is less constricted, whereas vocal fry is known for its association with glottal constriction (Gerratt and Kreiman, 2001), which is consistent with Švec et al. (1996) that a large open quotient distinguishes this kind of subharmonic vibration from vocal fry. Likewise, the speed quotients during period doubling were closer to those of modal voice, rather than lower (as expected for more constricted voice qualities). This also suggests that period doubling has a temporally balanced vibrating phase. As for the measures of peak increase in contact, the two alternating pulses in period doubling resembled modal voice and vocal fry one each (Figure 2.33). In sum, period doubling is a complex phenomenon, and the fuller picture requires two approaches: analyzing the meta-cycle synthetically, and the two smaller cycles analytically. The combination of both approaches better depicts period doubling as a phenomenon distinct from vocal fry and modal voice.

Where does period doubling fit within a taxonomy of creaky voice subtypes? Canonical characteristics of creak consist of low f0, constriction, and irregularity (Garellek, 2019); however,

period doubling does not show either attribute systematically. Occasional irregular pulses during period doubling were attested, but they are not as the same degree of aperiodic voice. Though the pulses also alternate between modes of more versus less constriction, the degrees of constriction during period doubling were similar or even lower than modal voice. Considering the multiple frequencies in period doubling, it would not be sufficient to conclude that period double is tied to a particular f0 target (e.g., low pitch), whereas vocal fry is associated with regular low pitch (Keating et al., 2015). Overall, period doubling does not fit in neatly to the taxonomy with either one of the established attributes. Nonetheless, period doubling may be seen intimately linked to low pitch (and possibly constriction), with its alternation in frequencies and voice qualities, rather than inherently low pitch or increased constriction. If it is the case, then low pitch may be a useful dimension to be added in the taxonomy – a question of whether the canonical characteristics confined to acoustic attributes should be expanded to include perceptual features. The existing creaky voice subtypes – prototypical, vocal fry, multiply pulsed voice (period doubling), and non-constricted creak can all be characterized by low pitch (aperiodic voice does not have a perceived pitch).

Lastly, the various spectral characteristics of period doubling remain a puzzle at the current stage. Several generalizable patterns were observed, such as competing strengths of H1 and H2, alternating harmonics, groupings of harmonics based on amplitude dips, and varying strengths of subharmonics; quantitative and signal processing models are needed to unfold and demystify the spectral differences within types of period doubling in future work.

2.8. Chapter summary

In this chapter I investigated the articulatory properties of period-doubled voice in typical Mandarin speech using a read speech corpus using EGG. I start by quantifying the characteristics of the waveform using proposed frequency and amplitude ratios to capture the relative durations and strengths of every other alternating cycle. We see that, on average, the two glottal cycles tend

to be 3:2 in frequencies and 2:1 in amplitudes. Amplitude-modulated and both amplitude and frequency-modulated period-doubled tokens have differences in the relationship between alternating pulses. I then investigate glottal constriction measures of period doubling taken from the EGG waveforms to make comparisons with vocal fry and modal voice. We see that, if viewed as a meta-cycle (a cycle consisting of a pair of cycles), period doubling largely resembles modal voice in terms of articulatory properties; but crucially, when divided accorded the two sub-cycles, period doubling is uniquely featured by alternations in contact durations, pulse shape symmetry, and speed of vocal fold contact, which distinguishes period doubling from other voices well. The alternating qualities and pitches likely contribute to the indeterminate and rough percept during period doubling. Further, the fact that period doubling does not exhibit degrees of constriction comparable to expectations of creaky voice poses challenges to the taxonomy of creaky voice. It is proposed that low pitch may be added as a relevant dimension to capture most subtypes of creaky voice.

Finally, in the frequency domain, the presence of subharmonics is attested, and various spectral patterns are found to signal subtypes of period doubling, complementing the findings in the waveforms, though their implications and mechanisms merit further investigation. Gathering from the novel patterns and distributions of articulatory features found in this study, period doubling may have a different linguistic distribution from other kinds of creaky voice (e.g., vocal fry) and likely affects pitch and lexical tone perception differently from other voice qualities. The discussion of these assumptions and predictions will be continued through studies in the following chapters.

Chapter 3

Production of period doubling: acoustic characteristics and linguistic distribution

3.1. Introduction

Period-doubled phonation is characterized by the presence of two cycles (in the time domain) alternating in frequency and/or amplitude between two states (high and low; long and short) in the speech signal (Titze, 1994). In Mandarin, the two glottal pulses are found to be articulated by alternating pitches as well as voice qualities (Huang, 2022; Chapter 2). Still, with the two sets of alternating frequencies, it remains unclear how period doubling differs in acoustics and perception from other voice types, and the role it might play in pitch and tone perception. For example, under the framework of creaky voice subtypes in Keating et al. (2015), common attributes of creaky voice include low f0, glottal constriction, and noise. Period doubling is expected to demonstrate acoustic characteristics of some of these attributes. But this is also true to various extents for other subtypes of creaky voice: for example, vocal fry is expected to be constricted and low in f0. So how does period doubling differ from fry? For one, the presence of more than one f0, unique to period doubling, is expected to lead to higher subharmonics in period doubling than vocal fry. For another, the presence of alternating frequencies does not necessarily imply a low f0, which is a primary feature of vocal fry. Therefore, investigating the acoustic properties of period doubling will help clarify its relation to other subtypes such as vocal fry, which has implications for taxonomy of creaky voice as a refined phonetic category and potentially different linguistic functions of creak subtypes.

This chapter thus probes the acoustic characteristics of period doubling using temporal and spectral measures, adopting Mandarin as a representative language with predictable non-modal phonation along with contrastive pitch contours. The predictable non-modal phonation includes vocal fry and period doubling, both of which are variably attested as discussed in Yu (2010), though how their acoustic characteristics differ and where the two types occur in the same contexts remain unexplored.

The first goal is to determine how period doubling differs acoustically from modal voice and vocal fry, in order to probe period doubling's fit within Keating et al.'s (2015) taxonomy of creaky voice subtypes. This will be done using both statistical models of single acoustic features and computational models of classification. Three families of acoustic measures, known to be representative of voice quality (Klatt, 1980; Kreiman et al., 2010; Garellek, 2019), are included: spectral tilt measures (H1*–H2*, H1*–A1*, H1*–A3*, and other measures in different frequency ranges, corrected for formant frequencies and bandwidths), periodicity measures (Harmonics-to-Noise Ratio, Cepstral Peak Prominence, Subharmonic-to-Harmonic Ratio), and energy measures (Energy, SoE: strength of excitation).

The second goal is to capture the linguistic distributions of period doubling and vocal fry, two subtypes of creaky voice allophonic to Mandarin tones, in order to probe how period doubling might be used linguistically. I start with probing the distributions of individual acoustic measures varying by the voice types, and then use all the measures to devise computational classification algorithms. Then, I present the prosodic and segmental distributions of period doubling and vocal fry, to infer the linguistic factors that drive the occurrence of period-doubled voice, and explore the different linguistic implications for subtypes of creaky voice.

3.2. Background

3.2.1. Production studies on period doubling and vocal fry

Keating et al. (2015) proposed a framework of classifying creaky voice into subtypes according to their acoustic properties. Vocal fry is typically characterized as having a low f0, glottal constriction (low H1–H2), and damped pulses (low noise, and narrow formant bandwidths). Period doubling, which falls into the category of multiply-pulsed voice, typically has irregular f0 (high noise), glottal constriction (low H1–H2), and high subharmonics (high SHR). Gerratt and Kreiman (2001) also described period doubling as one of the most common supraperiodic phonations with numerous interharmonics between harmonics of the f0. We have seen from the articulation data in Chapter 2 that period doubling does not simply entail increased glottal constriction, especially taking into consideration its alternating patterns in voice quality between two distinct glottal cycles. Thus, in this chapter, I review the acoustics of period doubling, in order to examine the similarities and differences between the acoustic attributes of period-doubled voice and vocal fry and modal voice.

Hedelin and Huber (1990) classified types of laryngealization in Swedish texts, among which vocal fry (termed 'creak') was defined as a pattern of glottal vibration consisting of a train of discrete low-frequency pulses at a frequency range between $20Hz$ and $50Hz$, with nearly complete damping of the vocal tract, and period doubling (termed 'diplophonoic phonation') was characterized by alternating strong and weak glottal excitations during phonation, typically sustained over longer periods of speech. They also found that period doubling mostly occurred at word junctures between voiced sounds and often cooccurred with other types of voicing (e.g., breathiness, aperiodic voice, devoicing). In English, Redi and Shattuck-Hufnagel (2001) developed categories of irregular voicing ('glottalization'), including vocal fry and period doubling, and found that period doubling occurred more frequently at the ends of utterances than the ends

of utterance-medial intonational phrases, and more frequently at boundaries of intonational compared to intermediate phrases.

Related automatic classification schemes of creaky voice distinguish vocal fry and period doubling, though using various terms (damping and diplophonia, Batliner et al., 1994; vocal fry and diphophonia, Martin, 2012). Using narrow band Fourier spectrum analysis on evaluation of sudden changes in the harmonic pattern of consecutive frames, Martin (2012) effectively detected these two creak subtypes by simulating the visual detection of subharmonics from spectrograms.

Other studies found that in English, vocal fry occurred more frequently in men versus women (Irons and Alexander, 2016) and in unstressed than stressed positions such that it is more likely to find vocal fry if the unstressed positions are further away from the stressed positions (Gibson, 2017).

3.2.2. Acoustic measures of voice quality

Any sound can be analyzed in both spectral and temporal domains. Table 3.1 details the acoustic measures output from VoiceSauce (Shue et al., 2011), along with the articulatory measures (obtained from EGGWorks; see Chapter 2 for a review of details) that I used in the current Chapter to probe acoustic properties of period doubling, vocal fry, and modal voice.

Table 3.1: Acoustic and articulatory measures extracted using VoiceSauce and EGGWorks. '*' indicates measures corrected for formant values and bandwidths.

	Measure	Explanation
Acoustic	H1*, H2*, H4*, H2K*	The amplitude of first, second, fourth harmonics and harmonic nearest $2000Hz$
	A1*, A2*, A3*	The amplitude of the harmonic closest to first, second, and third formants
	H1*–H2*	Difference between H1 and H2
	H2*–H4*	Difference between H2 and H4
	H1*–A1*, H1*–A2*, H1*–A3*	Difference between H1 and A1, A2, and A3
	H4*–H2K*	Difference between H4 and harmonic nearest $2000Hz$
	H2K*–H5K	Difference between harmonic nearest $2000Hz$ and that nearest $5000Hz$
	CPP	Cepstral peak prominence
	HNR05	Harmonics-to-noise ratio $< 500Hz$
	HNR15	Harmonics-to-noise ratio $< 1500Hz$
	HNR25	Harmonics-to-noise ratio $< 2500Hz$
	HNR35	Harmonics-to-noise ratio $< 3500Hz$
	SHR	Subharmonic-to-harmonic ratio
	PRAAT f0	f0 calculated from cross-correlation in PRAAT
	F1, F2, F3, F4 (snack)	First through fourth formant frequencies
	B1, B2, B3, B4 (snack)	First through fourth formant bandwidths
	Energy	Root-mean-squared energy
	SoE	Strength of excitation at glottal closure
	epoch	Instant of significant excitation/glottal closure
Articulatory	CQ	Contact quotient (threshold)
	CQ_H	Contact quotient (hybrid)
	CQ_PM	Contact quotient (derivative)
	CQ_HT	Contact quotient (Henry Tehrani method)
	peak_Vel	Cycle peak velocity value (peak increase in contact)
	peak_Vel_Time	Cycle peak velocity time
	min_Vel	Cycle minimum velocity value (peak decrease in contact)
	min_Vel_Time	Cycle minimum velocity time
	SQ2-SQ1	Duration between 10% and 90% of the closing slope (contacting duration)
	SQ4-SQ3	Duration between 10% and 90% of the opening slope (de-contacting/opening duration)
	ratio	Speed quotient: (SQ2-SQ1)/(SQ4-SQ3)

Vocal fry is famously known for its low fundamental frequency (Michel, 1968; Hollien and Michel, 1968), and based on the findings in Chapter 2, period doubling and vocal fry have similar f0s, which differ from the modal voice.

While f0 and H1–H2 are often correlated, spectral tilt measures including H1–H2 and beyond are the most widely-used family to describe voice quality characteristics. In general, the lower the spectral tilt (especially for H1–H2), the higher the degree of glottal constriction (Holmberg et al., 1995b). Studies have been using a full set or subset of spectral tilt measures ranging from low (H1–H2, H1–A1), mid (H1–A2, H1–A3, H2–H4), to high (H4–H2K, H2K–H5K) to explore the differences in categories of voice quality in the world's languages (English: Garellek and Seyfarth, 2016; Takhian Thong Chong: DiCanio, 2009; Chichimec: Kelterer and Schuppler, 2020; Yerevan Armenian: Seyfarth and Garellek, 2018; !Xóõ: Garellek, 2020; Itunyoso Trique: DiCanio, 2012; Phnom Penh Khmer: Kirby, 2014, among others), with substantially more uses of these measures in describing breathy voice (to name a few: Bickley, 1982; Hillenbrand et al., 1994; Wayland and Jongman, 2003). Also see Esposito and Khan (2020) and Garellek (2019, 2022) for overviews of these acoustic measures and the implications for their use. Notably, H1–H2 is also used in creaky voice detectors developed by Drugman et al. (2020).

Subharmonic-to-Harmonic Ratio (SHR) is a measure proposed by Sun and Xu (2002) to capture alternate cycles in speech signal, quantifying the subharmonic to harmonic ratio through spectrum shifting. Typically, the higher the SHR, the stronger the subharmonics present in the speech signal. A recent study evaluated the performance of a subharmonics detector using SHR in assessing the degree of subharmonics in speech samples and found that this metric is robust within certain adaptive ranges (Herbst, 2020). Because of the presence of subharmonics in period doubling, I expect to find a higher value there than in vocal fry and modal voice.

Harmonics-to-noise ratio or signal-to-noise ratio (HNR or SNR) and cepstral peak prominence (CPP) are measures of regularity and magnitude of harmonics above the noise floor with slightly different algorithms. Both being cepstral analysis, here HNR is calculated based on dif-

ferent frequency windows ranging from 500Hz to 3500Hz with a step size of 1000Hz, whereas CPP is calculated using all frequencies and a linear regression model of the noise floor. A lower value of these measures indicates a noisier and less periodic signal, probably due to aspiration or irregularity in f0. Though, a general caveat should be noted that the noise measures are easily influenced by the microphone or recording environment (Van der Woerd et al., 2020).

HNR was found to be useful in predicting creaky voice in low falling tones of Mandarin, Cantonese, and White Hmong (Kim et al., 2020). CPP was more commonly used to measure the acoustic properties of breathy vowels across languages (Gujarati: Esposito and Khan, 2012; White Hmong: Garellek and Esposito, 2021). HNR or CPP has also been used widely and found effective in assessing voice disorders (in continuous speech for disordered voices: Qi et al., 1999; treatment of hoarseness: Yumoto et al., 1982; benign vocal fold lesions, primary muscle tension dysphonia, vocal fold atrophy, or unilateral vocal fold paralysis before and after treatment: Gillespie et al., 2014; dysphonic vowels including breathy, hoarse, and strained voices: Wolfe and Martin, 1997; dysphonia: Mizuta et al., 2020; Sataloff et al., 2002; normal, intermittently dysphonic, or consistently dysphonic: Gaskill et al., 2017). Also see Ferrer et al. (2007) for a review of properties of CPP and its use in previous studies of voice quality measurements, and Fraile and Godino-Llorente (2014) for an analytic analysis. Using normal speaking voices, CPP was found reasonably reliable whereas HNR had a lower consistency in a study on the reliability of objective voice measures in clinical practice (Leong et al., 2013). HNR was also found useful for indexing vocal aging (Ferrand, 2002). Here I expect to find a lower value of the noise measures in the two subtypes of creaky voice than modal voice.

Energy measures include strength of excitation (SoE) and root-mean-squared energy. SoE represents "the relative amplitude of impulse-like excitation", measured at the instant of significant excitation of the glottal pulses, which typically takes place around the 'epoch', the instant of glottal closure (Murty and Yegnanarayana, 2008). SoE was found to be lower in creaky than modal voice in !Xóõ (Garellek, 2020), and higher in weakly-creaky vowels than strongly-glottalized

voice including glottal stops (Garellek and Esposito, 2021). It is thus expected that lower SoE will be seen in period doubling and vocal fry than modal voice.

Formants are relatively less-frequently used in the literature in assessing voice quality, assuming the independence between source and filter. However, it is possible that different voice source profiles induce different characteristics of formant frequencies and bandwidths. For example, Gobl (1989) found that creaky voice had stronger formants (higher amplitude relative to surrounding harmonics) and narrower formant bandwidths compared to modal and breathy voice. In addition, the relative measure between the amplitude of formants and harmonics such as A1–H1 and A2–H1 values were used by Gordon (2001) in describing the acoustics of creaky vowels in Western Apache and Hupa, and he found that the two values were larger for creaky vowels following ejectives than unaspirated stops. This suggests that creaky voice has a greater intensity at higher frequencies than lower frequencies, as expected if spectral tilt is lower, and it can be conditioned by aspiration. On the other hand, vocal tract resonances can have impacts on voice and gender perception (Bele, 2006; Levitt and Lucas, 2018; Coleman, 1971; Skuk and Schweinberger, 2014). The bitonal percept of period doubling often comes from its competing first and second harmonics which are an octave away from each other (Redi and Shattuck-Hufnagel, 2001). If a formant is closer to H2, it may boost the percept of the higher harmonic than the lower one.

These acoustic measures, taken together, serve as useful features to classify voice qualities. For example, Khan et al. (2015) found that perceived creaky voice in American English spoken by transgender men often had lower f0, HNR, and SHR, and *higher* spectral tilt measures such as H1–A1, H1–A2, and H1–A3, which would be classified as non-constricted creak.

3.2.3. Mandarin

Mandarin Chinese is a tone language with four contrastive pitch contours: Tone 1 (55, high), Tone 2 (35, rising), Tone 3 (21[4], low dipping), and Tone 4 (51, falling) (Chao, 1968). Besides

pitch, non-modal voice quality is a well-established phonetic correlate of the low-dipping Tone 3 and thus plays an important role in Mandarin. Creaky voice also appears in a number of positions in prosodic constituents (Kuang, 2018). Crucially, creaky voice largely covaries with f0 in production with extreme pitch ranges (Kuang, 2017), though exceptions have been found during coarticulation (Huang, ur). Creak is mostly seen in the lowest Tone 3, and can also occur on other tones with a low pitch target, Tones 2 and 4, implying its association with low f0 (Belotel-Grenié and Grenié, 1994, 2004; Davison, 1991; Kuang, 2017; Huang et al., 2018).

Li et al. (2020) found that one acoustic characteristics of sarcastic speech in Mandarin compared to sincere speech was the presence of multiple pulsing (shown by a higher SHR) with gender differences in the use of creak to express sarcasm. They also used a random forest classifier on nine voice quality parameters and found that SHR was most important for female and jitter (cycle-by-cycle frequency variation) for male. Further, the prosodic effect of tone and focus on creaky voice in Mandarin are found by Zheng (2006), but great individual variation exists when realizing creak.

Thus, Mandarin is selected as the target language for investigating the acoustic and distributional properties of period doubling and vocal fry. Given that period doubling and vocal fry are both common instantiations of creaky voice in Mandarin, the general expectation is that these two subtypes will not differ in terms of their locations. In other words, period doubling and vocal fry should have similar linguistic distributions – in low tones such as Tone 3 which is known to be characterized by creak, utterance finally, or in prosodic positions with low compressed f0 ranges such as post-focal positions.

3.3. Methods

3.3.1. Materials

The speech materials are from a Mandarin corpus of simultaneous audio and electroglottography (EGG) recordings collected to document a full range of tonal contextual variation in the language (Huang, ur). Although the materials were designed for another study, they adequately address the target questions here because a scripted read speech corpus with contextual variation is rich in identifying period doubling and other creaky voice types. The stimuli consist of a fixed carrier sentence with varying trisyllabic compound words: *wo3 tɕau1 nʲi3 WORD tsən3-mɤ0 ʂʷo1* 'I teach you WORD how to say.' In the compound words, each of the four Mandarin tones was flanked by varying Mandarin Tones 1–4, for a full range of $4 \times 4 \times 4 = 64$ combinations. Three sets of 64 sentences with two repetitions (384 sentences in total) were elicited per recording. The compounds are frequently-used words and familiar to native Mandarin speakers. Some words have left or right branching syntactically whereas others are fixed or metaphorical expressions. Though tone realization is not expected to be conditioned by the syntactic structure of the compound, neutral tones are occasionally seen in the corpus. All the lexical tones were labeled and neutral tones excluded based on their phonetic realization. All the materials can be found in the Appendix. Period-doubled phonation and vocal fry were identified anywhere in the phrase – during the target words or the carrier sentence, using the EGG signal, to avoid possible formant-induced interferences with the voicing signal.

3.3.2. Participants

Twenty native Mandarin speakers ($10F,\ \textit{mean age} = 20.1,\ \textit{range} = 18 - 22$) participated in the production experiment. Fifteen were from northern provinces and five from southern provinces in China. One speaker spoke a Sichuan dialect, and three speakers also spoke a Wu dialect

(Shanghainese or Yangzhounese). All speakers had moved to the United States before or for college. At the time of the experiment, the average time they have spent in the US was 3.7 years across all participants (ranging from 0-8 years). The speakers were recruited from the UCSD Psychology Subject Pool and received undergraduate course credit for their participation. No language or speech disorders were reported.

3.3.3. Procedure

Participants were recorded in a sound-attenuated booth at the Phonetics Lab at the University of California San Diego. The experiment was implemented and presented in PsychoPy (Peirce, 2007). Speakers were instructed to produce the sentences as if they were in a natural conversation to ensure the naturalistic quality of speech. After producing one sentence, they could proceed to the next sentence at their own speed. To avoid any strategy or fatigue in producing the sentences, picture fillers were used every four sentences. Participants needed to briefly describe the object that the picture showed. The experiment lasted for 40-50 minutes, and participants could take breaks during the experiment.

Audio and electroglottography (EGG) recordings were obtained simultaneously. Participants wore a Shure SM-10 head-mounted microphone and an electroglottographic collar attached with two electrodes right below their thyroid prominence. EGG was used to record the degree of contact between the vibrating vocal folds directly from the larynx. The EGG signal was recorded using a Glottal Enterprises EG2-PCS. Both audio and EGG signals were pre-amplified through a Focusrite Scarlett 8i6 preamplifier and digitized with the computer's sound card using Audacity at a $44,100Hz$ sampling rate and 32-bit float rate.

3.3.4. Analysis

The phrases and words in the audio recordings were segmented in PRAAT (Boersma and Weenink, 2022) and the EGG recordings were used to locate source pulses with period-doubled voice and vocal fry. Canonical period doubling is often characterized by sequences of two different cycles which differ in amplitude of the pulses, or length of the periods or frequency; in other words, by amplitude modulation or frequency modulation (Kreiman et al., 1993). Vocal fry is often defined by having low f0, glottal constriction, and high damping (Keating et al., 2015). Figures 3.1 shows sample period doubling located in the EGG signal and its audio waveforms. Figure 3.2 shows sample vocal fry in EGG and audio waveforms. I also extracted modal voice samples of $\sim 50ms$ from speech, which occurs in the same phrase as the instances of period-doubled voice for comparison.

The acoustic measures of the audio waveforms of both period doubling and vocal fry were extracted using a voice analysis program VoiceSauce (Shue et al., 2011) by taking measurements with a window size of 10 ms and step size of 1 ms over the tokens of period doubling, vocal fry, and modal voice identified from the EGG signal. PRAAT's f0 algorithm (Boersma and Weenink, 2022) was chosen for further acoustic measures, such as for calculation of the spectral tilt, because the default Straight algorithm within VoiceSauce has been found to best perform in optimal conditions such as with normal or typical voice. The longest period, or the length of the meta-cycle, is designated to be the fundamental period and act as the basis of calculating all other spectral measures. Using the canonical EGG waveforms as detailed in Chapter 2, there were 3784 tokens of period doubling and 1269 tokens of vocal fry identified in the corpus. An additional 1675 tokens of modal voice were selected for baseline comparison. Because EGG waveforms do not contain formant information, upon a close examination of the audio waveforms, I excluded any non-vocalic segments that were identified as one of the three voice types in the EGG recording to facilitate the processing of the acoustic measurements. And because many acoustic measurements depend on correct f0 estimation, every 1 in 10 files were checked manually and files containing

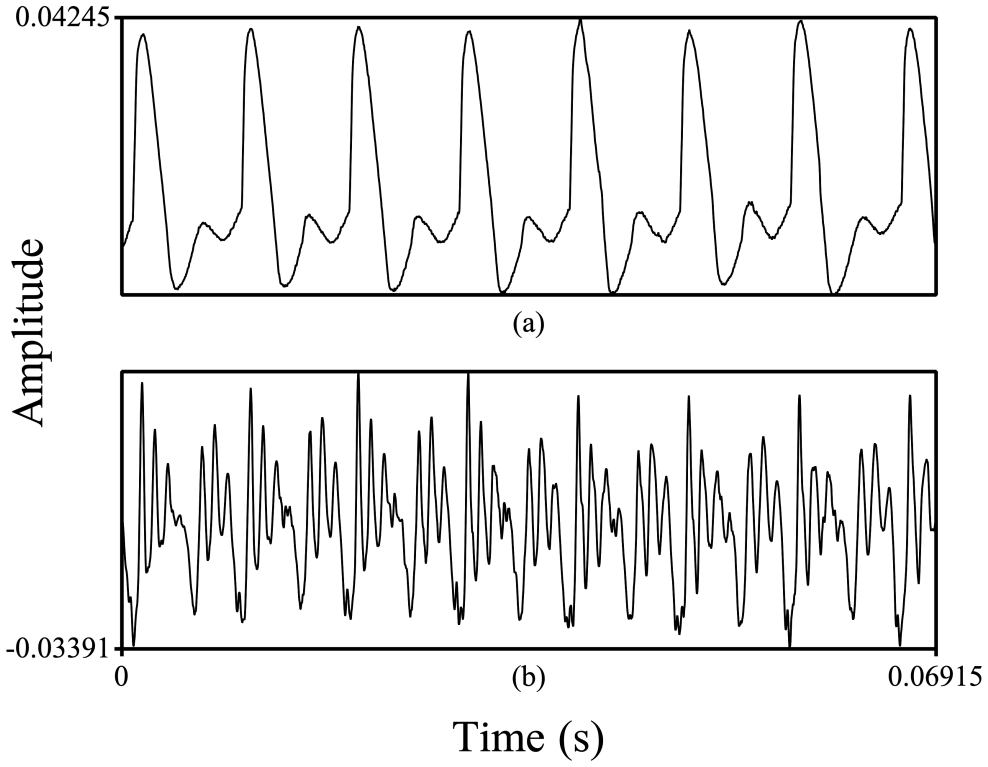


Figure 3.1: EGG waveform (a) and audio waveform (b) of period doubling.

incorrect f0 calculations were excluded (the error rate was around 3%). Also, based on the findings of the different frequencies in modal voice and period doubling in Chapter 2 (Figure 2.6), I excluded files with incorrect f0 calculations whose values fall into extremely unlikely ranges. For example, for women, files for which the f0 was larger than $150Hz$ in both period-doubled and vocal fry tokens, and smaller than $150Hz$ in modal tokens, were excluded; for men, f0 larger than $100Hz$ in both period-doubled and vocal fry tokens, and smaller than $60Hz$ or larger than $150Hz$ in modal tokens were excluded. Here I excluded 21.4% of the data from 1690 out of 7878 files. Further, I normalized the log-transformed f0 values using z-scores within each speaker and filtered out outliers deviating more than 2.5 standard deviations from the mean (another 67 files were removed, resulting in 6121 files in total). The resulting dataset have 3706 tokens of period-doubled voice, 817 tokens of vocal fry, and 1598 tokens of modal voice.

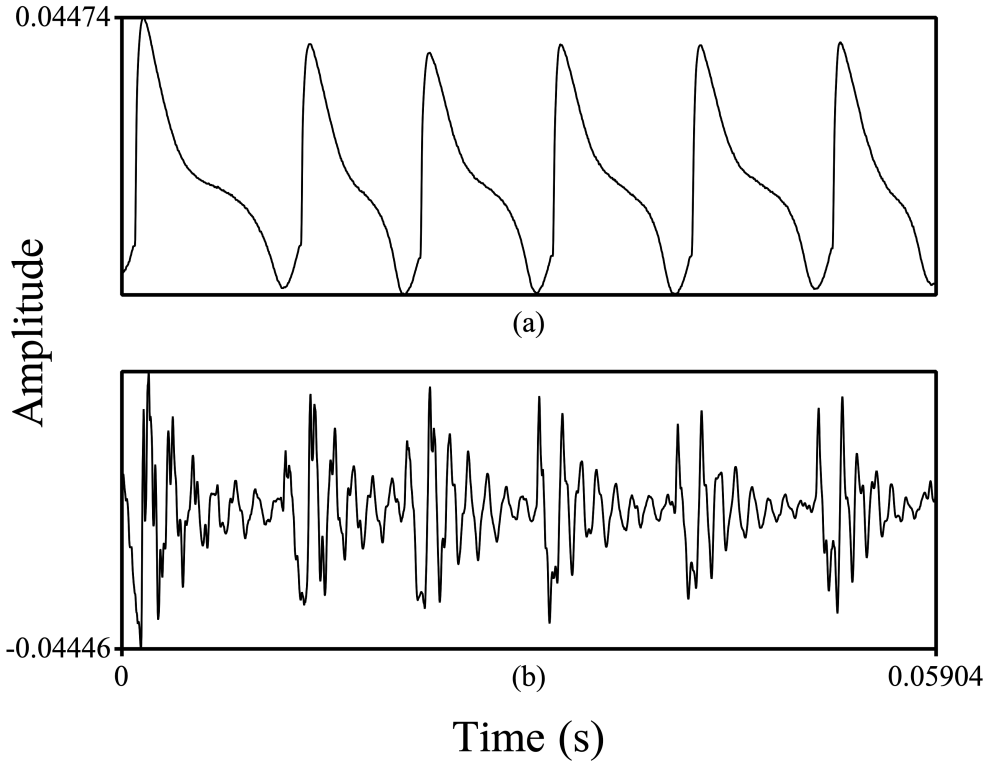


Figure 3.2: EGG waveform (a) and audio waveform (b) of vocal fry.

In the following Results section, I first present the specific acoustic measures in terms of their density distributions across three voice types. For each measure, z-score normalization is used to remove outliers by speaker. Second, I run linear mixed-effects regressions to predict how selected acoustic parameters vary as a function of voice type – how vocal fry and period doubling compare to modal voice and to each other.

3.4. Results: Acoustic analysis of PD, vocal fry, and modal voice

3.4.1. f0

Figure 3.3 shows the distribution of f0 across three voice types by gender; the separation in f0 ranges will likely result in differences in acoustic measures. Therefore, the following acoustic analyses are presented by gender. Consistent with the articulatory results, modal voice has the highest f0 while the two types of creaky voice have a clear separation from the clusters of modal voice. Despite having a similar pattern of f0 distributions across all voice types, for women, vocal fry has a lower range than period doubling; for men, vocal fry has a wider distribution and period doubling has a bimodal distribution.

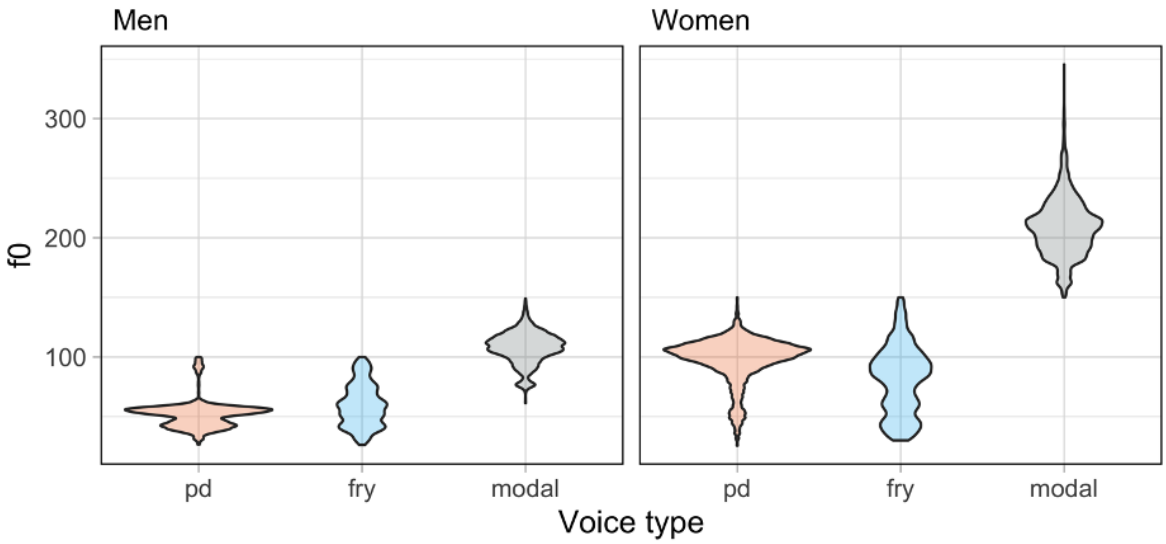


Figure 3.3: Distributions of fundamental frequency (f0) in period doubling, vocal fry, and modal voice, faceted by gender.

3.4.2. Spectral tilt measures

Given that vocal fry has the highest CQ values from the EGG signal, I expect the spectral tilt to be lowest during vocal fry, followed by period-doubled voice (though with a more variable distribution) and modal voice. For spectral tilt measures, I looked at H1*–H2*, H1*–A1*, and H1*–A2* for the lower range of the frequencies, H1*–A3* and H2*–H4* for the mid-range of the frequencies, and H4*–H2K* (the difference between H4 and the harmonic closest to $2000Hz$) and H2K*–H5K (the amplitude difference between the harmonic closest to $2000Hz$ and that to $4000Hz$) for the range of higher frequencies. H1*–H2* is known for its correlation with the glottal adduction such that a lower value indicates a shorter open phase of the glottal cycle. Thus, it is expected that H1*–H2* and H1*–A1* measures of period doubling should be higher than vocal fry because vocal fry has glottal constriction and possibly more energy in the higher frequencies which could lead to a lower spectral tilt. In the following analysis, I used separate linear mixed-effects model in both genders represented here to predict how each acoustic measure changes as a function of the type of voice (period doubling vs. vocal fry vs. modal) with the covariate of log-transformed f0, and random intercept and slope by subject, the structure of which is shown in (3.1). There are 1670 tokens from men and 4451 tokens from women; the asymmetric distribution is likely due to the higher f0 range that women employ. Because the model is run twice to be able to make all pairwise comparisons using modal voice and vocal fry as the baseline respectively, the alpha level is adjusted to $0.25 = 0.5/2$.

$$Acoustic\ measure \sim \log(f0) + type + (1 + type|subject) \tag{3.1}$$

Figures 3.4-3.6 show the patterns of lower range frequencies for modal voice, vocal fry, and period doubling. In Figure 3.4, the amplitude difference between corrected H1* and H2* clearly differentiates three voice types, with a similar pattern in both women and men. Contrary to

expectations, period-doubled voice has a lowest H1*–H2*, followed by vocal fry, and the modal voice has the highest value. This pattern holds for both genders.

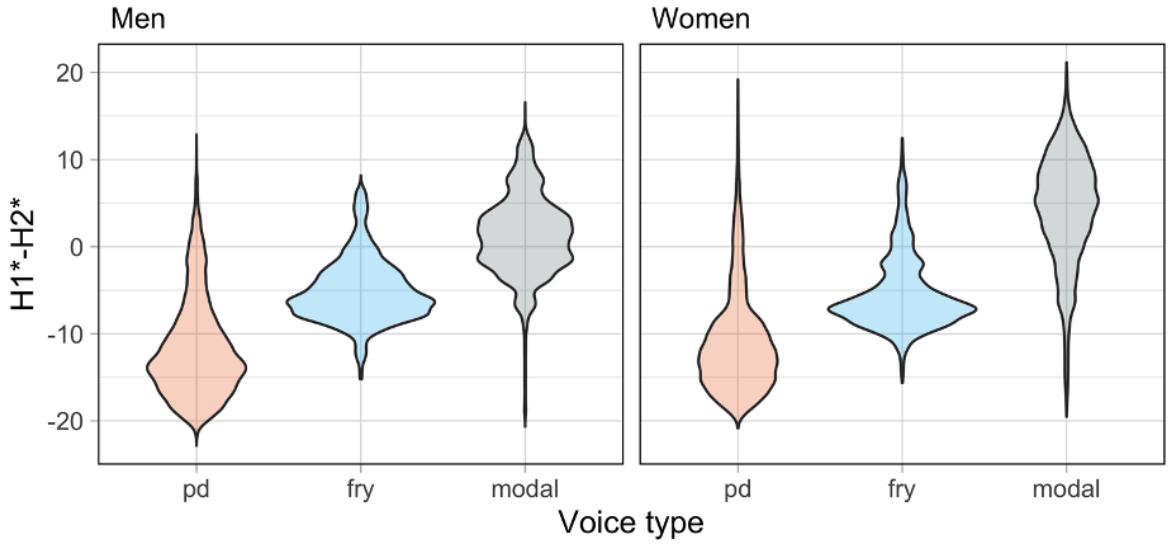


Figure 3.4: Distributions of H1*–H2* in period doubling, vocal fry, and modal voice, faceted by gender.

Table 3.2 shows the respective mean (SD) H1*–H2* values in these voice types. The linear mixed-effects model in women showed that both vocal fry ($\beta = -12.8, p < .001$) and period doubling ($\beta = -19.4, p < .001$) had lower values than modal, and period doubling was lower than vocal fry ($\beta = -6.60, p < .001$). The same model in men did not show any significant difference, probably due to a smaller sample size.

Table 3.2: Mean (SD) H1*–H2* in period doubling, vocal fry, and modal voice by gender.

Mean (SD) H1*–H2*	Period doubling	Vocal fry	Modal voice
Women	-10.9 (5.86)	-5.33 (4.29)	4.24 (6.36)
Men	-11.7 (5.64)	-5.06 (3.64)	1.32 (4.70)

H1*–A1* has a different distribution from H1*–H2* in that in women vocal fry and period doubling both have lower values than modal voice, whereas in men the distributions largely overlap, and vocal fry has a bimodal shape.

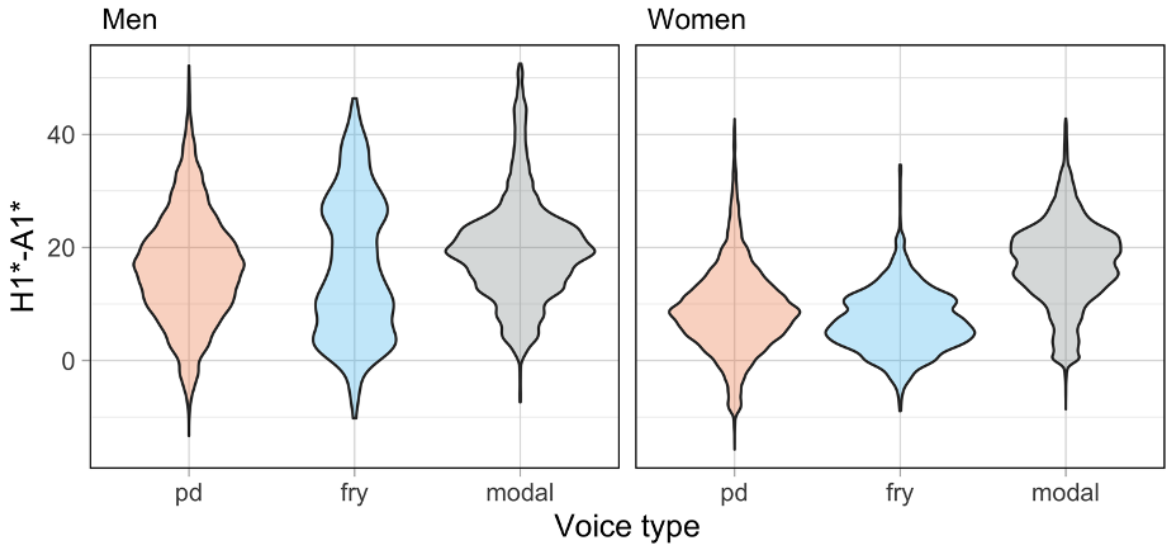


Figure 3.5: Distributions of H1*–A1* in period doubling, vocal fry, and modal voice, faceted by gender.

Table 3.3 shows the overall mean (SD) H1*–A1* values in these voice types. The linear mixed-effects model in women showed that period doubling had lower values than modal ($\beta = -5.52, p < .001$). The same model in men did not show any significant difference.

Table 3.3: Mean (SD) H1*–A1* in period doubling, vocal fry, and modal voice by gender.

Mean (SD) H1*–A1*	Period doubling	Vocal fry	Modal voice
Women	8.95 (7.46)	6.89 (5.40)	17.1 (7.51)
Men	16.5 (9.68)	16.4 (12.3)	19.6 (9.18)

H1*–A2* is interesting in that, for men, the two creak subtypes have a similar shape of bimodal distribution, whereas for women, there is more separation between period doubling and modal voice, with vocal fry in the middle.

Figure 3.6: Distributions of H1*–A2* in period doubling, vocal fry, and modal voice, faceted by gender.

Table 3.4 shows the overall mean (SD) H1*–A2* values in these voice types. The linear mixed-effects model in women showed that period doubling had lower values than modal ($\beta = -5.74, p < .001$). The same model in men showed that period doubling was higher than vocal fry ($\beta = 5.87, p < .025$). The modeled means are different from the raw means perhaps due to covariates and random effects such as f0 and speaker variation accounted for in the models.

Table 3.4: Mean (SD) H1*–A2* in period doubling, vocal fry, and modal voice by gender.

Mean (SD) H1*–A2*	Period doubling	Vocal fry	Modal voice
Women	7.89 (7.70)	7.07 (8.44)	15.4 (7.96)
Men	18.8 (10.3)	18.4 (11.1)	21.5 (8.56)

In terms of mid-range spectral magnitudes, H1*–A3* is similar to H1*–A1* in that in men vocal fry has a bimodal distribution while largely overlapping with modal and period doubling. For women, the distribution of H1*–A3* in modal is slightly more separable from vocal fry and period doubling, as shown in Figure 3.7.

Figure 3.7: Distributions of H1*–A3* in period doubling, vocal fry, and modal voice, faceted by gender.

Table 3.5 shows the overall mean (SD) H1*–A3* values in these voice types. The linear mixed-effects model in women showed that period doubling had lower values than modal ($\beta = -1.71, p < .025$). The same model in men showed that period doubling was higher than vocal fry ($\beta = 10.93, p < .01$).

Table 3.5: Mean (SD) H1*–A3* in period doubling, vocal fry, and modal voice by gender.

Mean (SD) H1*–A3*	Period doubling	Vocal fry	Modal voice
Women	-2.50 (8.82)	-3.40 (8.00)	4.40 (9.31)
Men	12.3 (14.1)	8.75 (17.1)	13.2 (13.1)

H2*–H4* has been shown to signal breathy versus modal voice quality (Garellek et al., 2013), similar to the function of H1*–H2* such that an increase in either measure would sound breathier. In Figure 3.8, H2*–H4* does not appear to be a good attribute in separating the three voice types in men, but period doubling shows a higher H2*–H4* than the other voice types in women. This complements findings with H1*–H2* that period doubling has a lowest H1*–H2*.

Figure 3.8: Distributions of H2*–H4* in period doubling, vocal fry, and modal voice, faceted by gender.

Table 3.6 shows the overall mean (SD) H2*–H4* values in these voice types. The linear mixed-effects model in women showed that both vocal fry ($\beta = 1.11, p < .001$) and period doubling ($\beta = 1.54, p < .001$) had higher values than modal, and period doubling was higher than vocal fry ($\beta = 4.32, p < .01$). The same model in men showed no difference.

Table 3.6: Mean (SD) H2*–H4* in period doubling, vocal fry, and modal voice by gender.

Mean (SD) H2*–H4*	Period doubling	Vocal fry	Modal voice
Women	6.80 (7.70)	2.14 (6.11)	1.98 (7.56)
Men	4.51 (5.14)	3.32 (5.47)	5.83 (6.72)

Then, for higher spectral-tilt measures, Figures 3.9 and 3.10 plot H4*–H2K* and H2K*–H5K across voices. It is less clear whether these higher-frequency spectral tilt measures play a role in differentiating different phonations, though Garellek (2019) suggests they have some relevance in signaling creaky versus non-creaky vowels in American English along with H1*–H2* and H2*–H4*, for example. In the current study, substantial variation is seen in H4*–H2K* values across all voice types. H4*–H2K* values do not appear to differ based on the voice types for women, but they differ for men such that period doubling concentrates more on the higher values with a bimodal distribution, similar to vocal fry, which is slightly higher than modal voice.

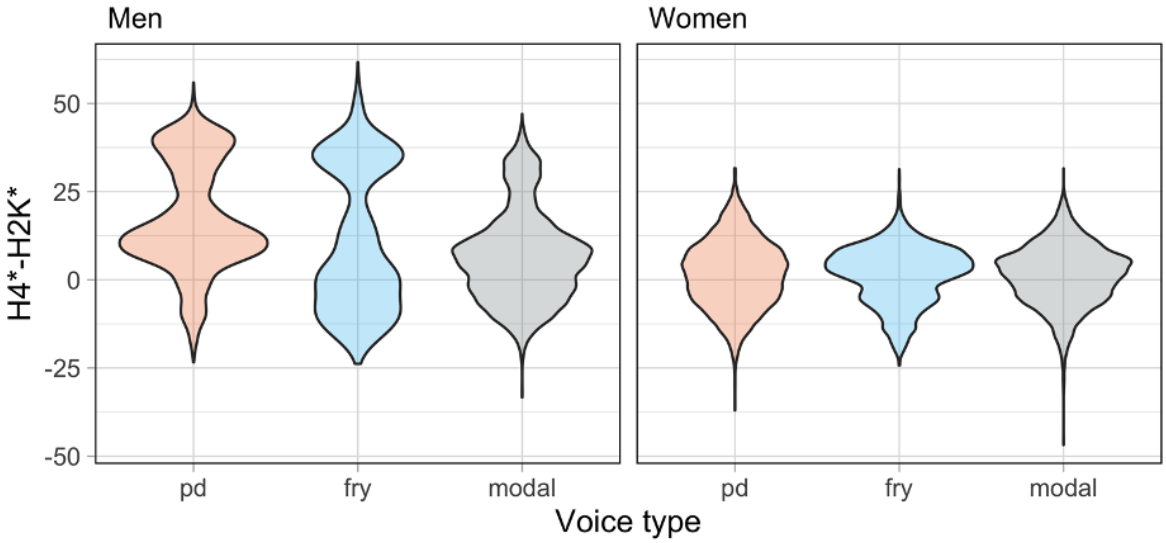


Figure 3.9: Distributions of H4*–H2K* in period doubling, vocal fry, and modal voice, faceted by gender.

Table 3.7 shows the overall mean (SD) H4*–H2K* values in these voice types. The linear mixed-effects model in women showed that period doubling had higher values than modal ($\beta = 1.88, p < .025$), potentially due to the long tail in modal voice. The same model in men showed period doubling was higher than vocal fry ($\beta = 11.65, p < .01$).

Table 3.7: Mean (SD) H4*–H2K* in period doubling, vocal fry, and modal voice by gender.

Mean (SD) H4*–H2K*	Period doubling	Vocal fry	Modal voice
Women	1.74 (9.55)	1.44 (8.16)	1.38 (8.55)
Men	17.6 (15.4)	11.3 (20.1)	6.80 (12.9)

As for H2K*–H5K, the distribution is similar to H4*–H2K*. For women, more values are concentrated in the higher range in period doubling than vocal fry and modal voice. For men, period doubling and vocal fry both have a bimodal distribution, compared to modal voice; period doubling also has a lower overall value than the other two voice types.

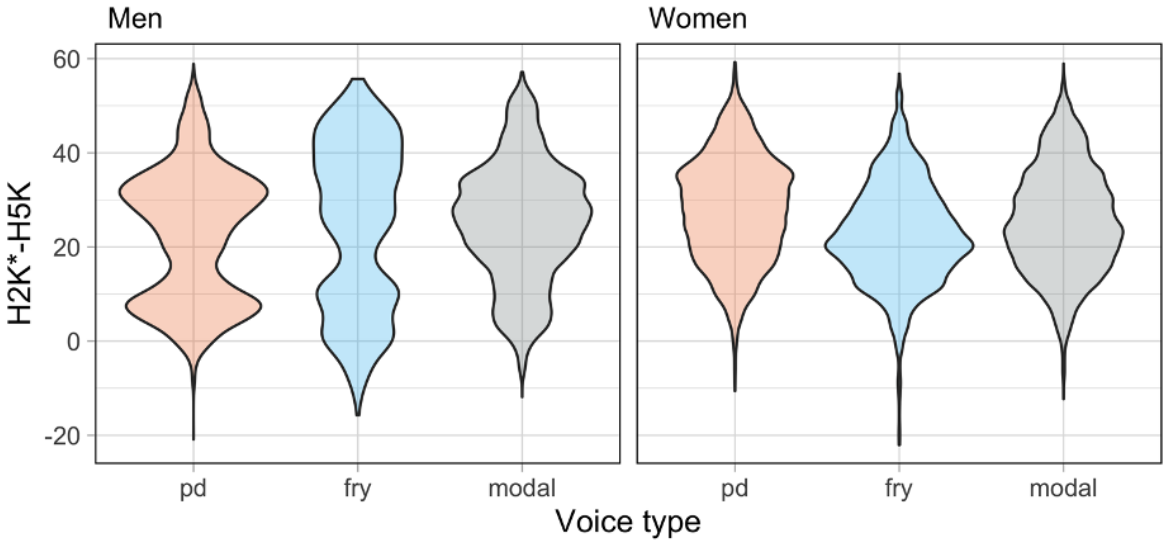


Figure 3.10: Distributions of H2K*–H5K in period doubling, vocal fry, and modal voice, faceted by gender.

Table 3.8 shows the overall mean (SD) H2K*–H5K values in these voice types. The linear mixed-effects model in women showed that period doubling had lower values than modal ($\beta = -3.94, p < .001$), which is opposite from the graph. Again, the modeled means may be different from the raw means due to f0 or speaker effects being factored out. The same model in men showed period doubling was lower than vocal fry ($\beta = -8.29, p < .01$).

Table 3.8: Mean (SD) H2K*–H5K* in period doubling, vocal fry, and modal voice by gender.

Mean (SD) H2K*–H5K	Vocal fry	Modal voice	Period doubling
Women	22.8 (10.6)	25.6 (11.2)	28.0 (11.1)
Men	24.3 (17.4)	25.2 (12.6)	22.4 (13.8)

In sum, based on the density visualization, in the lower range of harmonic slopes, H1*–H2* contributes to the largest separation for both genders, H1*–A1* and H1*–A2* also show some moderate separation at least in women. In the mid-range of harmonic slopes, H1*–A3* and H2*–H4* are more distinguishing for women whereas in the higher range of spectral slopes, both H4*–H2K* and H2K*–H5K show more clearer separation for men. Period doubling in general has lower H1*–H2*, mid H1*–A1*, H1*–A2*, and H1*–A3*, and higher H2*–H4* and H4*–H2K*, compared to vocal fry and modal voice.

Based on the results of H1*–H2*, I further plot the distributions of H1* and H2* across the three voice types, as shown in Figures 3.11 and 3.12. Given the multiple frequencies present in period doubling, the competing pitch percept is likely caused by a stronger H1 or H2 component, as reviewed in Section 2.6 of Chapter 2.

Compared to modal voice, both vocal fry and period doubling show a distribution that concentrates on a lower value, but the differences between the two creak subtypes are small.

Table 3.9 shows the overall mean (SD) H1* values in these voice types. The linear mixed-effects model in women showed that both vocal fry ($\beta = -6.76, p < .001$) and period doubling

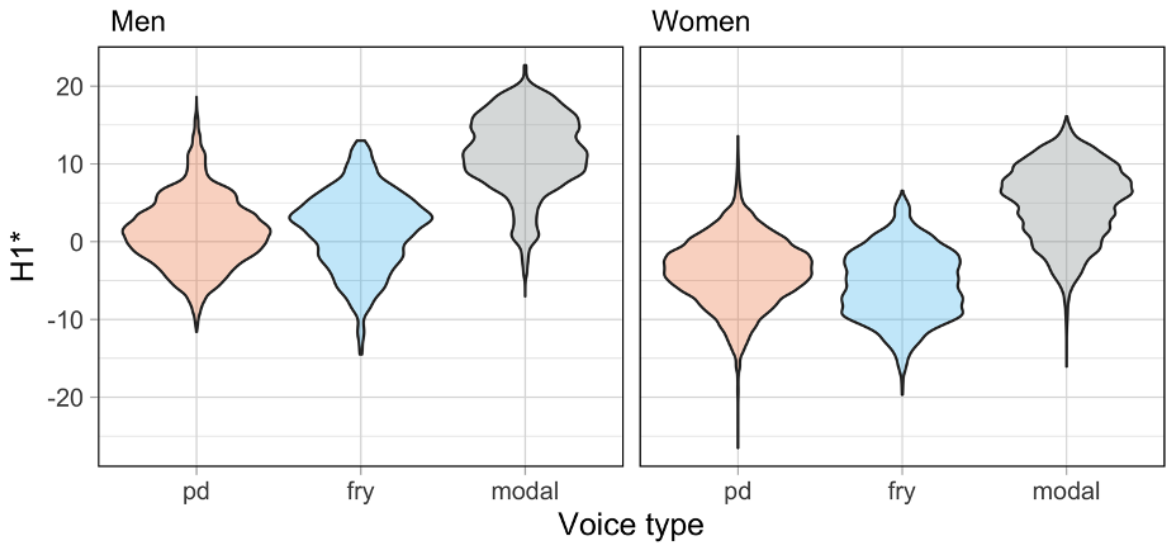


Figure 3.11: Distributions of H1* in period doubling, vocal fry, and modal voice, faceted by gender.

($\beta = -4.02, p < .001$) had lower values than modal, and period doubling was higher than vocal fry ($\beta = 2.74, p < .01$). The same model in men also showed that both vocal fry ($\beta = -4.70, p < .001$) and period doubling ($\beta = -3.13, p < .01$) had lower values than modal.

Table 3.9: Mean (SD) H1* in period doubling, vocal fry, and modal voice by gender.

Mean (SD) H1*	Period doubling	Vocal fry	Modal voice
Women	-3.80 (4.23)	-5.52 (4.40)	5.03 (4.76)
Men	1.32 (4.43)	1.52 (4.94)	11.4 (4.90)

H2* well distinguishes period doubling, in particular for women, compared to the other two voice types. PD has the highest value of H2*, which is consistent with the striking pattern of a prominent H2 and the subharmonics in the EGG spectra (see Chapter 2). Thus, the fact that period doubling has a lowest H1*–H2* value is a result of a lower H1 which is generally

associated with creaky voice, and more crucially, a higher H2, comparable to the original H1 in modal voice.

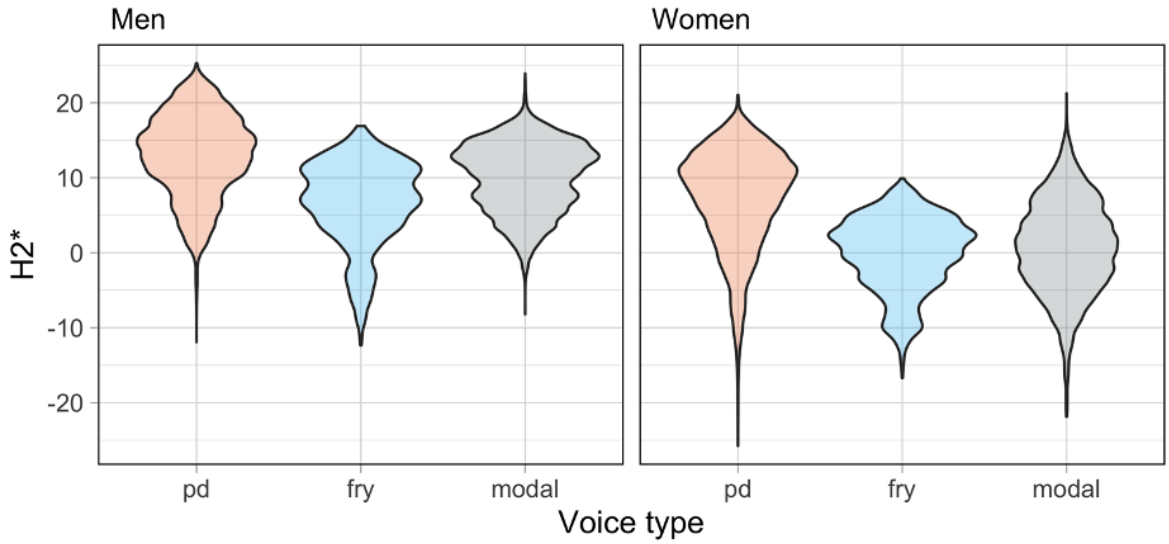


Figure 3.12: Distributions of H2* in period doubling, vocal fry, and modal voice, faceted by gender.

Table 3.10 shows the overall mean (SD) H2* values in these voice types. The linear mixed-effects model in women showed that both vocal fry ($\beta = 6.15, p < .01$) and period doubling ($\beta = 1.54, p < .001$) had higher values than modal, and period doubling was higher than vocal fry ($\beta = 9.22, p < .001$). The same model in men showed that vocal fry had lower values than modal ($\beta = -5.14, p < .01$).

Table 3.10: Mean (SD) H2* in period doubling, vocal fry, and modal voice by gender.

Mean (SD) H2*	Period doubling	Vocal fry	Modal voice
Women	7.11 (6.64)	-0.189 (4.93)	0.794 (6.15)
Men	13.0 (5.50)	6.52 (5.64)	10.1 (4.55)

To investigate the correlation of H1*–H2* and the contact quotient, I used files that contain complete data from both the audio and EGG recordings. There were 1107 files (vocal fry: 196 tokens; modal voice: 301 tokens; period doubling: 610 tokens). The correlation crossing four types of CQ and H1*–H2* is shown in Table 3.11. For the CQ (Henry Tehrani) method, files that contain a zero were excluded before calculating the correlation. Surprisingly, no strong correlation was found in any of the CQ measures with H1*–H2*. Likewise, previous studies showed moderate correlations between OQ/CQ and H1*–H2*, though they were still better than other correlations between OQ and f0 or other spectral tilt measures such as H1*–A3* (DiCanio, 2009; Kreiman et al., 2012). Figure 3.13 plots CQ (hybrid) and H1*–H2* against each other, showing a clear separation of the three voice types: period doubling is characterized by a lowest H1*–H2* and mid CQ, vocal fry is characterized by a high CQ and mid to low H1*–H2*, and modal voice has a higher H1*–H2* and mid CQ.

Table 3.11: Correlation across CQ measures and H1*–H2*.

	CQ (derivative)	CQ (threshold)	CQ (hybrid)	CQ (HT)
Period doubling	-0.21	0.18	0.11	0.26
Vocal fry	-0.10	-0.10	-0.10	0.02
Modal voice	-0.05	0.22	0.27	0.25

3.4.3. Periodicity measures

To probe the extent of noise and its relative strength compared to the periodic voicing harmonics, I look at harmonics-to-noise ratio (HNR) $<500Hz$ as a measure of noise around f0 and cepstral peak prominence (CPP; Hillenbrand et al., 1994) as a measure of noise over the entire spectrum. Both measures rely on cepstral analysis: HNR $<500Hz$ uses the frequency window of $500Hz$, and CPP uses all frequencies and measures the cepstral peak relative to a linear regression model of the noise floor rather than the average. It is expected that the two creaky voice subtypes have lower HNR and CPP values than the modal voice.

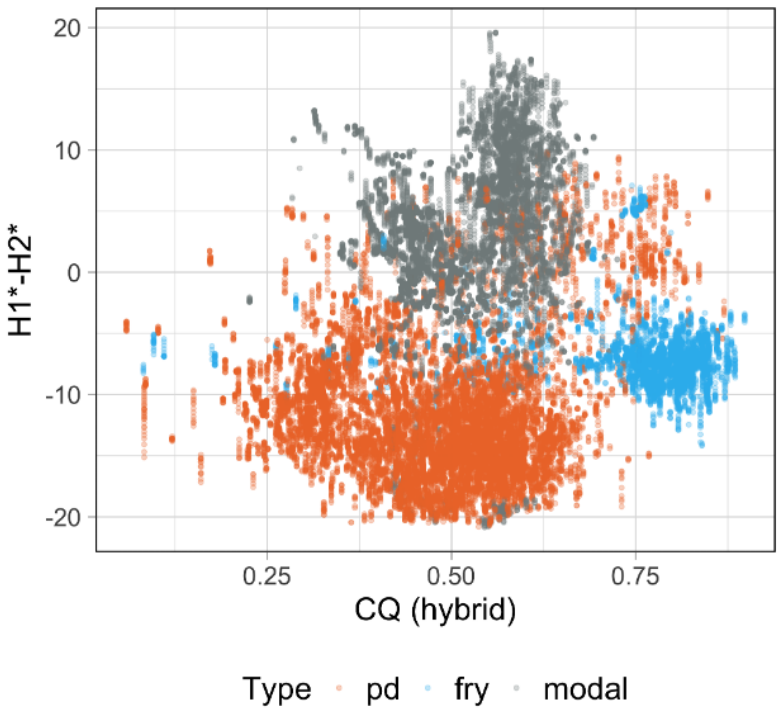


Figure 3.13: Scatter plot of CQ (hybrid) and H1*–H2* values varying by voice types.

Figure 3.14 shows that vocal fry and period doubling both have a lower overall mean HNR $<500Hz$, which is clearly separable from modal. Period doubling largely overlaps with vocal fry, particularly for men, whereas in women, period doubling spans a larger range with more variance.

Table 3.12 shows the overall mean (SD) HNR ($<500Hz$) values in these voice types. The linear mixed-effects model in women showed that both vocal fry ($\beta = -8.03, p < .001$) and period doubling ($\beta = -6.27, p < .001$) had lower values than modal. The same model in men did not show any significant differences.

Table 3.12: Mean (SD) HNR ($<500Hz$) in period doubling, vocal fry, and modal voice by gender.

Mean (SD) HNR ($<500Hz$)	Period doubling	Vocal fry	Modal voice
Women	16.7 (6.98)	13.3 (4.55)	29.6 (7.98)
Men	5.61 (4.59)	6.53 (4.88)	13.4 (6.22)

Figure 3.14: Distributions of HNR ($<500Hz$) in period doubling, vocal fry, and modal voice, faceted by gender.

Figure 3.15 shows that CPP less clearly differentiates the three voice types in both genders. However, modal voice tends to have higher values than the two creaky voice subtypes, as expected.

Table 3.13 shows the overall mean (SD) CPP values in these voice types. The linear mixed-effects model in women showed that both vocal fry ($\beta = -2.08, p < .001$) and period doubling ($\beta = -1.69, p < .001$) had lower values than modal. The same model in men showed that period doubling had lower values than modal ($\beta = -1.23, p < .01$).

Table 3.13: Mean (SD) CPP in period doubling, vocal fry, and modal voice by gender.

Mean (SD) CPP	Period doubling	Vocal fry	Modal voice
Women	19.4 (2.36)	19.6 (2.35)	21.2 (2.34)
Men	18.8 (2.10)	18.8 (1.91)	20.4 (2.36)

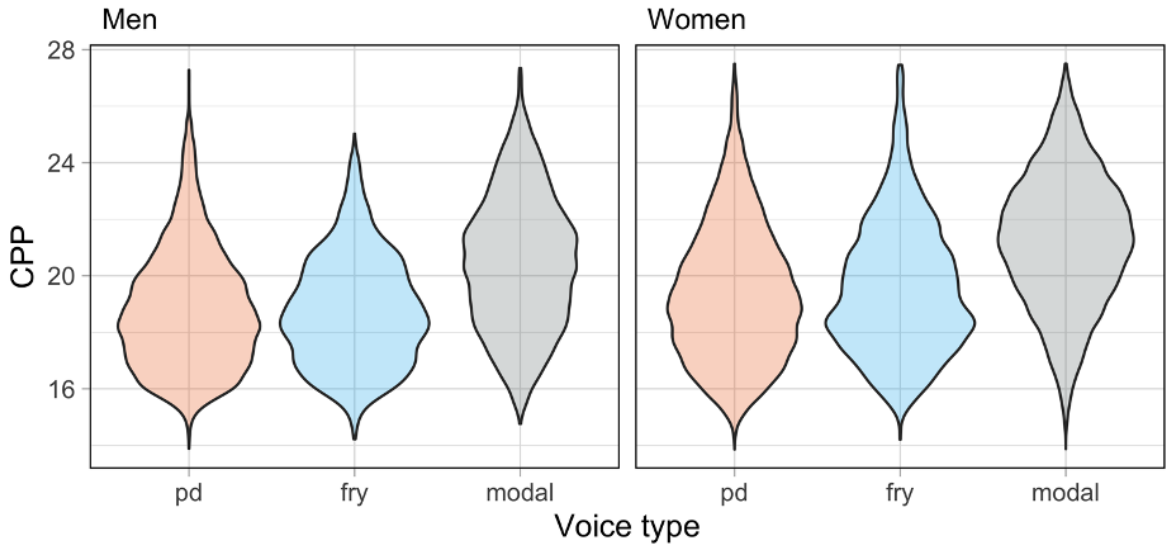


Figure 3.15: Distributions of CPP in period doubling, vocal fry, and modal voice, faceted by gender.

Thus, the noise and periodicity measures well differentiate creaky voice as a group from modal voice without further distinguishing the two subtypes. As expected, noise is one of the representative characteristics of creaky voice.

Next, I show the violin plots of subharmonic-to-harmonic ratio (SHR). SHR is defined as the ratio of the magnitude of subharmonics with respect to harmonics, which reflects the degree of deviation of a type of voice from modal voice (Sun and Xu, 2002). I expect that the SHR of period doubling should be higher than vocal fry given that vocal fry only has one regular set of harmonics. Thus, vocal fry should have fewer (and weaker) subharmonics in the spectrum. To ensure representative tokens with a correct SHR value, I trimmed SHR whose f0 values are less than 1.5 times of the PRAAT f0. Figure 3.16 shows the different distributions of SHR in vocal fry, period doubling, and modal voice. It is predicted that modal voice has fewer subharmonics for a strong and stable fundamental frequency – the harmonics. Period doubling then is expected to have a larger value of SHR given the presence of subharmonics. However, contrary to the

expectation, in women, the distribution of SHR in period doubling largely overlaps with that of vocal fry, though modal voice shows a clearer separation from the two subtypes of creaky voice by having most values concentrated in the lower range. In men, period doubling shows a larger concentration of higher SHR values than vocal fry and modal voice.

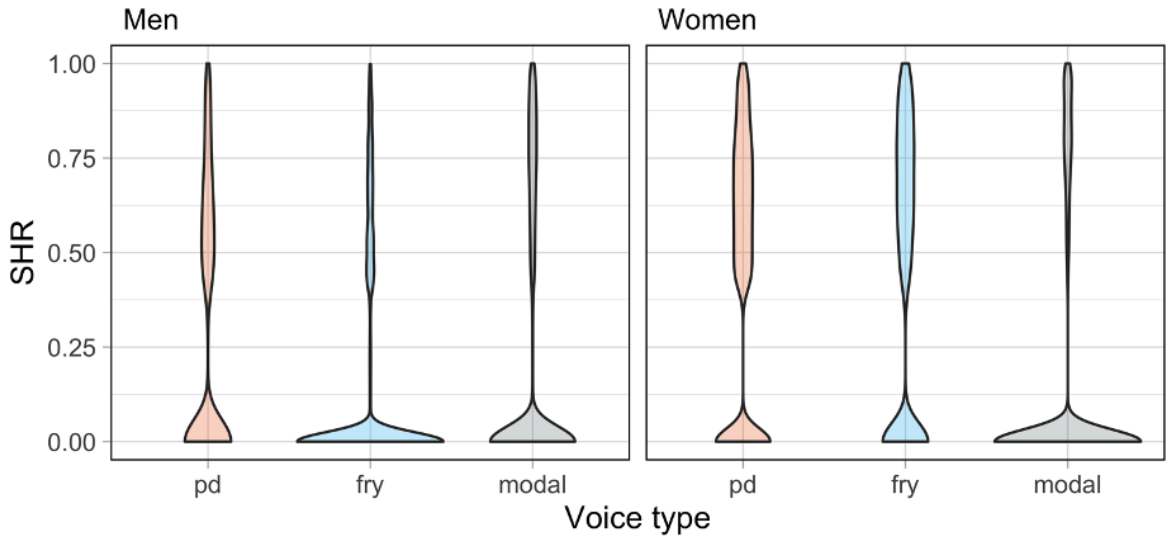


Figure 3.16: Distributions of SHR in period doubling, vocal fry, and modal voice, faceted by gender.

Table 3.14 shows the overall mean (SD) SHR values in these voice types. The linear mixed-effects model in women showed that both vocal fry ($\beta = 20.4, p < .01$) and period doubling ($\beta = 28.3, p < .001$) had higher values than modal. The same model in men did not show differences.

Table 3.14: Mean (SD) SHR in period doubling, vocal fry, and modal voice by gender.

Mean (SD) SHR	Period doubling	Vocal fry	Modal voice
Women	0.456 (0.343)	0.440 (0.367)	0.158 (0.322)
Men	0.289 (0.338)	0.156 (0.283)	0.222 (0.347)

Contrary to expectations, period doubling was not found to have a higher mean SHR than vocal fry statistically (though a trend may be inferred from visualization); this is similar to the results of the above two noise measures that distinguish the two types of creaky voice from modal voice, with no further distinctions between creak subtypes.

3.4.4. Energy

Lastly, I compare the distributions of strength of excitation (SoE) and energy over the spectra of three voice types, as shown in Figures 3.17 and 3.18. It is expected that modal voice will have the highest strength of excitation and energy throughout the spectra because modal voice typically does not have damped voicing anywhere in its glottal pulses.

The distributional patterns of SoE across these voice types are similar in both genders in that vocal fry is more separated from modal voice with a narrower concentration around lower values. Both period doubling and modal voice span a larger range of variation.

Table 3.15 shows the overall mean (SD) SoE values in these voice types. The dependent variable is scaled up by multiplying by 1000 to avoid singularity and precision issues in R. The linear mixed-effects model in both genders showed that vocal fry had lower values than modal (in women: $\beta = -21.99, p < .001$; in men: $\beta = -6.56, p < .025$) and period doubling (in women: $\beta = -17.97, p < .001$; in men: $\beta = -4.60, p < .01$).

Table 3.15: Mean (SD) SoE in period doubling, vocal fry, and modal voice by gender.

Mean (SD) SoE		Period doubling Vocal fry	Modal voice
Women	0.0442 (0.0223)	0.0187 (0.0120)	0.0635 (0.0252)
Men	0.0156 (0. 00978)	0.00844 (0.00593)	0.234 (0.0119)

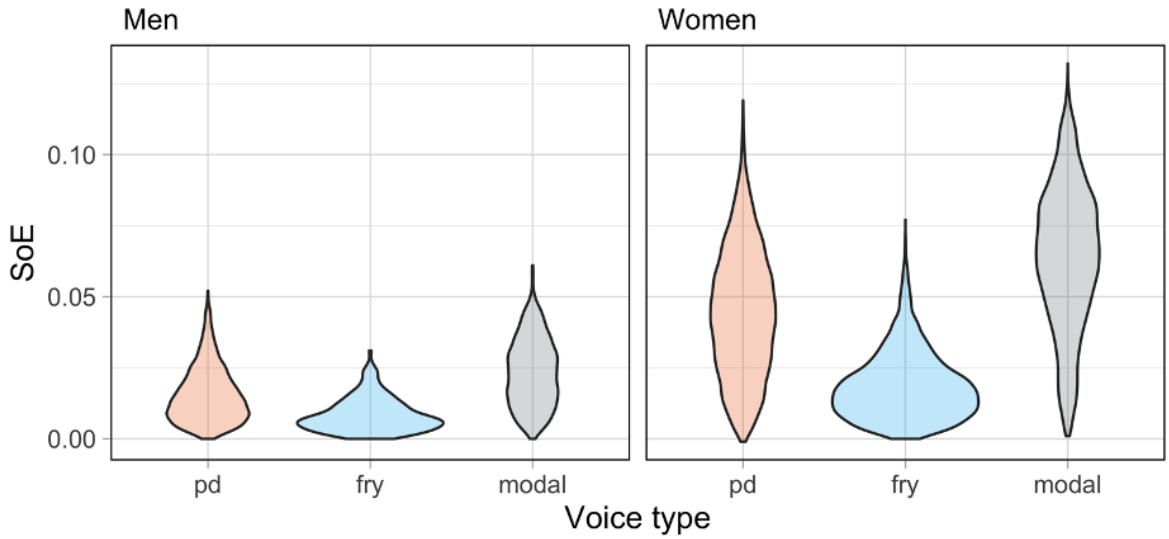


Figure 3.17: Distributions of SoE in period doubling, vocal fry, and modal voice, faceted by gender.

As for root-mean-squared energy, the three voice types in the original distribution are right-skewed with long tails. The data were thus log-transformed. Vocal fry again has overall lower values than period doubling and modal voice.

Table 3.16 shows the overall mean (SD) energy (log-transformed) values in these voice types. The linear mixed-effects model in both genders showed that both period doubling (in women: $\beta = -0.23, p < .001$; in men: $\beta = -0.29, p < .01$) and vocal fry (in women: $\beta = -0.87, p < .001; \beta = -0.42, p < .025$) had lower values than modal. In addition, in women, period doubling had higher values than vocal fry ($\beta = 0.64, p < .001$).

Table 3.16: Mean (SD) Energy in period doubling, vocal fry, and modal voice by gender.

Mean (SD) Energy (log-transformed)	Period doubling	Vocal fry	Modal voice
Women	-0.813 (0.496)	-1.10 (0.469)	-0.898 (0.523)
Men	-0.634 (0.395)	-0.846 (0.420)	-0.497 (0.402)

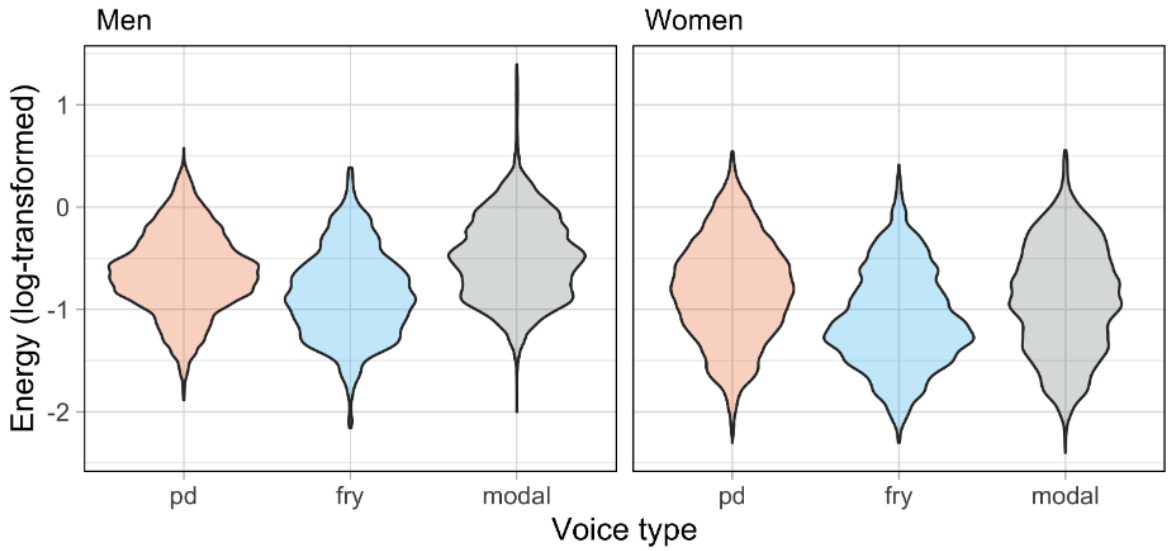


Figure 3.18: Distributions of root-mean-squared energy (log-transformed) in period doubling, vocal fry, and modal voice, faceted by gender.

Together, period doubling seems to have higher energy measures than vocal fry and tends to be lower than modal voice for both SoE and root-mean-squared energy. Both measures indicate that vocal fry has the lowest energy, which is expected due the presence of damping between pulses of vocal fry. The energy measures are thus a better predictor to distinguish the three voice types.

Table 3.17 summarizes all the individual acoustic features discussed in this section, demonstrating the presence or absence of significant links across voice types by gender. Probably due to a larger sample size, all measures in women show significant differences between at least two voice types. In particular, H1*–H2*, H1*, H2*, and H2*–H4* in women show a three-way separation. More measures do not show significance in men, and several measures distinguish two voice types among three: H1*, H2*, H1*–A2*, H1*–A3*, H4*–H2K*, H2K*–H5K*, CPP, and SoE.

Table 3.17: Summary of acoustic measures across three voice types by gender. '0' represents baseline; '−−, −, +, ++' represent sign and extent of differences from the baseline and pairwise comparisons; '/' represents no difference in any pairwise comparisons. For example, in women, H1*–H2* in period doubling is the lowest (−−), followed by vocal fry (−), compared to modal voice (0). Bolded measures demonstrate a three-way distinction.

Women				Men		
Vocal fry	Modal voice	Period doubling	Measure	Vocal fry	Modal voice	Period doubling
−	0	−−	**H1*–H2***	/	/	/
−−	0	−	**H1***	−	0	−
+	0	++	**H2***	−	0	/
/	0	−	H1*–A1*	/	/	/
/	0	−	H1*–A2*	0	/	+
/	0	−	H1*–A3*	0	/	+
+	0	++	**H2*–H4***	/	/	/
/	0	+	H4*–H2K*	0	/	+
/	0	−	H2K*–H5K	0	/	−
−	0	−	HNR	/	/	/
−	0	−	CPP	/	0	−
+	0	+	SHR	/	/	/
−	0	0	SoE	−	0	0
−−	0	−	Energy (log)	−	0	−

3.5. Computational classification

In this section, the three voice types are classified using different machine-learning approaches to quantify the importance of the acoustic and articulatory features, and predict how these features differentiate modal and the two creaky voice subtypes, vocal fry and period doubling.

3.5.1. Machine learning methods and analysis

Because of the exploratory nature of the analysis among three voice types, I first use t-distributed stochastic neighbor embedding (t-SNE), a dimensionality reduction technique, to compare the similarity or dissimilarity among all tokens in high-dimensional datasets (Van der Maaten and

Hinton, 2008). If clear separations of clusters are seen, it suggests that *a priori* labels from human production data can be recognized and generalized by machines. I can then apply machine classification methods to further explore the existing patterns and categories of the voice types in question.

Regarding the specific statistical and machine learning approaches, given that the dependent variable is categorical, I first use logistic regression to model the contributions of individual features in predicting the voice types. However, considering the high dimensionality and sparsity in the dataset – larger number of predictors and fewer rows of observations in the dataset, a simple logistic regression will not converge and will likely result in overfitting issues; that is, if the dependent variable is explained by too many independent variables that are potentially correlated and have multicollinearity, substantial variance across models will be seen as a result and thus the model will lose any generalized power. Thus, I employ a regularization technique that trades bias for variance to improve the deviance and misclassification error of logistic regression models. I chose Lasso regularization which has the potential to shrink coefficients of predictors to zero through a penalty term (indicated by alpha $= 1$ in the specification in the `glmnet` library in R), to avoid multicollinearity and reduce overfitting issues (James et al., 2013). This way Lasso also enables variable selection by yielding sparse models that involve only a subset of the predictors. Another model parameter, λ, which devises the amount of misclassification error, was determined by cross validation (Browne, 2000), a resampling method that uses some portion of the data to train model parameters, and then uses the best model generated to fit the rest of the data, with multiple iterations.

Cross validation was performed by spliting the dataset into a training set ($\sim 33.3\%$) and test set ($\sim 66.7\%$) using the *sample*() function in R. The choice of this split is to favor the predictive power of the model to generalize with sparse data. In a real world, labeled data are fewer than unlabeled ones. Thus, choosing a split with a smaller training set may be viewed as an effort to mimic future classification problems. In fact, a reverse split was also tested, and

the accuracy, precision, and recall scores were comparable for these datasets in this chapter. The models were run on the training set using the *cv.glmnet*() function in the `glmnet` library in R, and finalized using the *glmnet*() function with the λ that produced the minimum classification error among the three voice types, and then being tested in the test set using the *predict*() function. The distributions of the three voice types are shown in Table 3.18. The training set and the test set thus have similar distributions of the number of different voice types.

Table 3.18: Distributions of period doubling, vocal fry, and modal voice in training and test sets.

		Period doubling	Vocal fry	Modal voice
Acoustic	Training	1086	220	540
	Test	2211	418	1063
	Total	3297	638	1603
Acoustic + articulatory	Training	133	62	111
	Test	282	117	213
	Total	415	179	324

Second, I use random forest (Breiman, 2001), an ensemble method that aggregates decision trees to devise machine classification boundaries. Since the model makes use of a combination of decision trees, it is generally assumed that the majority of the outcomes of the decision trees leads to more confidence in classification than single decision trees. It also allows for non-linear decision boundaries, handles class imbalance better, and can be intuitively used for visualization and calculation of feature importance. Thus, the advantage of the random forest method lies in its better interpretability of the variable importance and the predictions.

Third, I use support vector machine (SVM) with radial basis function (RBF) kernel which produces non-linear decision boundaries. For a discussion of the relationship between logistic regression and SVM, see James et al. (2013). SVM has been found to be robust to individual observations, and tolerate incorrect classifications in the training set to trade for generalization in the test set (James et al., 2013).

Thus, I compare logistic regression, random forest, and SVM with RBF kernel by evaluating their model performances in terms of accuracy, precision, and recall scores, to be detailed in the following results section.

I use two datasets to evaluate the machine classification algorithms: the first one includes only acoustic features and the second includes both acoustic and articulatory features. The acoustic dataset consists of temporal and spectral measures of voice quality extracted from VoiceSauce, and the combined acoustics and articulation dataset also contains the phonatory measures obtained from EGGWorks in addition to the acoustic features (see Table 3.1).

Recall that every acoustic measure was taken at a step size of 1ms, so the length of the measures of each observation depends on the length of the file. Because the temporal dynamics of each of the measures are not of interest, I took the mean of each measure for each single token for temporal normalization, and then use the built-in function *scale*() in R to standardize the values to a distribution around a mean of 0 with a standard deviation of 1 to avoid individual or unit variations. Each row corresponds to one token. Uncorrected formants and bandwidths were excluded from the predictors. Based on the differential observations between women and men, gender is included as an independent variable to explain any incurred variance attributable to different sexes, coded binarily as 1 (men) or 0 (women). Among the various f0s and formants output by VoiceSauce, only the f0 calculated based on the PRAAT algorithm and the formants and bandwidths calculated using Snack (Sjölander, 2004; http://www.speech.kth.se/snack/) were retained. To recapitulate, Table 3.1 shows the finalized acoustic and articulatory measures used in assessing different machine classification methods, with detailed explanation. In the following subsections, I discuss logistic regression with Lasso regularization, random forest, and SVM. I used `glmnet` library for the logistic regression with Lasso regularization, `randomForest` for random forest, and `e1071` for SVM in R.

3.5.2. Results: Initial visualization using t-SNE clustering

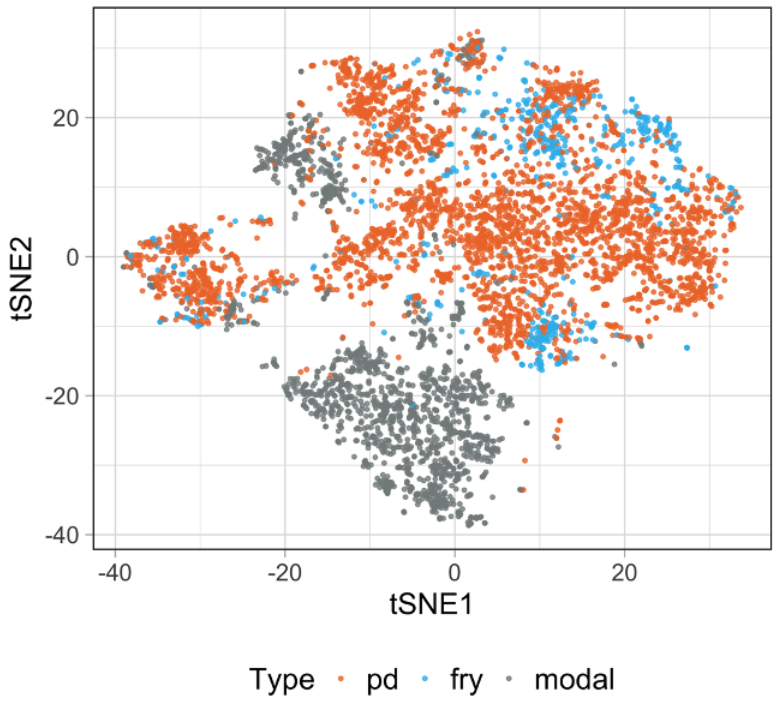
To explore any generalizable patterns and potential clustering of tokens in period doubling, vocal fry, and modal voice, I first employed t-distributed stochastic neighbor embedding (t-SNE) to reduce the dimensionality in the acoustic dataset (5538 rows x 32 columns; each row corresponds to a unique file, and each column corresponds to an acoustic or articulatory measure), using the mean values of all the available acoustic parameters. Gender is not included as a dimension but analyzed as meta information instead. I used the *Rtsne*() function with perplexity = 50 and initial dimension = 26 in the `Rtsne` library.

Despite an imbalanced dataset, different clusters can be detected from the t-SNE visualization, as shown in Figure 3.19. The number of tokens of each voice type in women and men are also shown in Table 3.19. In Figure 3.19a, clusters of modal voice are concentrated in the bottom-center, whereas the two subtypes of creaky voice spread from the top to middle areas of the space. Vocal fry and period doubling are more similar to each other than to modal voice, based on the visualization. Even the two voices seemingly pattern together, small clusters of vocal fry tokens are seen occasionally, in sparse areas less occupied by tokens of period doubling. But, this could be due to an imbalance in the dataset, and it is expected that more separation will be seen if more data were available. Consistent with the density patterns of each individual acoustic measures, period doubling occupies the space somewhere in between the majority of modal voice and that of non-modal phonation: it shares the characteristics of both voices, being more similar to modal voice while maintaining its similarity with vocal fry; vocal fry is less similar to modal voice and more similar to period doubling. Small clusters are observable in Figure 3.19a, I further plotted the voice clusters by gender, as shown in Figure 3.19b. Specifically, in men's production, period doubling and vocal fry pattern together, though with a seemingly bimodal distribution, separated from modal voice. In women, both creak subtypes also overlap and exhibit a unimodal distribution separated from modal voice.

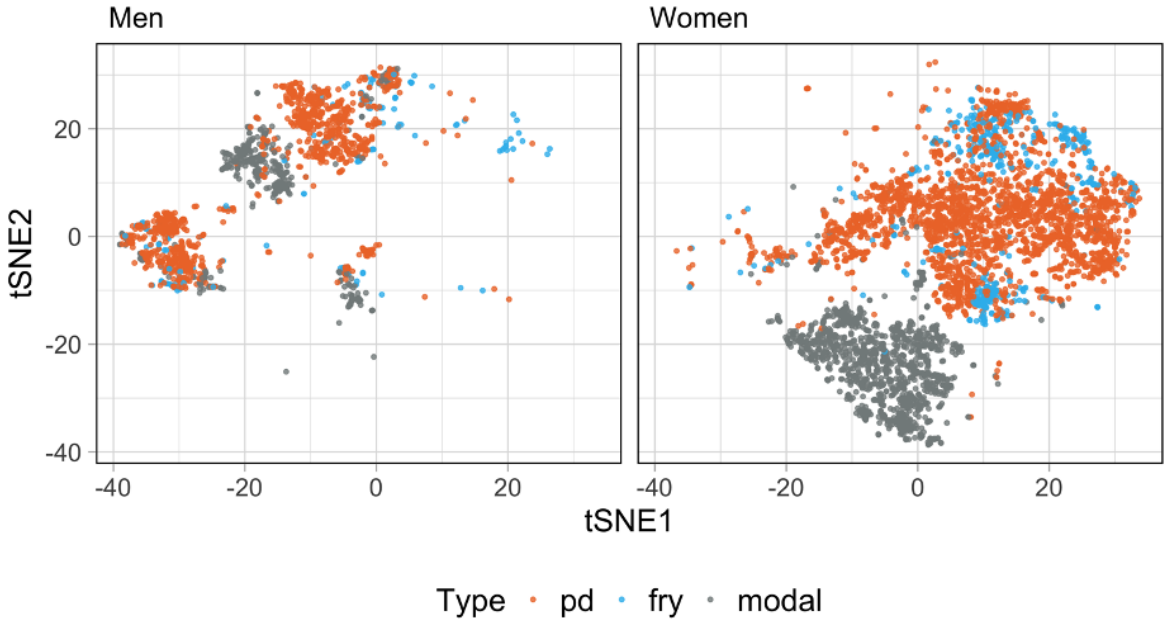
Table 3.19: Distributions of PD, vocal fry, and modal voice in women and men in the acoustic dataset.

	Period doubling	Vocal fry	Modal voice
Women	2354	482	1175
Men	943	156	428
Total	3297	638	1603

To investigate the importance of acoustic and articulatory features in clustering voice types, I also plot the t-SNE graph based on both acoustic and articulatory measures. The files that contain both acoustic and EGG measures are fewer in this dataset (918 rows x 43 columns; each row corresponds to a unique file, and each column corresponds to an acoustic or articulatory measure), after excluding any miscalculation of the acoustic or EGG measures. In Figure 3.20, despite the smaller dataset and different distributions across genders, we observe a clearer separated pattern among these voice types when using both acoustic and articulatory features than only acoustic measures to structure the space based on similarity or dissimilarity of all the tokens. Generally speaking, modal voice occupies the bottom left to middle feature space, period doubling concentrates on the middle center, and vocal fry clusters at the top right corner. The number of tokens in each voice type in women and men are shown in Table 3.20.

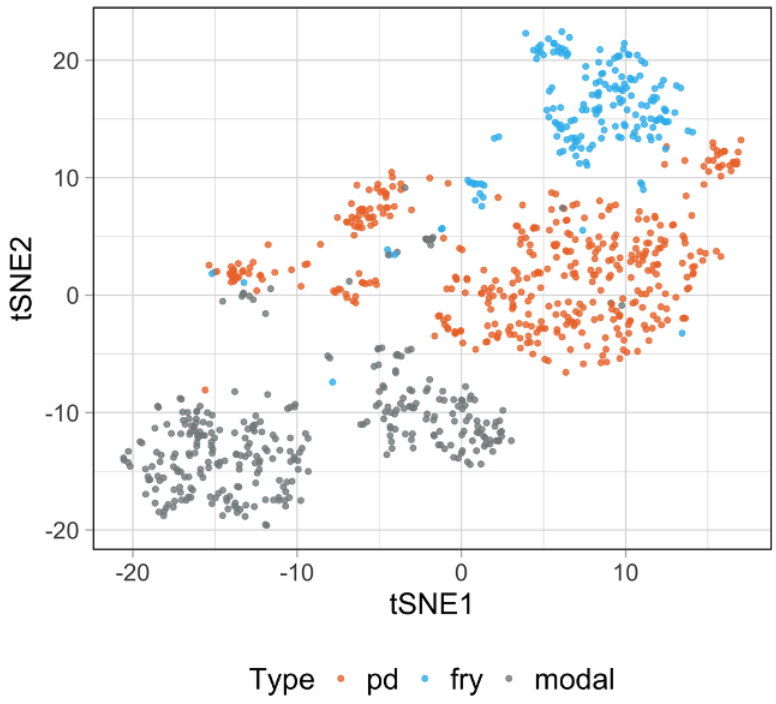


(a)

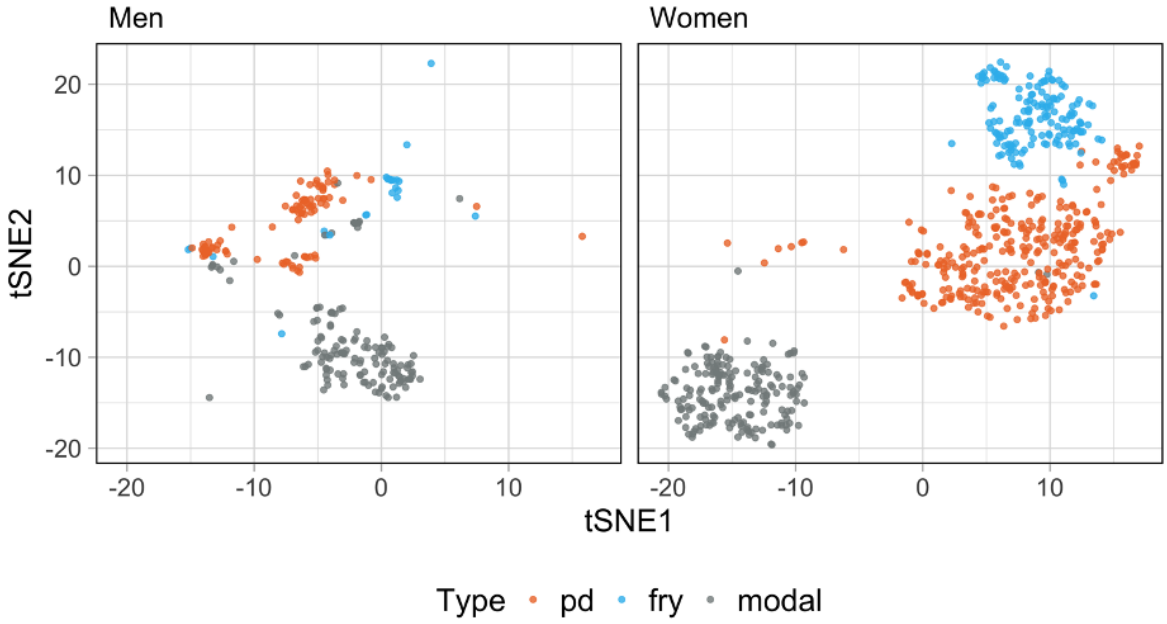


(b)

Figure 3.19: t-SNE visualization of tokens of period doubling, vocal fry, and modal voice using all the acoustic parameters as detailed in Table 3.1 and Section 3.4. The top panel plots the overall clustering based on voice types; the bottom panel plots the clustering faceted by gender.



(a)



(b)

Figure 3.20: t-SNE visualization of tokens of period doubling, vocal fry, and modal voice using all the acoustic and articulatory parameters as detailed in Table 3.1 and Sections 3.4 and 2.5. The top panel plots the overall clustering based on voice types; the bottom panel plots the clustering faceted by gender.

However, the two dimensions in t-SNE are not explicit, thus impossible to interpret the exact contribution of each measure or as an ensemble cluster. The next section makes use of various computational classification methods to capture the determinant acoustic and articulatory features among modal voice, vocal fry, and period doubling.

Table 3.20: Distributions of PD, vocal fry, and modal voice in women and men in the acoustic and articulatory dataset.

	Period doubling	Vocal fry	Modal voice
Women	324	154	187
Men	91	25	137
Total	415	179	324

3.5.3. Results: Classification of modal voice, vocal fry, and period doubling

Given the separation of the three voice categories as manifested by t-SNE, I further probe the contribution and importance of each of the acoustic features using various machine learning approaches including logistic regression with Lasso regularization, random forest, and support vector machine (SVM). The structure is as follows: I report the performance of each of the three computational models first in the acoustic dataset (5538 observations x 33 features), and then in the combined acoustic and articulatory dataset (918 observations x 44 features). In the following analysis, gender coded as a binary variable is included as a covariate to account for the variances in voice clusters as shown by t-SNE. All classification models are first devised using the training set, and then evaluated in the test set.

First and foremost, I focus on the correlational structure of acoustic and/or articulatory features measured from vocal fry, period doubling, and modal voice, as shown in Figures 3.21 and 3.22. Several groups of measures are highly correlated, they are, spectral prominence within or outside the same group, such as corrected formant amplitudes and spectral tilt measures (A1c-A3c, H1A1c–H2KH5Kc, etc.); HNR measures over different frequency window sizes

($HNR05 - HNR35$); formants frequencies (sF1, sF2, etc.); SoE and epoch (used to locate the significant excitation to calculate SoE). Due to the considerable number of features to be assessed and the correlation and collinearity among these features, Lasso regularization is used to shrink coefficients of less informative predictors and reduce potential overfitting issues.

In the logistic regression model, the dependent variable is the type of voice category, and the independent variables are all the acoustic measures described in Table 3.1. The general formula of logistic regression for both datasets is given in (3.2), and the R code used for regularization is in (3.3)-(3.5). The parameter value, alpha $=$ 1 in (3.3)-(3.4), stands for Lasso regularization. Through cross validation, I chose the lambda (λ) which generated the smallest misclassification error for the multinomial distribution [see (3.3)]. (3.4) uses this lambda and evaluates the predictors with potential shrinking. Finally, the model structure generated in (3.4) is used in (3.5) to predict the performance of the training classification on the test set.

- $Voice\_type \sim acoustic/articulatory\ measures$ (3.2)
- $cv.glmnet(train\_measures,\ train\_type,\ family =$ 'multinomial', $alpha = 1,$
  $type.measure =$ 'class') (3.3)
- $lasso\_optimal = glmnet(train\_measures,\ train\_type,\ family =$ 'multinomial',
  $alpha = 1,\ lambda = lambda.min,\ type.measure =$ 'class') (3.4)
- $predict(lasso\_optimal,\ s =$ 'lambda.min', $newx = test\_measures,\ type =$ 'class') (3.5)

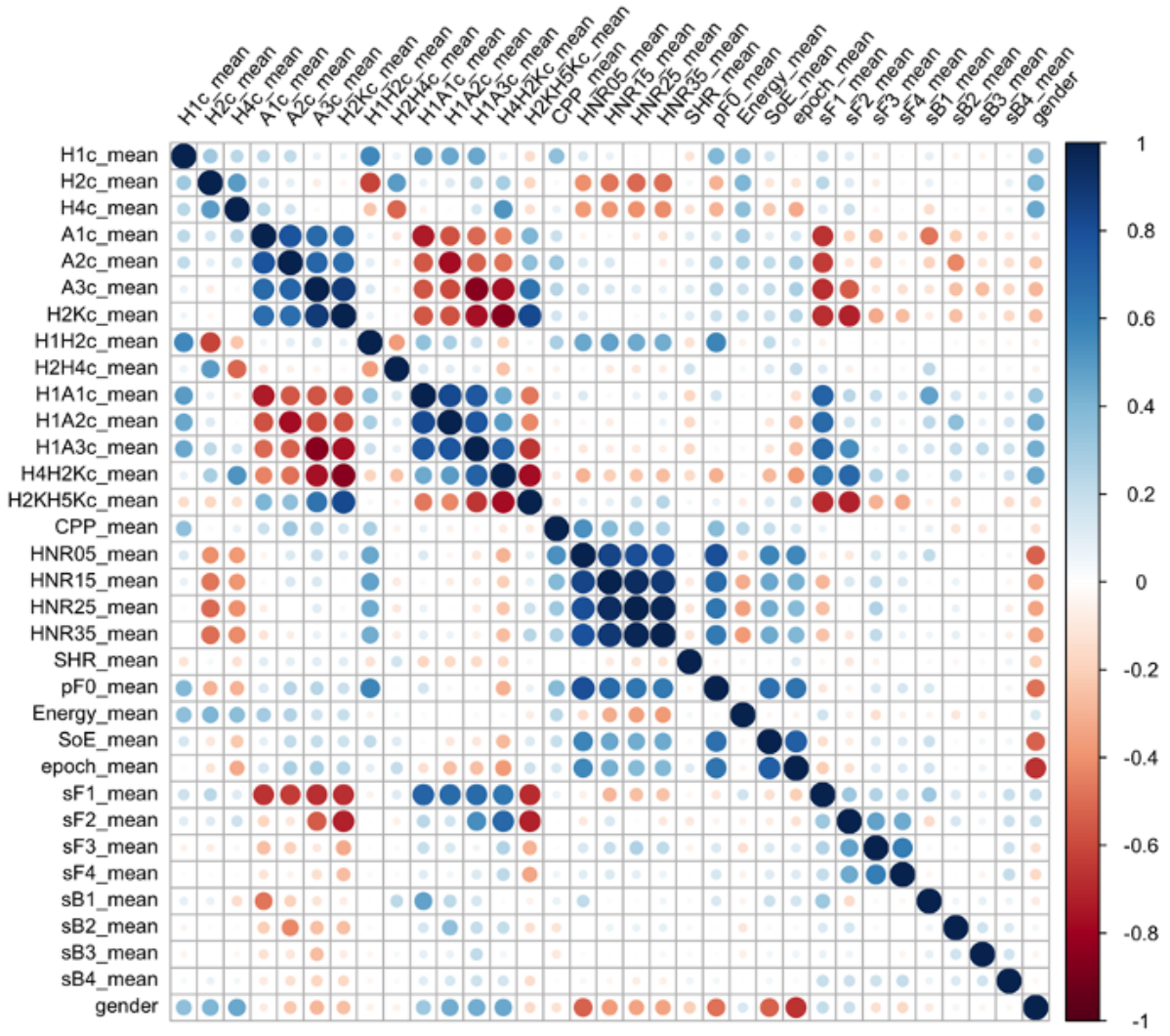


Figure 3.21: Pearson correlation coefficients among acoustic features. ‘c stands for corrected for formant values and bandwidths.

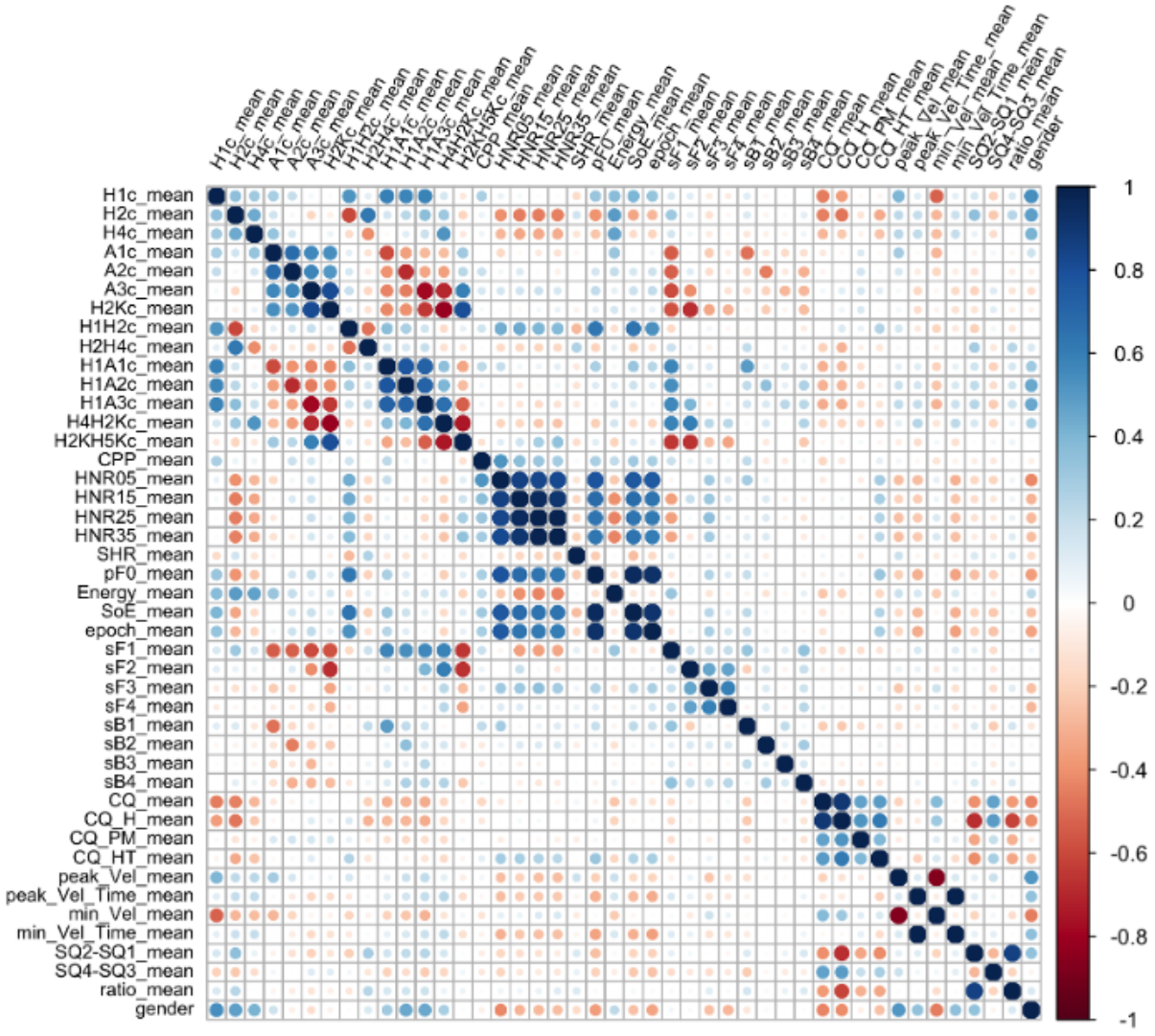


Figure 3.22: Pearson correlation coefficients among acoustic and articulatory features. 'c' stands for corrected for formant values and bandwidths.

Figure 3.23 plots the relationship between the value of $\log(\lambda)$ and the misclassification error. The minimum misclassification error (0.0915) is achieved when $\log(\lambda)$ approaches -8, shown by the first dashed vertical line. This lambda explains 81.81% of the multinomial deviance, which is a form of likelihood ratio to assess how well the model fits. The top line of numbers indicates the number of nonzero predictors, and the resulting number of nonzero predictors is 17. This demonstrates the determination process of choosing the optimal lambda according to the misclassification error of the model.

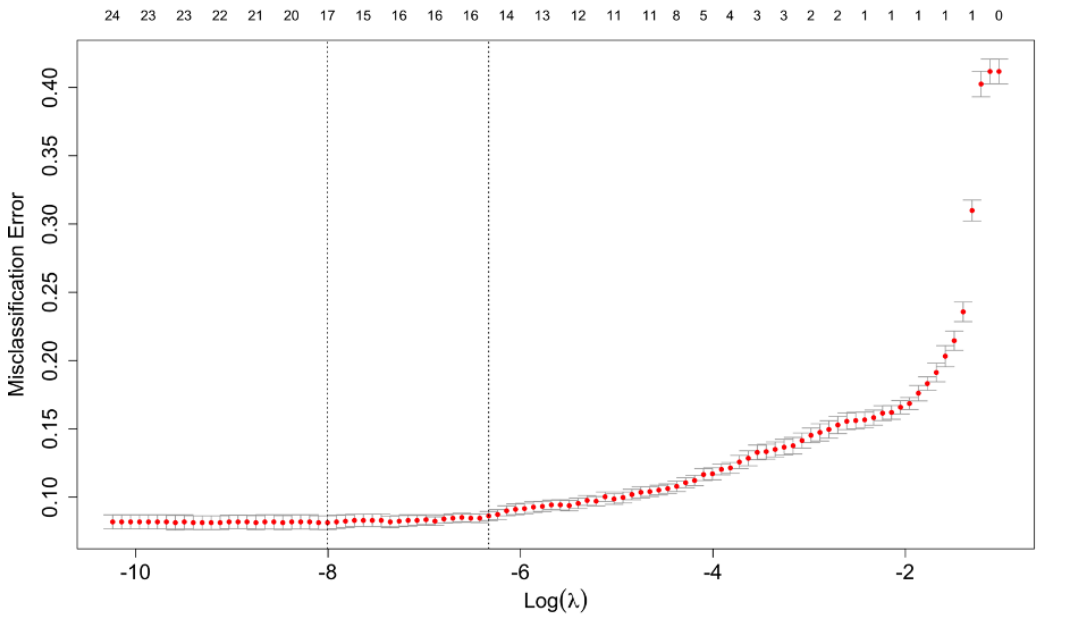


Figure 3.23: Lasso selection process of lambda based on misclassification error of period doubling, vocal fry, and modal voice using the acoustic features. The first dashed line indicates the lambda that gives minimum mean cross-validated misclassification error; the second dashed line indicates the lambda that gives the most regularized model (most zero predictors) with a cross-validated misclassification error within one standard error of the minimum.

Table 3.21: Confusion matrix of the predicted and the actual voice types in the test set of acoustic features using logistic regression with Lasso regularization.

Actual / Predicted	Period doubling	Vocal fry	Modal
Period doubling	2077	101	33
Vocal fry	177	236	5
Modal	9	13	1041

Using the selected model structure with the optimal lambda, the overall prediction accuracy is 0.9085 in the test set, and the confusion matrix is shown in Table 3.21. Given the class imbalance caused by fewer tokens of vocal fry, I also calculated the macro average precision (which accounts for Type I error, leading to false positives) and recall (which accounts for Type II error, leading to false negatives) of the classifier to evaluate its performance. The formulas of these evaluation metrics are shown in (3.6). For example, the precision of the predicted tokens of period doubling

is defined as the proportion of the true positives (2077) over the total number adding the true positives (2077) and false positives ($185 = 176 + 9$); the recall of the predicted tokens of period doubling is defined as the proportion of the true positives (2077) over the total number adding the true positives (2077) and false negatives ($134 = 101 + 33$). As a result of the calculations, the macro average precision across the three voice types is 0.8524 and macro average recall is 0.8278.

- $Accuracy = True\ Positive / All\ Tokens$
- $Macro\ Average\ Precision = Average(Precision(pd) + Precision(fry) + Precision(modal))$
- $Macro\ Average\ Recall = Average(Recall(pd) + Recall(fry) + Recall(modal))$
- $Precision = True\ Positive / (True\ Positive + False\ Positive)$
- $Recall = True\ Positive / (True\ Positive + False\ Negative)$ (3.6)

A detailed output of the coefficients of the acoustic predictors is shown in Table 3.22. The goal of the logistic regression is to predict a particular voice type given all the available acoustic features. With the feature elimination given by Lasso regularization, the remaining non-zero coefficients signal the "more important" predictors that contribute to a certain voice category. For each voice type, the coefficients of the predictors shrunk to zero may be different, reflecting the varying levels of importance of the various acoustic features in characterizing different categories. For instance, H4 and A1 are distinguishing features associated with period doubling (non-zero), but not with vocal fry and modal voice (shrunk to zero). H4–H2K is positively correlated with period doubling ($\beta = 0.22$), but negatively so with vocal fry ($\beta = -1.04$).

Table 3.22: Non-zero coefficients of different acoustic features predicting period doubling, vocal fry, and modal voice using Lasso regularized logistic regression. '.' means that a particular feature was shrunk to zero for a given voice type. Other features which descended to zero in all voicing types are not listed.

	Period doubling	Vocal fry	Modal
H2c_mean	.	-1.002	.
H4c_mean	0.10	.	.
A1c_mean	0.04	.	.
A2c_mean	.	0.13	.
A3c_mean	0.11	.	.
H2Kc_mean	.	0.002	.
H1H2c_mean	-0.29	.	.
H2H4c_mean	.	.	0.003
H1A1c_mean	-0.66	.	0.24
H1A2c_mean	.	.	0.13
H1A3c_mean	.	.	0.18
H4H2Kc_mean	0.23	-1.03	.
H2KH5Kc_mean	0.73	-0.47	.
CPP_mean	-0.44	.	.
HNR05_mean	.	-2.59	1.20
HNR15_mean	-1.59	1.56	.
HNR25_mean	0.71	.	.
HNR35_mean	.	-0.96	.
SHR_mean	-0.42	.	0.29
pF0_mean	-1.99	.	8.62
Energy_mean	.	-0.66	0.28
SoE_mean	0.56	-1.35	.
epoch_mean	0.77	.	.
sF1_mean	.	0.38	-0.80
sF2_mean	0.82	.	-0.11
sF3_mean	.	-0.05	0.24
sF4_mean	0.03	-0.03	.
sB1_mean	0.62	-0.002	.
sB2_mean	-0.13	.	0.23
sB3_mean	-0.13	.	0.20
sB4_mean	.	-0.20	0.34
gender	-0.05	.	5.67

This procedure as outlined in (3.2)-(3.5) is repeated for the combined acoustic and articulatory dataset. Figure 3.24 and Table 3.23-3.24 show the results in detail.

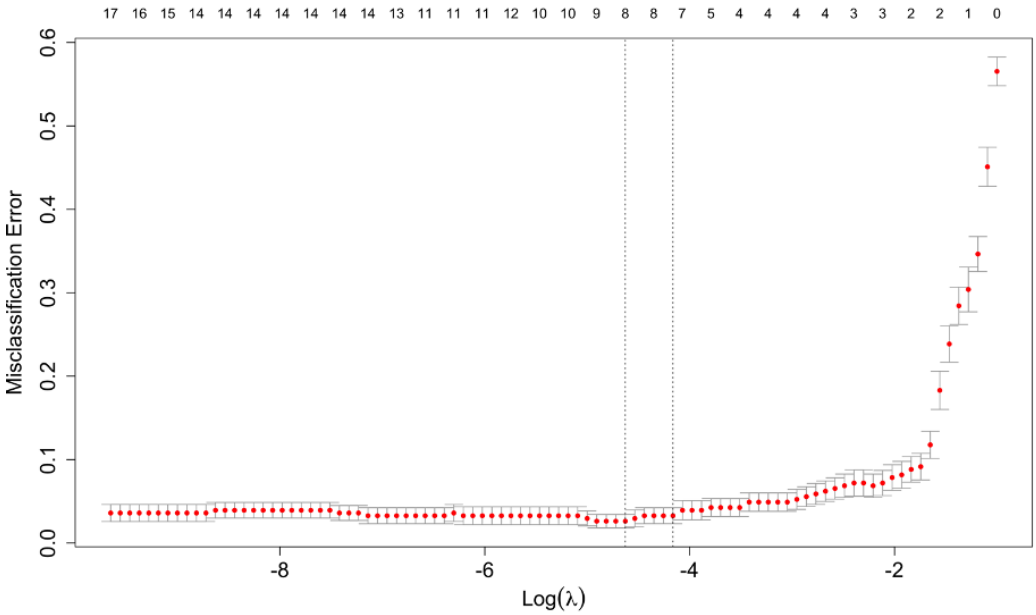


Figure 3.24: Lasso selection process of lambda based on misclassification error of period doubling, vocal fry, and modal voice using the acoustic and articulatory features. The first dashed line indicates the lambda that gives minimum mean cross-validated misclassification error; the second dashed line indicates the lambda that gives the most regularized model (most zero predictors) with a cross-validated misclassification error within one standard error of the minimum.

The minimum misclassification error (0.0212) is achieved when log(λ) is around -4.63, shown by the first dashed vertical line in Figure 3.24. This lambda explains 91.42% of the multinomial deviance. The top line of numbers indicates the number of nonzero predictors, and the resulting number of nonzero predictors is 8.

Using the selected model structure with the optimal lambda, the overall prediction accuracy is 0.9788 in the test set of the combined features, and the confusion matrix is shown in Table 3.23. Similar to the acoustic dataset, this dataset of combined features is slightly skewed by the smaller number of tokens of vocal fry. Using the same formulas in (3.6), the macro average precision across the three voice types is 0.9819 and macro average recall is 0.9698, which is much better than the performance of the logistic regression model using only acoustic features.

Table 3.23: Confusion matrix of the predicted and the actual voice types in the test set of acoustic and articulatory features using logistic regression with Lasso regularization.

Predicted \ Actual	Period doubling	Vocal fry	Modal
Period doubling	281	1	0
Vocal fry	5	109	3
Modal	4	0	209

A detailed output of the coefficients of the different predictors is shown in Table 3.24. Compared to Table 3.22, a larger number of predictors are shrunk to zero when articulatory measures are added to the model. This reflects that the addition of the phonatory dimension helps trim the model to be simpler and become more interpretable, and we can concentrate on a smaller number of defining features for each of the voice types. For example, articulatory measures do not constitute modal voice but show up in subtypes of creaky voice (e.g., vocal fry and period doubling but not modal voice make use of CQ and SQ measures), but a high coefficient of f0 certainly defines modal voice.

Table 3.24: Non-zero oefficients of different acoustic and articulatory features predicting period doubling, vocal fry, and modal voice using Lasso regularized logistic regression. '.' means that a particular feature was shrunk to zero for a given voice type. Other features which descended to zero in all voicing types are not listed.

	Period doubling	Vocal fry	Modal
H1c_mean	.	.	0.38
H4c_mean	0.11	-0.25	.
H1H2c_mean	-0.50	.	0.93
H42Kc_mean	0.05	.	.
pF0_mean	.	.	2.29
sF4_mean	.	0.10	-0.07
sB1_mean	0.10	-0.54	.
sB2_mean	.	0.30	.
CQ_mean	.	0.74	.
CQ_H_mean	.	1.40	.
CQ_PM_mean	-0.14	.	.
CQ_HT_mean	-0.90	.	.
min_Vel_mean	.	0.63	.
SQ4-SQ3_mean	-0.47	0.12	.
ratio_mean	.	-1.23	.
gender	-1.81	.	.

Next, I turn to the random forest classifier to investigate the importance of the selected acoustic and/or articulatory features. The formula is shown in (3.7). I additionally tuned the optimal value of *mtry*: a parameter that determines the number of variables randomly sampled as candidates at each split, the tuning formula of which is shown in (3.8). *mtry* is tuned based on the Out-of-Bag error estimate: a validating measure of the random forest model using bootstrapping techniques; more specifically, the error estimate calculates the mean prediction error on each training sample using decision trees that do not contain the sample during bootstrapping. An *mtry* parameter at 10 gives an Out-of-Bag estimate of error rate at 6.93%.

- $randomForest(x = train\_x,\ y = train\_y,\ ntree = 500,\ proximity = T,\ mtry = 10)$ (3.7)
- $tuneRF(train\_x,\ train\_y,\ stepFactor = 0.5,\ plot = TRUE,\ ntreeTry = 150,$
  $trace = TRUE,\ improve = 0.05)$ (3.8)

After applying the model parameters determined by (3.7) on the test set, the confusion matrix is shown in Table 3.25. The overall accuracy is 0.9282, the macro average precision is given by 0.9036, and the macro average recall is given by 0.8536.

Table 3.25: Confusion matrix of the predicted and the actual voice types in the test set of acoustic features using random forest.

Predicted \ Actual	Period doubling	Vocal fry	Modal
Period doubling	2136	50	25
Vocal fry	151	263	4
Modal	25	10	1028

The following Figure 3.25 shows the importance of individual acoustic features evaluated by the random forest model in the training set. Two criteria are given: mean decrease in accuracy and mean decrease in Gini index. The former measures the impact of each feature by showing how much the permutation of each feature decreases accuracy of the model; the latter measures values of each feature and sort them based on total decrease in node impurity (an optimal condition of the nodes in the decision trees). The top three features are the same using either evaluation metric: f0, H1*–H2*, and H1*. The spectral tilt measures are followed by energy and noise measures such as SoE and HNR.

The random forest model also outputs a multidimensional scaling plot (MDS) using the training set shown in Figure 3.26. A relatively clear separation of the three voice types can be observed, with overlap in the middle-bottom of the graph.

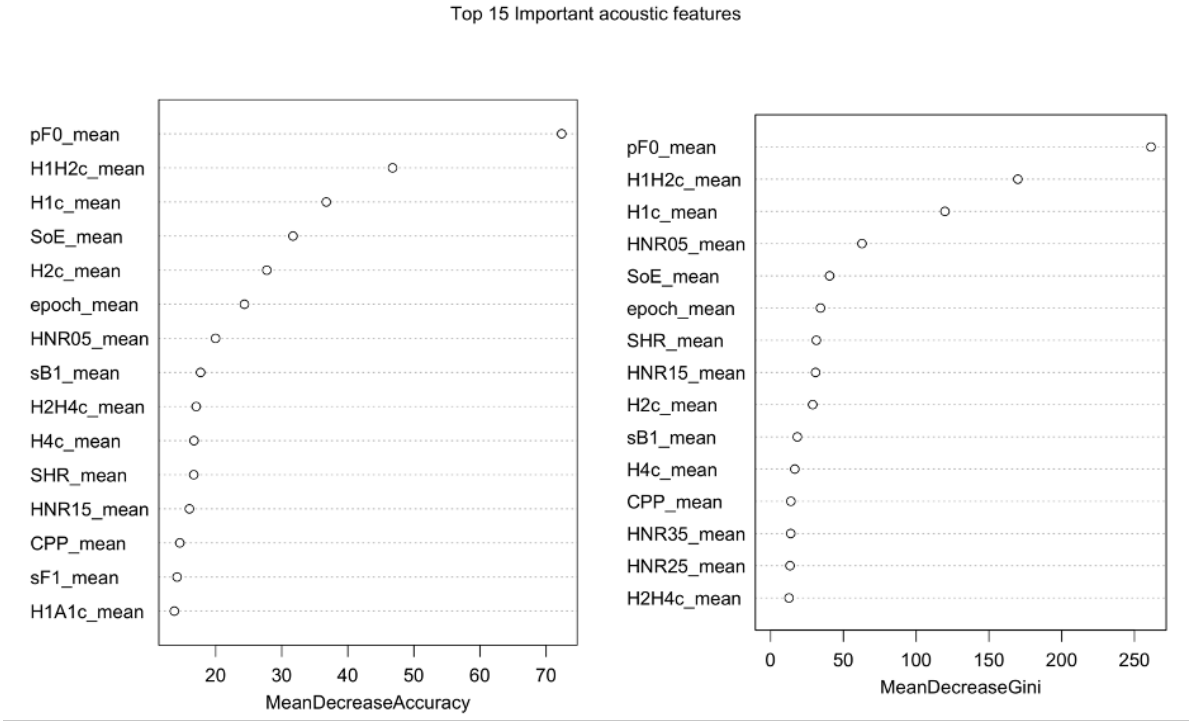


Figure 3.25: Top 15 important acoustic features in the training set of the random forest model.

As for the dataset using the combined acoustic and articulatory features, the procedure is similar, using the formulas in (3.7)-(3.8), except that the tuning parameter of mtry is set at 12 for an optimal Out-of-Bag estimate of error rate at 1.63%. Table 3.26 shows the confusion matrix, and Figures 3.27-3.28 display the ranks of variable importance and the MDS plot. The overall accuracy is 0.9771, the macro average precision is 0.9792, and the macro average recall is 0.9669, which outperforms the model using only acoustic features.

Table 3.26: Confusion matrix of the predicted and the actual voice types in the test set of acoustic and articulatory features using random forest.

Actual / Predicted	Period doubling	Vocal fry	Modal
Period doubling	281	1	0
Vocal fry	2	108	7
Modal	4	0	209

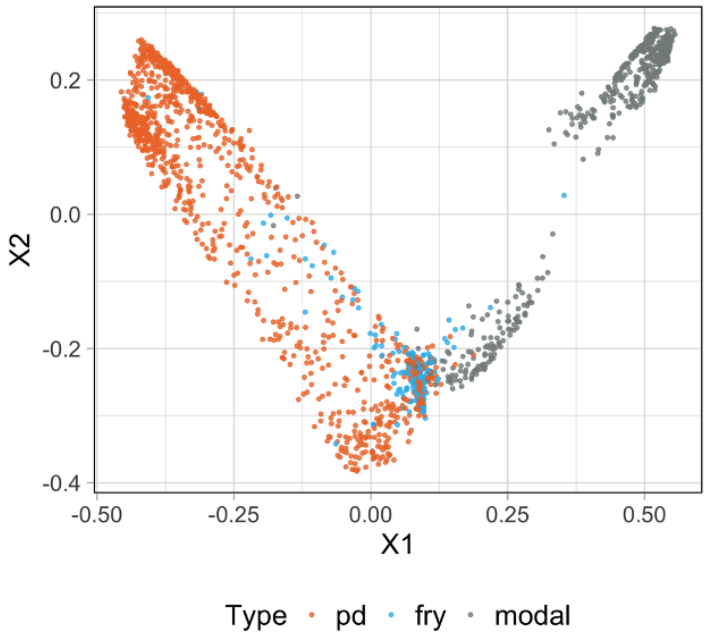


Figure 3.26: Multidimensional scaling plot using acoustic features in the training set of the random forest model.

According to the ranking of acoustic and articulatory features in Figure 3.27, the most important features in this combined dataset are H1*–H2*, the duration between the 10% and 90% of the opening phase (SQ4-SQ3), and SoE, followed by CQ measures and H1*. F0 is less important when the articulatory measures are added in the model.

Though with fewer data points, Figure 3.28 shows a better separation of the three voice types using both acoustic and articulatory features than only the acoustic features.

Lastly, given the nonlinear clustering of the three voice types and the considerable dimensions of the feature space in the data, I used SVM with a radial kernel, which is found to be effective in data with a higher number of dimensions than the number of observations, to compare the classification results of the random forest models. SVM finds a hyperplane that separates the data points of different voice types by projecting the features to higher dimensional space. The formula used is shown in (3.9). Two parameters need to be tuned in SVM: *gamma*, which determines the decision boundary, and *cost*, which penalizes misclassification error. Here I used the

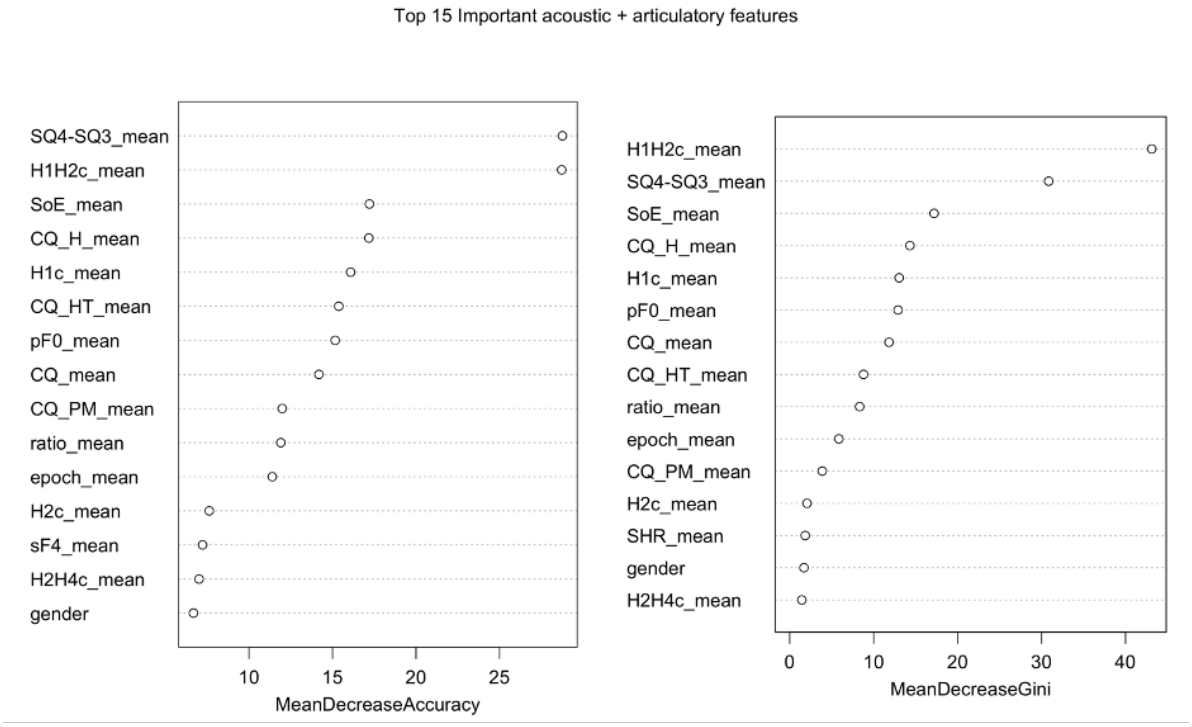


Figure 3.27: Top 15 important acoustic features in the training set of the random forest model.

default *gamma* in the R library `e1071` given by 1/33 (number of features in the acoustic dataset). Using *tunesvm*() shown in (3.10), the *cost* parameter is tuned at 10 for the acoustic model, and at 1 for the combined model.

- $svm(train\_y \sim train\_x,\ method = \text{‘C-classification’},\ kernel = \text{‘radial’}, scale = F,\ cost = 10)$

(3.9)

- $tune.svm(x = train\_x,\ y = train\_y,\ type = \text{‘C-classification’},\ kernel = \text{‘radial’},\ cost = 10^{-1:2})$

(3.10)

The following Table 3.27 displays two confusion matrixes predicted by the SVM in both datasets. For the acoustic model, the overall accuracy is 0.9163, the macro average precision is 0.8588, and the macro average recall is 0.8693, lower than the outcomes of the random forest

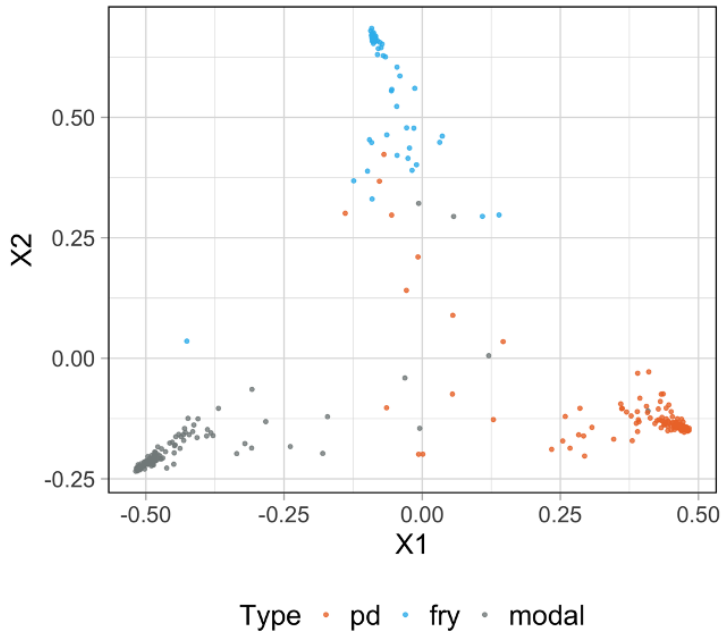


Figure 3.28: Multidimensional scaling plot using acoustic and articulatory features in the training set of the random forest model.

model. For the combined model, the overall accuracy is 0.9706, the macro average precision is 0.9572, and the recall is 0.9734.

Table 3.28 summarizes all the accuracy, macro precision and recall scores using Lasso regularized logistic regression, random forest, and radial SVM classifiers in the acoustic and combined acoustic and articulatory datasets.

Table 3.27: Confusion matrixes of the predicted and the actual voice types in the test set of acoustic features (a) and combined features (b) using radial SVM.

(a)

Predicted \ Actual	Period doubling	Vocal fry	Modal
Period doubling	2073	129	23
Vocal fry	98	280	10
Modal	40	9	1030

(b)

Predicted \ Actual	Period doubling	Vocal fry	Modal
Period doubling	280	7	4
Vocal fry	2	105	0
Modal	0	5	209

Table 3.28: Summary of overall accuracy, and macro average precision and recall scores using logistic regression, random forest, and radial SVM. Bolded numbers indicate the highest number in a particular score.

		Accuracy	Precision	Recall
Acoustics	Logistic regression with lasso	0.9085	0.8524	0.8278
	Random forest	**0.9282**	**0.9036**	0.8536
	Radial SVM	0.9163	0.8588	**0.8693**
Acoustics + articulatory	Logistic regression with lasso	**0.9788**	**0.9819**	0.9698
	Random forest	0.9771	0.9792	0.9669
	Radial SVM	0.9706	0.9572	**0.9734**

Overall, random forest models are shown to best predict and classify the three voice types in the acoustic dataset, and logistic regression model performs the best in the dataset with combined acoustic and articulatory features. Radial SVM models tend to perform better at recall scores by reducing the number of false negatives. Nevertheless, the differences in model performances are small – all models perform comparably well. The models using acoustic features alone already show reasonable separation whereas the models using a larger feature set with both acoustic and articulatory features effectively distinguish period doubling, vocal fry, and modal voice from each other. The most important acoustic measures are f0, H1*–H2*, and H1*, and the

most important articulatory features are the particular SQ measure (duration between 10% and 90% of the glottal opening slope) and CQ (hybrid), as established by combining the results of random forest models and logistic regression with Lasso.

Considering the mapping between perception and acoustics, articulatory measures are hardly accessible to listeners when encountering speech signals. Though adding the phonatory dimension better differentiates subtypes of creaky voice and modal voice in production, it remains unclear whether and how phonatory characteristics are transmitted to influence people's perception. It implies that listeners may show less robust categorization choices than a machine does with all the available acoustic and articulatory features. Yet, given that period doubling and vocal fry are distinguishable from each other and modal voice (Gerratt and Kreiman, 2001), it is possible that the perceptual product is not only acoustics, but a combination of both vocal folds vibration and vocal tract resonances.

3.6. Linguistic distributions

The second goal of this chapter is to investigate the distributions of period doubling and vocal fry as a function of segmental and prosodic factors. This section reviews the locations with respect to different linguistic environments: phrase, tone, and segment profiles where period doubling and vocal fry occur. Recall that period-doubled and vocal fry instances were located using the EGG waveform. Given that voice quality is suprasegmental, which spans across multiple segments rather than being restricted to a fixed one, in the corresponding audio recordings, period doubling and vocal fry were found in different syllables or part of the syllables including consonants or vowels. Both vocalic and non-vocalic segments were included for the analysis of tone and phrasal distributions, and only vowels were considered for the analysis of segmental distribution. Manual coding was done for the phrasal position, tone, and segment of the utterance where period doubling and vocal fry occur.

3.6.1. Prosodic analysis

The scripted corpus consists of 384 fixed carrier sentences with varying trisyllabic compounds: 'I teach you WORD how to say.' The compound words exhibit 64 different tonal combinations (4 tones x 4 tones x 4 tones; see Table 3.A1 in Appendix for a complete word list). To analyze the phrasal and tonal distributions of period doubling and vocal fry, I labeled the prosodic position, tone, and segmental profile of the identified tokens within the elicited utterances. Neutral tone is not considered in the analysis of tonal distribution. The explanations of the prosodic positions are in Table 3.29.

Table 3.29: Coded prosodic positions in the distribution analysis.

	Prosodic position	Explanation	Group
wo3	UI	Utterance initial	Utterance initial
tɕau1	PS2	Pre-stimulus 2 (further)	
nʲi3	PS1	Pre-stimulus 1 (closer)	
SYLL1	PI	Phrase initial (first syllable)	Utterance medial
SYLL2	PM	Phrase medial (second syllable)	
SYLL3	PF	Phrase final (third syllable)	
tsən3-mɤ0	AS	After-stimulus	Utterance final
ʂʷo1	UF	Utterance final	

The sentence structure with the coded positions is shown in example (3.1), and the waveform of an example sentence is shown in Figure 3.29.

(3.1) wo3 tɕau1 nʲi3 SYLL1 SYLL2 SYLL3 tsən3-mɤ0 ʂʷo1
UI PS2 PS1 PI PM PF AS UF
'I teach you WORD how-to say'

I then used a custom PRAAT script to log the information of prosodic position, tone, and segment associated with the tokens identified as period doubling or vocal fry. All the distribution data were processed and analyzed in R. There are 5848 tokens of period doubling and 1574 tokens of vocal fry used in the prosodic analysis.

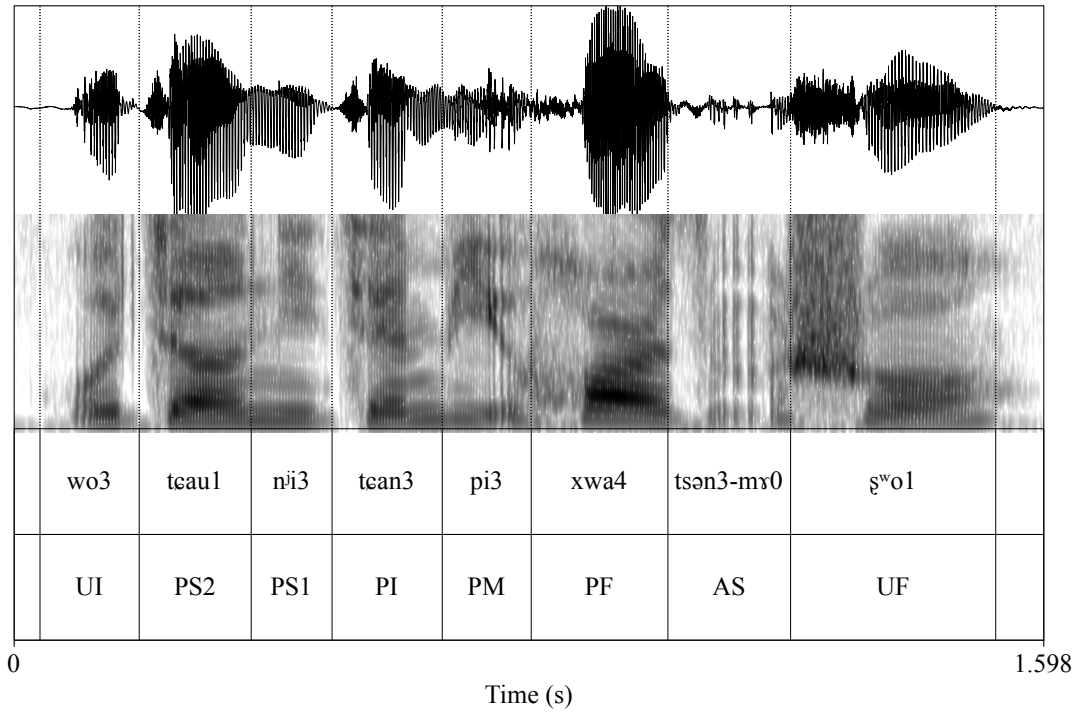


Figure 3.29: Waveform and spectrogram of a sample stimuli carrier sentence: 'I teach you black-board drawing how to say' with coded prosodic positions.

3.6.2. Results: Tonal distribution

First, I report the tonal distributions of period doubling and vocal fry. Because the tones in the fixed carrier phrase outside the sentence-medial compound do not vary, the analysis of the tonal distribution only applies to the compound stimuli. The phrasal positions associated with each of the words in the entire carrier phrase are separately analyzed in the following subsection.

Based on findings from previous studies that creaky voice is typically seen in Tones 3 and sometimes in Tones 2 and 4 because of a low pitch target (Belotel-Grenié and Grenié, 1994, 2004), both subtypes of creaky voice are expected occur in tones with lower pitch: Tones 2, 3, 4 more than Tone 1. Figure 3.30 shows the distribution of the two creak subtypes in different tones and sentence-medial positions. I coded the three syllables of the sentence-medial compound words

(SYLL1-SYLL3) as phrase-initial (PI; SYLL1), phrase-medial (PM; SYLL2), and phrase-final (PF; SYLL3).

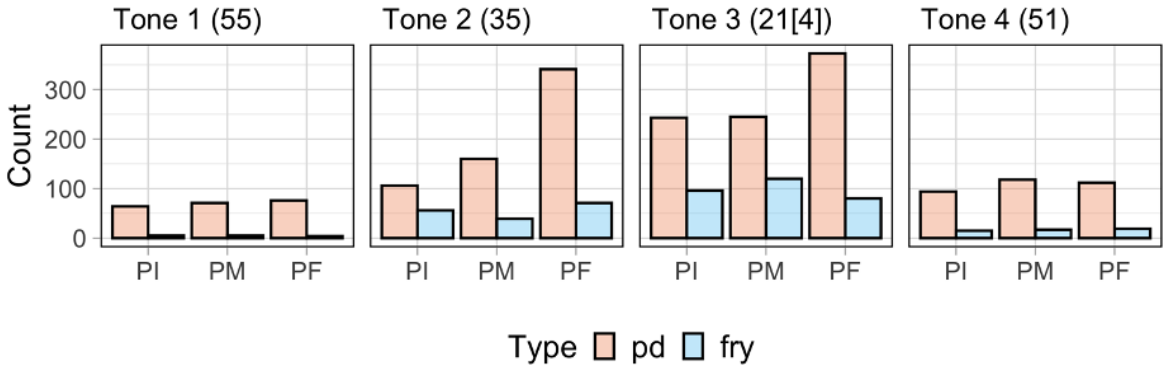


Figure 3.30: Bar plot of raw count of period doubling and vocal fry across tones in sentence-medial positions. PI = phrase initial, PM = phrase medial, PF = phrase final.

Overall, Tones 2 and 3 have more tokens of both subtypes of creaky voice than Tones 1 and 4. Vocal fry is rarely observed for Tone 1. For Tones 2 and 3, period-doubled tokens have a trend of a gradual increase as a function of the phrasal positions from the left-edge (PI) to the right-edge (PF) in trisyllabic sequences. This pattern is not attested in Tones 1 and 4 with an even distribution, probably due to fewer occurrences of creak for those tones. In contrast, the distribution of vocal fry does not seem to be conditioned by particular sentence-medial phrasal positions given that few differences are seen across these positions.

3.6.3. Results: Phrasal distribution

As detailed in Section 3.6.1, eight phrasal positions were coded. I first plot the three groupings of these positions based on larger prosodic breaks: utterance initial (consisting of all pre-stimulus syllables), medial (stimulus), and final (all post-stimulus syllables), as shown in Figure 3.31. Considering the difference in the frequency counts of period doubling and vocal fry, and those in women and men, their occurrences are plotted in percentage according to the prosodic breaks

within subset of different genders. Here an increasing trend is observed for both vocal fry and period doubling, and for both women and men.

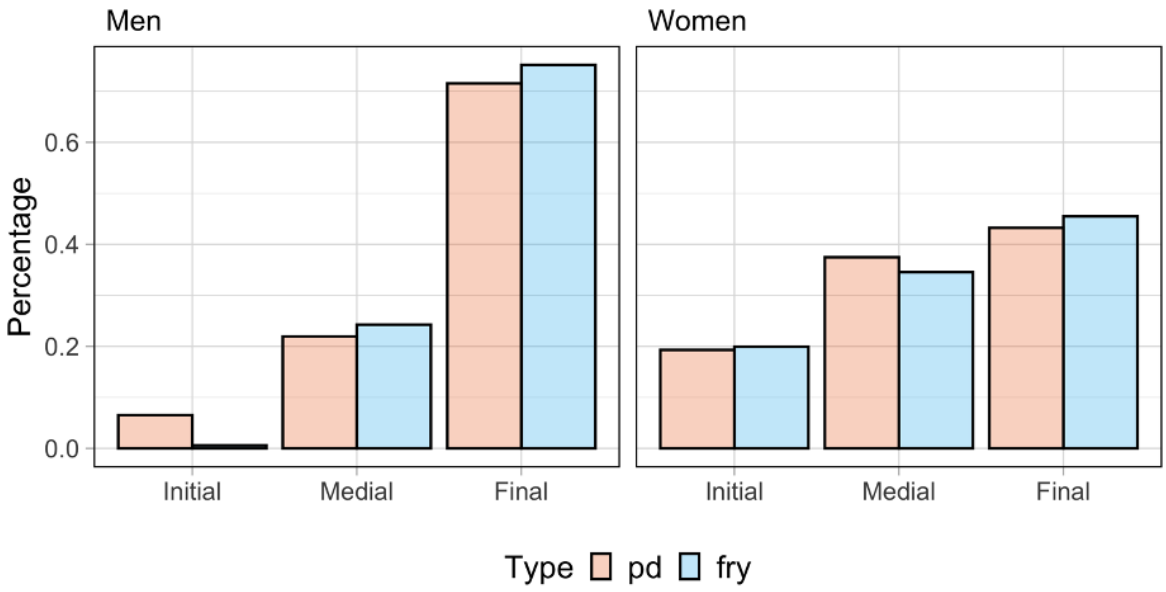


Figure 3.31: Bar plot (in percentage) of period doubling and vocal fry across different utterance positions: initial, medial, and final by gender.

Then, within each group, I analyze the occurrences of period doubling and vocal fry in more specific locations throughout the carrier phrase. The results shown in Figure 3.32 are presented with all the positions together for detecting finer-grained trends from the utterance initial to final position as the sentence unfolds in time. The percentage of both voice types are calculated based on the specific prosodic position within each gender subset.

The majority of the creaky tokens are associated with the AS and UF positions, which is consistent with the findings in Belotel-Grenié and Grenié (2004) that creaky voice in Mandarin frequently occurs in the final and penultimate positions. Interestingly, for both women and men, a linear increase is only observed in period-doubled voice starting from PS2 to UF. While period doubling is mostly found in UF, vocal fry mostly occurs in the post-stimulus, or the post-focal position. Post-focal compression is a well-documented phenomenon, characterized by compressed

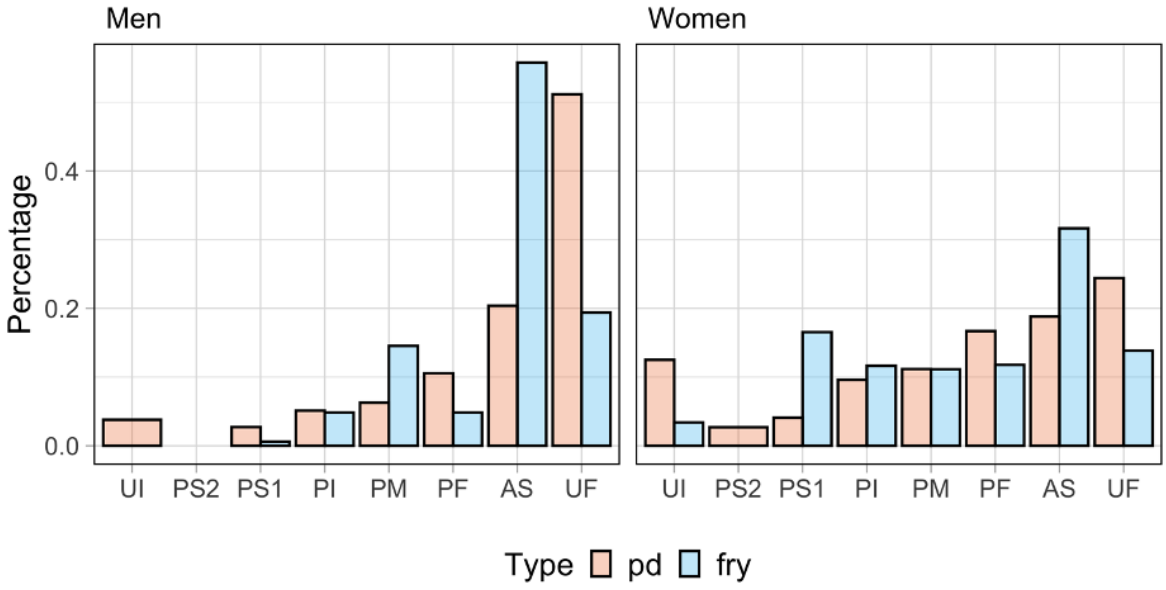


Figure 3.32: Bar plot (in percentage) of period doubling and vocal fry across different phrasal positions throughout the sentence by gender. UI = utterance initial (wo3); PS2 = pre-stimulus (tɕau1); PS1 = pre-stimulus (nʲi3); PI = phrase initial (Syll1); PM = phrase medial (Syll2); PF = phrase final (Syll3); AS = after stimulus (tsən3-mɤ0); UF = utterance final (ʂʷo1).

f0 range, shortened duration, and lowered intensity (Xu, 2011). Thus, it is likely that the compressed f0 range given rise by the post-focal position induces more vocal fry. The hypothesis is that period doubling is more strongly driven by utterance edges than is vocal fry, as seen in UI and UF, because the former reflects unstable voicing: towards the end of the utterance, voicing becomes progressively less stable. Consequently, the occurrences of period doubling in utterance-final position could be a byproduct of downdrift or automatic downstep as f0 progressively lowers towards the utterance edge without necessary constriction (see discussions about non-constricted quality during period doubling in Sections 2.5 and 2.7, Chapter 2). In contrast, vocal fry is typically triggered by a low and compressed f0 range with more probable constriction. The fact that the occurrences of vocal fry are more restricted and mostly occur in the penultimate position suggests that vocal fry could signal a stronger linguistic role such as a marking weak prosodic element post-focally.

Moreover, tonal influences are observed in combination with prosodic effect. For example, fewer occurrences of vocal fry than period doubling are found in UF position associated with a high-pitched Tone 1, and PS1 with Tone 3 shows more instances of vocal fry at least for women. These are consistent with the findings of tonal distribution in Figure 3.30. However, UI also with Tone 3 has substantially fewer occurrences of vocal fry than PS1 possibly due to the phrasal effect. Thus, prosodic position seems to be a stronger driving factor than tone in favoring the occurrences of the two creaky voice subtypes.

3.6.4. Results: Segment distribution

Creaky voice as a suprasegment, is often realized continuously through multiple phones and even syllables. Period doubling and vocal fry identified from the EGG signal are thus associated with both consonants and vowels in the audio waveforms. Given that a diversity of consonants was found in these tokens with creak, and to narrow down the range of investigation for potential patterns, here I only focus on vowels that carry the two subtypes of creaky voice. The goal is to probe the influence of vowel quality on occurrences of period doubling and vocal fry. On the one hand, the hypothesis is that vocal fry may be favored by low vowels which are more likely to induce constriction because of tongue retraction as the jaw lowers (Esling et al., 2019). For example, a recent study by Sands and Lubera (2017) found that vocal fry in American English occurred more frequently in low vowels and front vowels when eliciting words in isolation. On the other hand, given that period doubling may be an indicator of vocal instability and was found to be non-constricted in Chapter 2, I expect it not to be favored by low vowels and thus may happen more frequently in non-low vowels.

However, the design of the corpus did not control for the type or frequency of the vowels, and the range of vowels includes [i], [y], [u], [e], [o], [ɤ], [a], [ye], [ou], [au], [ei], [ai], [ie], [iu], [io], [ia], [uo], [ua], [ui], [ue], [iau], and [iou]. Moreover, period doubling or vocal fry may happen in a particular segment which could be a vowel by itself, or part of a diphthong or triphthong,

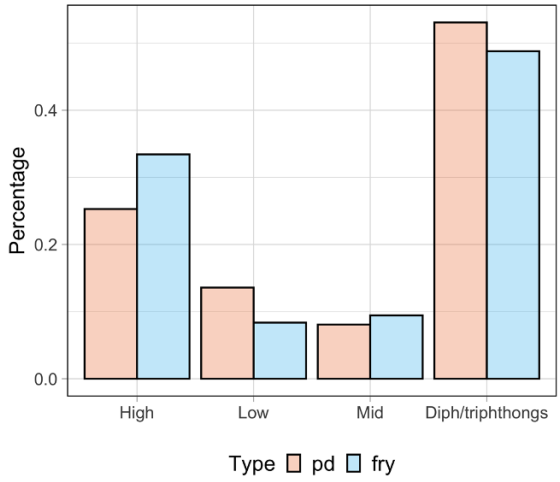
making it challenging to assess how voice quality changes as a function of vowel quality. As a preliminary analysis, I classify the monophthongs into low, mid, and high vowels and left the other diphthongs and triphthongs unclassified (Table 3.30) and plot their distributions in Figure 3.33. The percentages are calculated based on the distribution of different vowel categories within period doubling or vocal fry.

Table 3.30: Classification of vowels used in the distribution analysis.

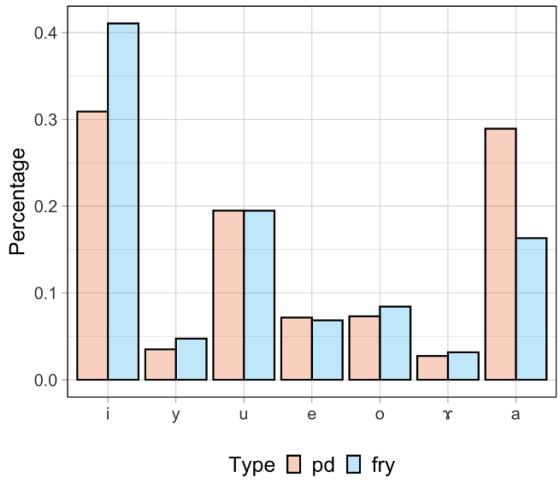
Vowel	Front	Back
High	[i, y]	[u]
Mid	[e, o]	[ɤ]
Low		[a]
Diphthong/ triphthong	[ye], [ou], [au], [ei], [ai], [ie], [iu], [io], [ia], [uo], [ua], [ui], [ue], [iau], [iou]	

It seems that vocal fry and period doubling more frequently occur in vowels [i, a, u] than other mid or high vowels [e, o, ɤ, y]. However, the fact that vowels [i, a, u] are more abundant than the other vowels may only suggest they are more prevalent in the stimuli and carrier phrases. The same applies to the four most frequently-occurring diphthongs: [au, ou, ai, ei].

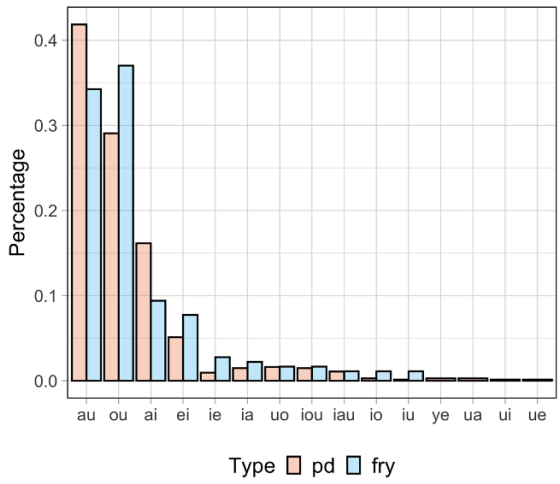
Vocal fry involves increased constriction and thus should occur more frequently among low vowels, on the assumption that these are more constricted (Esling et al., 2019). Besides low vowels, Sands and Lubera (2017) also found that vocal fry appeared frequently in front vowels, which are expected to have decreased constriction. In addition, creak is realized more easily in vowels that are noisier, represented by noise measures such as HNR and CPP (e.g., Awan et al., 2012 found that high vowels [i, u] had lower CPPs than low vowels [ɑ, æ]; recall that a lower value of CPP or HNR means more noise and less periodic). Nevertheless, in the present discussion, [i, a, u] have frequent instances of both creak subtypes, suggesting that vocal fry is not necessarily realized in low vowels and period doubling in non-low vowels. Because of the nature of the study design, it is more prudent to conclude that voice quality is not found to be affected by vowel quality, especially given the noisy dataset not particularly controlled for investigating the association between vowel quality and voice type.



(a)



(b)



(c)

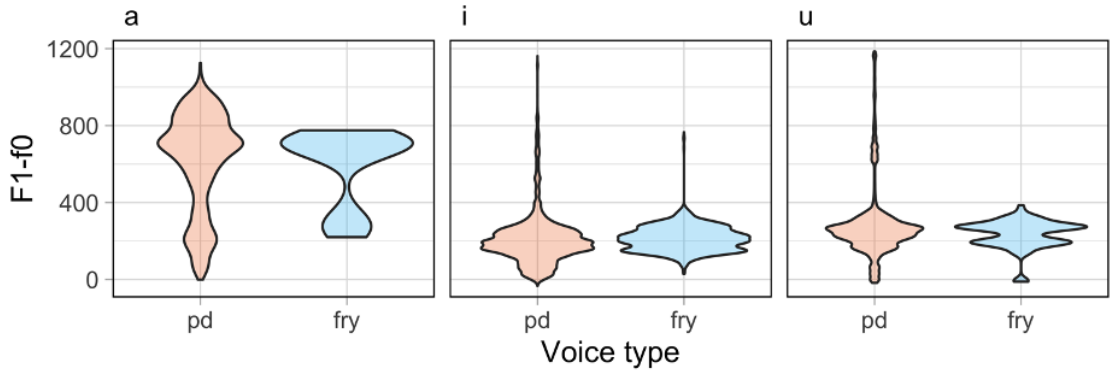
Figure 3.33: Bar plot (in percentage) of monophthongs by vowel height (a, b), diphthongs and triphthongs (c) across tokens of period doubling and vocal fry .

3.6.5. Results: The relationship between vowel formants (filter) and harmonics (source)

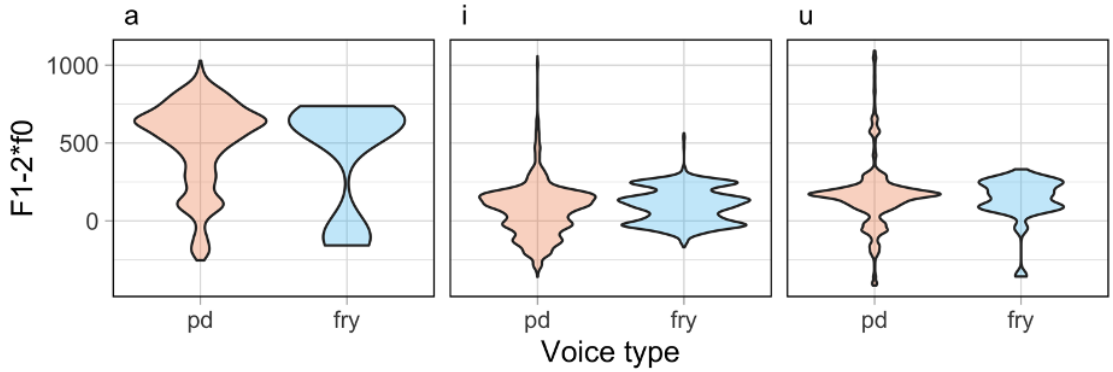
Recall that H1*–H2* was the most effective measure in distinguishing period doubling from vocal fry and modal voice; here, I probe how the harmonic amplitudes are affected by formants in different vowels. To map the influence of filter, or the vocal tract shape, onto the characteristics of period doubling identified from the EGG signal, I look into the formant-f0 interaction in period-doubled and vocal fry tokens. In particular, I use both the amplitude and frequency measures, H1*–A1* and H2*–A1* (in dB), and $F1 - f0$ and $F1 - 2 \times f0$ (in Hz) as difference measures between harmonic and formant amplitudes and frequencies, to investigate the relationship between formant and f0, especially given the stronger subharmonics and H2* in period doubling. To achieve this, I extracted vowels of [a, i, u] that exhibited creaky voice. There were 378 vocal fry tokens and 743 period doubling tokens.

Figure 3.34 shows the distribution of $F1 - f0$ and $F1 - 2 \times f0$ 'vicinity' by plotting the linear distance between F1 and f0 in three vowels [a, i, u] in tokens of period doubling and vocal fry. The high vowels [i, u] have a shorter distance than the low vowel [a], indicating a higher vicinity of f0 and F1. This is expected because high vowels have a lower F1. When subtracting $2 \times f0$ from F1, the values are smaller for all vowels and both voice types. This shows that the first vowel formant is closer to $2 \times f0$, which might amplify H2 during period doubling and vocal fry, consistent with the acoustic evidence of a lower H1*–H2* in both voices (when H2* is stronger, H1*–H2* should be lower, assuming a constant H1*). The two figures also do not demonstrate differences based on the voice types – their distributions are largely similar.

Next, I plot the distance between H1* and A1*, or H2* and A1* in Figure 3.35. There is not much difference in [i] tokens between the two measures, but [a] and [u] both have a higher H2*–A1* than H1*–A1*, meaning that the amplitude of the first formant is closer to H1*. In addition, period doubling has a larger value of H2*–A1* than vocal fry, whereas their average values of H1*–A1* are similar, meaning that its H2* is stronger in period doubling.

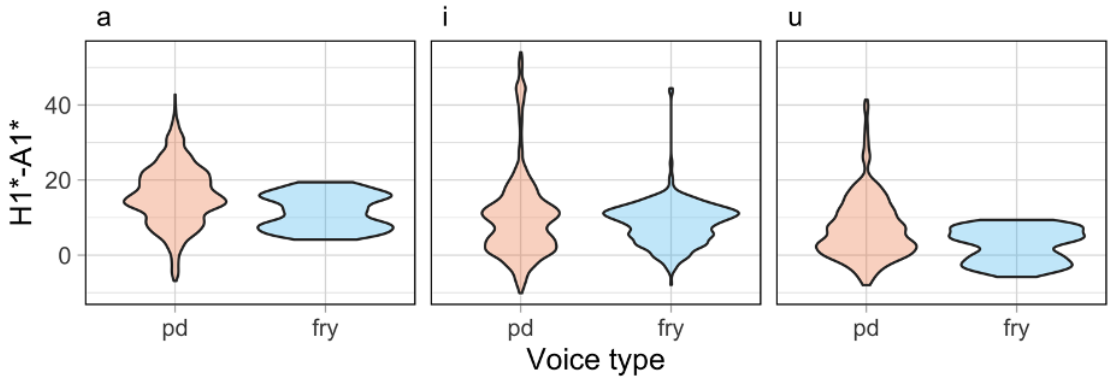


(a)

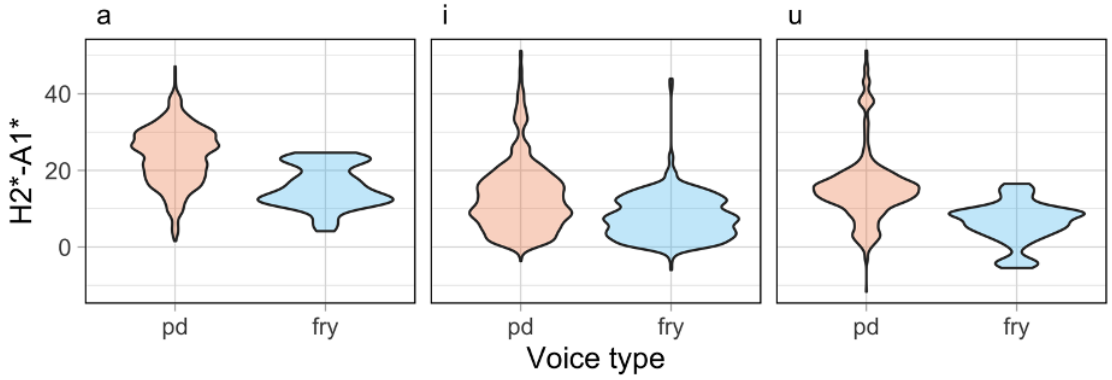


(b)

Figure 3.34: The distribution of the distance between $F1$ and $f0$ (Hz) (a) and between $F1$ and $2 \times f0$ (Hz) (b) in [a, i, u] across tokens of period doubling and vocal fry.



(a)



(b)

Figure 3.35: The distribution of H1*–A1* (a) and H2*–A1* (b) in [a, i, u] across tokens of period doubling and vocal fry.

3.7. Discussion

From the perspective of articulation, I have established that period doubling has alternating glottal pulses of different amplitudes, frequencies, and voice qualities, making it distinct from vocal fry and modal voice. In this chapter, I first probed the acoustic characteristics of period doubling as compared to the other two types of voice, by visualizing their distributions of individual acoustic features. I showed that period doubling and vocal fry have similar f0 ranges, both lower than that of modal voice. H1*–H2* is found to be most effective at signaling a three-way distinction among period doubling, vocal fry, and modal voice. Notably, the higher H2* distinguishes period doubling from the other two voices, and this contributes to the lower H1–H2. HNR and SoE show reasonably different values in period doubling compared to vocal fry and modal voice; so do spectral tilt measures including H1*–A1* and H2*–H4*, though the distinction for these measures is less salient. To simulate a holistic description of voice qualities, these findings based on the single acoustic features are further confirmed and generalized in a computational classification analysis by pooling together all available acoustic features obtained from the automated voice program. There I showed that f0, spectral tilt measures such as H1*–H2*, H1*, H2*, energy measures such as SoE, noise measures such as HNR, as well as SHR and the first formant bandwidth play an important role in delineating boundaries of period doubling, vocal fry, and modal voice in machine learning, whereas formants do not play a role throughout.

Moreover, to complement the findings in articulation of period doubling as detailed in Chapter 2, I applied the machine learning approaches in another dataset of files containing both acoustic and articulatory measures. With both acoustic and articulatory measures, several acoustic measures that are crucial in the acoustic dataset are still robust in classifying and predicting these voice types: H1*–H2*, H1*, H2*–H4*, and SoE. The articulatory measures SQ4-SQ3, the duration of the opening phase, and contact quotient are found to be distinguishing features as well. F0, epoch, and gender are also found to be important. Note that substantially more

observations come from women (<3 times than men), and many of the acoustic measures that exhibit a statistical significance are based on the subset of women, though trends are observed from the subset of men. Here, gender is self-reported by participants in a language background questionnaire; 'men' and 'women' are mostly referred to 'male' and 'female'. The difference in acoustic measures between genders as shown in the figures in this chapter is likely to be attributed to their biological and anatomical differences. From a physiological perspective, women might be using period doubling more than men because producing period doubling would be easier than vocal fry, which requires maintaining articulating a very low f0 with a constricted glottis. However, it is also possible that period doubling and vocal fry have different sociolinguistic functions as subtypes of creak because of their different linguistic distributions. This remains unanswered and awaits future studies.

By comparing the various machine learning methods including regularized logistic regression, random forest, and radial SVM, we can devise computational approaches to quantifying the acoustic features (and, sometimes, both acoustic and articulatory features) to predict the three voice types. Using mere acoustic features, these models performed reasonably well in classifying the three voice types; with the addition of articulatory features, model performances have been improved substantially. Random forest reached highest accuracy and macro precision scores in the acoustic dataset and provided insight into the most distinguishing and defining features, whereas Lasso regularized logistic regression performed best in the combined acoustic and articulatory dataset. In either dataset, radial SVM always performed better in macro recall scores by reducing the number of false negatives. Because vocal fry has the disproportionately fewer number of tokens, a common pitfall across all algorithms is the misclassification of vocal fry, compared to other voices, regardless of the dataset. More balanced datasets are thus desirable for a machine learning problem, but it may also be worth in a follow-up study to closely examine the files that are easily misclassified as another type to reduce potential conflation, or motivate theoretical questions such as how acoustically distinct we could establish two voice categories, how

to allocate importance of articulatory characteristics and acoustic attributes when distinguishing different voice types, etc.

One central goal of the current study is to bridge the gap between acoustic and articulatory parameters of certain voice qualities, as the two realms are often studied separately. Here I used both acoustic and articulatory features to study the similarities and differences among different voice types. In terms of classification and prediction of the labels of the voice categories, in future work, one could inspect the characteristics of voice qualities shown by the source, as modeled using the EGG signal, and the filter, reflected in the audio signal. For example, we could devise a classifier trained on EGG labels and tested on audio output, namely, using EGG waveform to predict acoustic parameters of a particular voice type. The relation between the source and filter could be evaluated by machine-learning approaches which take articulatory features of EGG pulses to form a selection criterion and in turn test on the corresponding audio chunks to see how well the EGG signal can predict the audio signal in terms of the occurrence of period doubling. A question to ask further is, what aspects of vocal fold articulation during period doubling lead to changes in the audio signal? Then we could study how changes in vowel quality interact with the overall phonatory pattern. A caveat here might be that the EGG signal, though a model of the source, will not be the same as the aerodynamic source being filter by the vocal tract to establish the more direct mapping and correspondence.

The fact that period-doubled voice shares acoustic characteristics of both vocal fry and modal voice obscures the acoustic differences between modal and non-modal phonation. The linear increasing trend of the amount of period doubling towards the end of the utterance is comparable to a similar phonation-ending gesture that Slifka (2006) has found for both irregular and regular phonation at the end of an utterance. The utterance-final irregular phonation is usually produced with short intervals of adduction followed by longer intervals of abduction and/or with incomplete closure of the vocal folds, rather than adducted like vocal fry (Slifka, 2006). Recall that during period doubling, the spread of contact quotient appears to be variable, with an overall

lower value than the modal voice. It is likely then, period doubling is associated with utterance finality without being necessarily irregular, at least in the sense of having an irregular single f0. Relatedly, Davidson (2019) has found that listeners performed poorly on identifying partial creak at the final 40-50% of the utterance, suggesting that creak is least salient when it occurs utterance-finally. The perception results would correspond to the present production findings in that possibly at the ends of utterances, voicing is generally less stable for listeners to easily differentiate non-modal from (expected) less regular but modal phonation. Or, utterance-final creak is harder to recognize when realized as period doubling with properties of both regular and irregular voicing, its varying degree of amplitude/frequency modulation, and a lesser degree of constriction.

Further, this chapter gives insight to data elicitation for speech production studies. Voice quality distinctions at the end of the utterance would be challenging to discern, because of the irregular voicing irrespective of the targets of voice quality. Thus, studies probing the influence of tone on voice quality would want to minimize phrasal effects, because for example, the realization of period doubling in Mandarin Tone 3 or other tones will result in similar kinds of articulation and voicing instability. The underlying voicing differences in these lexical tones are obscured. Thus, more follow-up studies are needed to investigate whether the non-modal phonation at the end of a phrase is due to phrasing, or to aerodynamic-articulatory constraints; that is, whether it is planned or involuntary. If period doubling does in fact increase due to voicing instability, a practical implication is that researchers who study voice quality associated with factors other than phrasal positions are recommended to record speech data in non-final contexts rather than isolation, to avoid edges of the utterance which induce unstable voicing such as period doubling.

3.8. Chapter summary

As both subtypes of creaky voice, how period doubling differs from vocal fry in terms of acoustic attributes and linguistic functions is not entirely clear. This chapter expands the articulatory findings and explores the defining characteristics of period doubling in the audio signal via an

examination of single acoustic features both individually and in multi-feature models. By comparing period doubling to vocal fry and modal voice as they occur in Mandarin, I found that as single features, H1*–H2*, H1*, H2*, and H2*–H4* most effectively distinguish period-doubled voice from other voices in the language. Using machine learning approaches, spectral tilt measures in the lower frequency range, and energy and noise measures typically contribute to the most variance in separating these three voice types. When evaluating both acoustic and articulatory features, spectral tilt measures remain one of the most significant contributors whereas articulatory constriction measures are a meaningful addition to improve the model performance in all machine learning methods. Thus, to adequately capture the distinctions among voice types, a combination of acoustic and articulatory dimensions is necessary.

This chapter also reports phrasal, tonal, and segmental distributions of period doubling and vocal fry in Mandarin. Both creak subtypes are more often found with Tones 2 and 3, which are associated with lower pitch, than with Tones 1 and 4. Vowel distributions seem to be largely random rather than patterned. Importantly, period doubling is found abundantly at the edges of utterances, with a gradual increase towards the end of the utterance, likely associated with sentence-level downdrift. In contrast, the incidence of vocal fry does not increase linearly towards the utterance-final position and instead concentrates more on the penultimate position, which is also post-focal. This suggests that vocal fry may be a product of f0 compression and perhaps signaling a stronger linguistic role. Period doubling is therefore considered a semi-regular type of voicing that commonly occurs phrase-finally, to compare with phrasal-final non-constricted creak as described in Slifka (2006). I suggest that it may be driven by vocal instability; as voicing becomes unstable at low subglottal pressure, period doubling is more likely to occur. This hypothesis, if proved correct, has ramifications for speech production studies, such that data elicitation is recommended in non-final contexts to avoid conflation of modal and non-modal phonation. These results complement the articulatory findings of period doubling; together, period doubling is established as a distinct subcategory within creaky voice, with predicted different linguistic functions.

Chapter 4

Perception of period doubling: pitch, voice quality, and tone

4.1. Introduction

The physical property termed fundamental frequency (f0) forms the basis of pitch perception. Typical modal voice possesses a single f0 that is simple and straightforward to identify as the primary correlate of pitch. According to pitch perception models, listeners can effortlessly uncover pitch from speech with a missing f0; the lowest harmonic (or f0) can be derived from the spacing of higher harmonics (multiples of f0), lending support to a spectral approach to pitch perception (Plack and Oxenham, 2005). On the other hand, by comparing each cycle to every other in a waveform, listeners can detect periodicities to determine the f0 for pitch; this is known as the temporal approach (de Cheveigné, 2005), and is typically used at low frequencies (Wightman and Green, 1974; McClaskey, 2016). However, a problem arises when multiple (sub)harmonics and periods are present in the spectral and temporal domains, respectively; when they occur, so does the possibility of multiple f0s, as both spectral and temporal pitches may be heard (de Cheveigné, 2005; also see Yost 2009 for a review of pitch perception models, in particular temporal ones using autocorrelation).

This study probes a type of voice, period doubling, that carries at least two simultaneous periodicities, leading to an indeterminate pitch with a low and rough quality (Keating et al., 2015; Yu, 2010; Schreibweiss-Merin and Terrio, 1986). In Keating et al. 2015, it is considered to be a special case of 'multiply-pulsed voice', a subtype of creaky voice, which instantiates

the presence of multiple f0s (in period doubling, two f0s) caused by the abundant subharmonics along with harmonics in the signal. In Gerratt and Kreiman 2001, it was described as one of the supraperiodic phonation types, meaning it is neither periodic as modal voice nor aperiodic as noise; and according to Klatt and Klatt (1990), period doubling is observed in $\sim 25\%$ of normal speakers' utterances. Still, production and perception studies are limited to a large extent. Past production studies only documented period doubling as having alternating long and short periods and high and low amplitudes, increased noise (Titze, 1994; Gerratt and Kreiman, 2001), and glottal constriction (Keating et al., 2015). In a recent study, Huang (2022) (itself an earlier version of Chapters 2 & 3) found that articulatorily period doubling is captured by two alternating glottal cycles with distinct pitches and voice qualities, and acoustically it is best distinguished by a low H1–H2 compared to vocal fry and modal voice. Perceptual findings from psychoacoustic studies suggest that pitch perceived during period doubling will lower as the stimulus f0 drops and as the degree of modulation within period doubling increases (Sun and Xu, 2002; Bergan and Titze, 2001). Based on the acoustic and articulatory characterization from Huang (2022) and the previous chapters, here I further probe the following questions regarding perception during period doubling. How do people perceive pitch during various types and extents of period doubling, and how does this voice quality affect tone perception? Using a standard artificial language learning and shadowing paradigm with implicit categories of 'high' and 'low' tones, I test listeners' ability to perceive and imitate tonal stimuli manipulated with period doubling. The goal is to investigate the role of period doubling in pitch and tone perception and production in a (pseudo-)linguistic context.

4.1.1. Perception of period doubling

The perception of period doubling involves both pitch and voice quality percepts. In studies of pathological voices, period doubling is described as the presence of two cooccurring tones that are similar to one another in quality and loudness, but which are perceived as two separate

pitches (Ward et al., 1969; Schreibweiss-Merin and Terrio, 1986). Specifically for synthesized period-doubled signals, the perceived pitch becomes lower as the amount of amplitude and frequency modulation increases, which can be predicted by measures in the frequency domain (e.g., subharmonic-to-harmonic ratio; Sun, 2002); however, perceived pitch differs across fundamental frequencies, and modulation types; for example, a lower f0 prompts earlier identification of a lower pitch, and the pitch drops more quickly in frequency- than in amplitude-modulated tokens (Sun and Xu, 2002; Bergan and Titze, 2001). Bergan and Titze (2001) additionally found that the performance of professional vocalists had higher intrasubject agreement on an ABX task and increased accuracy in a pitch matching task than people with none to moderate music training. They also found that sopranos and tenors tended to choose the fundamental more frequently, whereas altos and baritones tended to choose the (lower-frequency) subharmonic more often, suggesting a correspondence with their own natural speaking f0. However, the two most relevant studies examined only amplitude-modulated or only frequency-modulated period doubling; in natural speech, combined frequency and amplitude modulation is more commonly seen than the two types in isolation (Huang, 2022). Moreover, only English native speakers were recruited. The present study thus includes the combined modulation and asks how the three modulation types affect pitch perception. Besides music experience, linguistic background is considered, so that two groups of native Mandarin and English listeners are tested to investigate whether speaking a tone language influences pitch perception during period doubling.

In both synthetic and typical speech, two fundamental frequencies are found in period doubling, yet the voice is often heard as rough. This percept likely results from an indeterminate pitch (Yiu et al., 2002; Keating et al., 2015). Given that a defining characteristic of period doubling is its presence of subharmonics, researchers studying the relationship between subharmonics and roughness in connected speech (Omori et al., 1997; Kramer et al., 2013; Murphy, 2000). In particular, Omori et al. (1997) have found that the degree of roughness is related to the frequency and power of subharmonics, such that the stronger the amplitude and the lower the frequency, the rougher the voice is perceived to be. But Kramer et al. (2013) found that the degree of irregularity

and the percentage of low f0 values, but not the power of subharmonics, were salient cues to the degree of roughness.

Further, studies of the relationship between roughness and acoustic attributes have found that a low fundamental frequency, noise, and irregularities in the voice induce and reinforce a percept of roughness. Fraj et al. (2012) assessed the perception of simulated disordered voice using additive pulsatile noise, jitter, shimmer and frequency/amplitude tremor. They found that the degree of roughness is positively associated with pulsatile noise and frequency jitter while negatively related to the f0.

Period doubling appears to be perceptually distinct from other subtypes of creaky voice, at least for the clinical population. Gerratt and Kreiman (2001) found that, using samples from speakers with vocal pathology, period doubling can be easily distinguished from aperiodic voice. Previously, period-doubled voice had been found to be perceptually distinct from noisy and other rough or breathy disordered voices (Kreiman et al., 1993). It is also unclear how period doubling affects tone perception. For example, Mandarin tones are often realized with creaky voice, itself manifested as vocal fry and period doubling (Yu, 2010). However, in Mandarin tone identification, Huang (2020) found that resynthesized tones with period doubling (created by the 'double pulsing' parameter in the Klatt synthesizer) had a negative impact on tone identification under the interaction with noise, even with the frequently-creaky dipping Tone 3. It did not affect identification in the noise-free condition. Thus, period doubling could hinder rather than facilitate tone identification. It might be perceived as roughness or as competing pitches that disrupt tone perception.

How the acoustic configurations of period doubling might lead to the bitonal or indeterminate pitch percept remains a puzzle unique to this voice quality. For example, vocal fry is typically characterized by having a lower pitch, but not an indeterminate one. In the current perception experiment, I use a two-alternative forced choice task, and resynthesized stimuli to mirror the types of modulation (amplitude and/or frequency), and the various degrees of modulation, as

attested in the previous production study (Huang, 2020; Chapter 2). The questions to be explored here are, how perceived pitch (or pitches) varies according to modulation types and the extent of amplitude and/or frequency modulation, and how period doubling is used to identify linguistic tones.

4.1.2. Shadowing and imitation studies

To further assess the perception of voice and pitch during period doubling, another approach that complements a perception experiment is to elicit productions of period doubling. Period doubling is often found in natural speech and likely to be produced spontaneously (see discussions in previous Chapters 2 and 3). However, it is not entirely clear that speakers would follow instructions of producing such type of voice if prompted. Thus, I employ a shadowing task which immediately follows the perception experiment to investigate how speakers imitate period doubling and the acoustic characteristics of their productions. As reviewed in Lambert (1992), shadowing is a paced auditory tracking task to reproduce the stimulus presented, like imitation. It requires speakers to go through several phases: perceiving the stimulus as input, transducing the input to output, and producing the stimulus as output. It is expected that through imitation, we are able to infer how speakers perceive aspects of pitch and voice during period doubling, and the connection between perception and production.

Shadowing or imitation methods have been widely adopted in intonational studies that involve pitch production. For example, Tilsen et al. (2013) showed that English speakers imitated intonational gestures by controlling f0 targets of intonational tones within speaker-specific values to achieve partial imitation. D'Imperio et al. (2014) found that Bari Italian speakers were able to change both tonal alignment and pitch levels during imitation of pitch accents. In speech shadowing studies that are not designed to imitate pitch, imitators were still found to successfully mimic global and local f0 targets with certain amount of deviations (Bailly, 2003; Eriksson and Wretling, 1997); similarly, in a study investigating phonetic convergence during deliberate and

unconscious imitation, speakers followed the pitch of voices they were exposed to without being instructed (Garnier et al., 2013). It has also been used in speech quality evaluation (Pierce and Silbiger, 1972), promoting English intonation acquisition (Hsieh et al., 2013), and phonetic accommodation studies to study accuracy of phonetic and phonological imitation (e.g., Gessinger et al., 2021).

In the current shadowing experiment, because there is no ground truth of pitch given the often indeterminate percept during period doubling, shadowing is used to explore how period doubling affects people's perception and production of linguistic tones, rather than evaluating how well people can reproduce a known pitch and/or voice. In particular, considering that period doubling is characterized as having a bitonal and rough percept, it remains to be seen how speakers will imitate its pitch and voice quality depending on the modulation types and extents. More variability is expected if bitonal percept is dominant, and rough sounding quality is expected if roughness overrides the tonal percept of period doubling.

4.1.3. Research questions and hypotheses

Together, the present study asks the following research questions: how do people perceive and produce period-doubled voice with different modulation types and varying degrees of modulation? Does participants' production match their perception? that is, are they consistent in perceiving and imitating pitches during period doubling? Finally, do speakers with different language and/or music backgrounds behave differently in perceiving and producing period doubling?

On the one hand, based on the previous findings of a stronger frequency modulation effect on perceived pitch of period doubling, and assuming an additive effect of both modulation types, I expect that the perceived/imitated pitch will be the lowest with concurrent amplitude and frequency modulation. Additionally, I expect that the threshold of the modulation extent that drives listeners from perceiving a higher pitch to a lower pitch would occur earlier with con-

current amplitude and frequency modulation, followed by frequency modulation and amplitude modulation. I further expect to find a negative correlation between the modulation degree and perceived/imitated pitch, such that the *higher* the modulation degree, the more likely period-doubled tones will be perceived as *low tones* or shadowed as low pitch. When the modulation degree is higher, the amplitude and/or frequency differences between adjacent pulses are larger; the weaker pulse is thus harder to detect, and the result should be that listeners are less likely to treat the weaker pulse as a repeating cycle of the stronger pulse, so the perceived pitch is expected to be lower, as it is derived from a pair of adjacent pulses.

On the other hand, high tones may be chosen when tokens have a weaker degree of modulation. In this case, every glottal pulse can be similar to the following one, and a higher fundamental frequency can be retained based on the calculation of a single pulse. I also expect that, if a lower pitch or higher pitch is generally preferred, subjects will categorize or reproduce the corresponding pitch with more consistency and less variation. Alternatively, if a pitch is nearly impossible to define given the current stimulus, the vocal production could be very challenging and within-subject variability is expected.

It might also be expected that participants' language background will affect the pitch perception or production, with speakers of a non-tonal language (here, English) behaving differently from speakers of a tone language (i.e., Mandarin), in particular in showing more variability. Researchers have found that speakers of a tone language perform better in pitch perception (Bidelman et al., 2011; Creel et al., 2018; Alexander et al., 2011). Finally, music experience may affect pitch perception and production. Waters et al. (2021) found that musicians were better at vocal imitation, and Leonard et al. (1988) showed that singers responded to pitch changes significantly faster than the nonsingers. Therefore, I expect that musicians will be less affected by modulation and perceive and produce the pitch of period doubling more consistently and closer to the fundamental frequency rather than the subharmonic.

4.2. Experiment 1: Pitch perception during resynthesized period-doubled tones

4.2.1. Procedure

The experiment was implemented in PsychoPy (Peirce, 2007) in a sound-attenuated booth at the UCSD Phonetics Lab. Stimuli were played from Focusrite Scarlett 8i6 pre-amplifier connected to a lab computer and participants used a SHURE SRH440 or AKG K-55 stereo headphones throughout. A standard artificial language learning paradigm was used. Three phases of the experiment are shown in Figure 4.1. Participants were told they were learning a novel tone language which has two tonal categories, represented iconically as ↑ and ↓. The participants went through a familiarization phase where they heard 40 tokens of a single pitch within the reference ranges between half of the f0 and the f0 of the experimental stimuli. They had to attend to the presentation of the stimuli and follow the instructions by pressing an up or down arrow key upon hearing the sound to get familiarized with the corresponding category ↑ or ↓. Then they were tested in a training phase to categorize the aforementioned 40 tokens into either ↑ or ↓, with feedback provided. They were required to pass the training phase with a cumulative accuracy at the minimum of 75% after categorizing at least two rounds of the training tokens (80 tokens). Then they proceeded to the two repetition test blocks to categorize 380 synthesized tokens of period doubling into the same categories. The two categories were always given by symbols without explicit explanation. Order of stimulus presentation was fully randomized across listeners and phases. The entire experiment was strictly timed, and participants were warned if they responded too slow (after 1.3 seconds upon hearing the audio). Two misses on the same token were counted as an invalid response. Between the two test blocks, participants were prompted to take a break. The entire perception experiment lasted approximately 20 minutes.

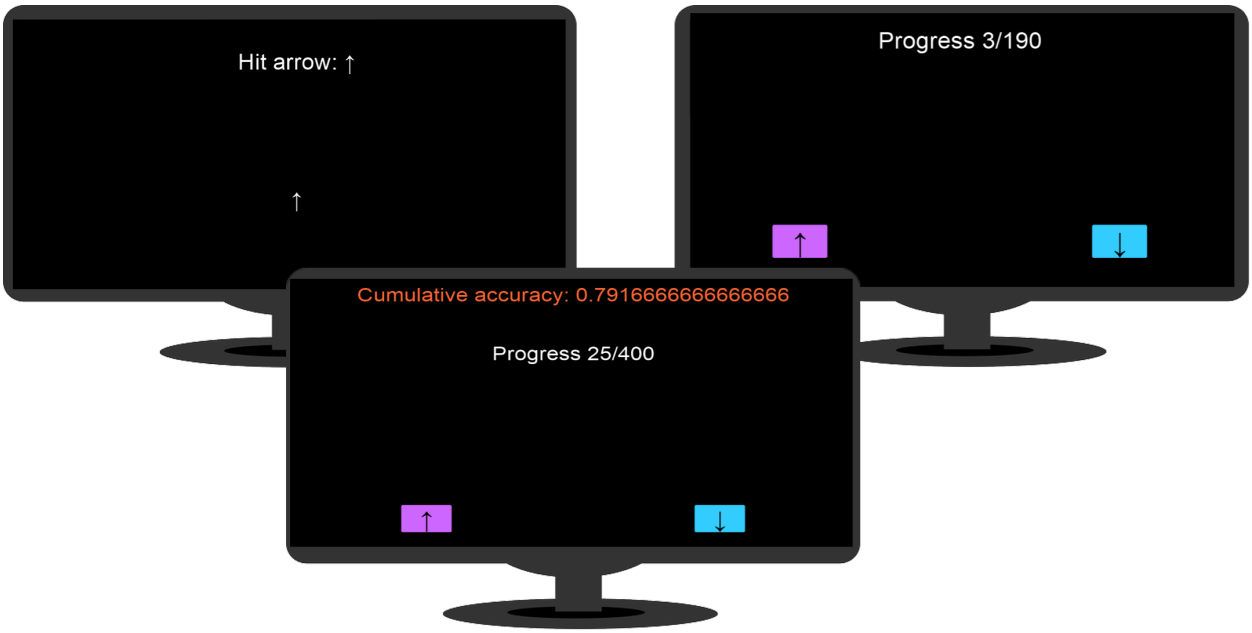


Figure 4.1: Experimental procedure from left to right: familiarization, training, testing phases.

4.2.2. Stimuli

Period-doubled stimuli The experimental stimuli consisted of resynthesized tokens of a vowel [a] produced and recorded by the author. One pulse of the vowel was extracted in the middle of the [a] from a region of stable formants. All stimuli contain a basis formed by duplicating and/or manipulating this extracted pulse according to the experimental condition; the two pulses are concatenated, and then the result of which is duplicated several times and concatenated again to generate a complete sound (illustrated in Figure 4.2). To reproduce the characteristics of period doubling, three experimental conditions were created based on the empirical ratios calculated from electroglottography in a scripted Mandarin corpus (Huang, 2022; Chapter 2): amplitude modulation, frequency modulation, and combined amplitude and frequency modulations of every other cycle. The resulting stimuli have alternating pulses of "long-short-long" periods and/or "high-low-high" amplitudes (as described in Titze 1994; Gerratt and Kreiman, 2001). The steps of creating the stimuli are detailed below, achieved using a custom PRAAT script (Boersma and Weenink, 2022) to automate the process.

1) In the non-modulated condition, duplicate the extracted pulse from [a] to produce two identical pulses. (Figure 4.2a)

2) In the amplitude modulation condition, the amplitude of the first pulse is retained, and that of the second pulse (a2) is reduced based on the amplitude ratio a1/a2. (Figure 4.2b)

3) In the frequency modulation condition, the duration of the first pulse (d1) lengthens and that of the second pulse (d2) shortens based on the frequency ratio d1/d2, while maintaining a fixed presumptive fundamental frequency given by $2/(d1 + d2)$. (Figure 4.2c)

4) In the combined amplitude and frequency modulation condition, the first pulse lengthens, and the second pulse both shrinks and lowers in amplitude. (Figure 4.2d)

5) Concatenate the two modified pulses to form the basis, and duplicate the basis, and repeat the contatenation and duplication process until the entire duration of the sound reaches 1 second.

6) Excise 300 ms of the stimuli from the midpoint of the vowels to mirror the length of a typical syllable.

7) Scale intensity to 70 dB.

As for the specific ratios used for amplitude and/or frequency modulation, the range of values were based on a previous study on production of period doubling in a scripted corpus (Huang, 2022; also in Chapter 2). Here I chose the range of ratios that covers at least two standard deviations from the median of either distribution. Table 4.1 shows the specifics adopted to create the stimuli. When a ratio is equal to 1, it means there is no modulation because the two amplitudes or frequencies are the same, which is expected to create a modal-sounding tone. On the other hand, extreme values (4 for amplitude ratio; 3 for frequency ratio) were included to anchor the other end of the distributions to make sure the tokens do indeed sound period-doubled. I would expect to see a sigmoid-like shape in the perceptual results along with the varying degrees

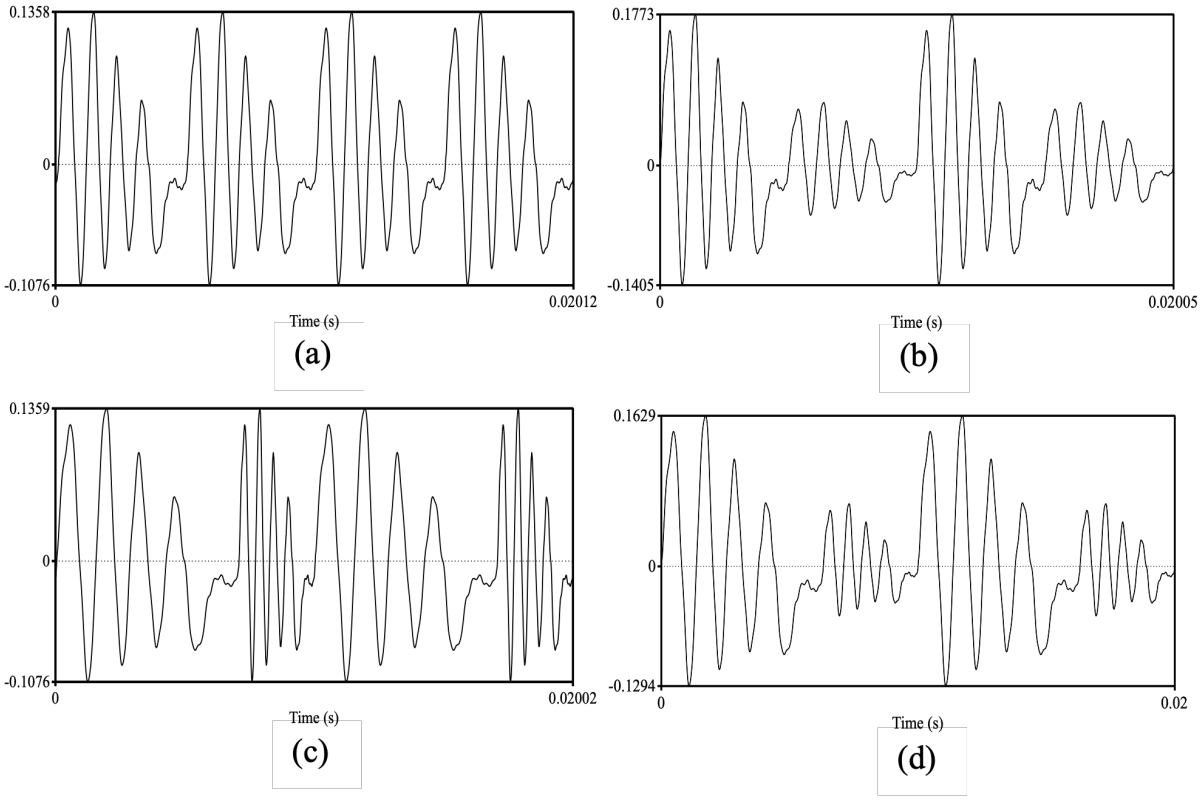


Figure 4.2: Resynthesized period-doubled pulses of 200Hz: original unmodified token (a), amplitude-modulated at 2.4 (b), frequency-modulated at 2.4 (c), and amplitude- (ratio: 2.4) plus frequency-modulated (2.4) (d).

of modulation when the voice of the tokens changes from modal (a single pitch) to period doubling (multiple pitches). Note that when the ratio increases, the modulation extent increases, and an increasingly stronger percept of the lower pitch (half of the original one) is expected, because the two alternating pulses become more and more distinct to facilitate the lower pitch percept cued by the lower f0, given by $1/(d1+d2)$.

Table 4.1: Amplitude and frequency ratios used for modulation based on empirical data.

	Range	Step	Median (sd)	Median + 2sd	Total
Amplitude ratio	Range: (1, 3); Extreme: 4	0.2	1.43 (0.73)	2.88	11
Frequency ratio	Range: (1, 2.6); Extreme: 3	0.1	1.38 (0.56)	2.50	17

For the amplitude modulation, a step of 0.2 was chosen based on the consideration to avoid over-generating tokens and the pilot findings that differences across varying degrees of amplitude modulation at a step size of 0.1 were less salient. A total of $(11 \times 17 + 3) \times 2 = 380$ tokens are thus used in the experiment. This includes the full range of pure amplitude-modulated, frequency-modulated, and combined amplitude- and frequency-modulated tokens.

According to Sun and Xu (2002) and Bergan and Titze (2001), tokens were perceived as having a lower pitch around modulation extent of $40\% \sim 50\%$, depending on the type of modulation. The range adopted here covers up to 44.4% (frequency modulation) - 50% (amplitude modulation) with the extreme values targeted 50% for frequency modulation and 60% for amplitude modulation extent.

Training stimuli The training stimuli used in the familiarization and training phases have the same segmental profile [a] as the period-doubled tokens, except that they were modal, without any amplitude or frequency modulation. The stimuli were resynthesized using the same extracted pulse with duplication and concatenation, while their pitches were manipulated and controlled for using the overlap-add method in PSOLA through the *Manipulation* function in PRAAT. The steps are shown below:

1) Duplicate the extracted pulse from [a] and concatenate these pulses and repeat the process to render a token of 1 second.

2) Extract the pitch tier of the token in PSOLA.

3) Change the pitch tier using a formula and replace the pitch with its original one to generate a training token.

4) Scale intensity to 70 dB.

Depending on the experimental condition, 20 pitch values each were randomly sampled in Gaussian distributions around different means and standard deviations, see Table 4.2. For ex-

ample, in one condition, 40 training tokens were generated from the distribution around $200Hz$ and $100Hz$ ($=$ 200/2), and in the other condition, 40 training tokens were generated from the distribution around $300Hz$ and $150Hz$ ($=$ 300/2). The standard deviations were chosen considering the non-linearity of pitch perception after converting hertz to semitones to better simulate a comparable distance between different pitches by the human ear.

Table 4.2: Gaussian distributions (mean, SD) used to generate pitch values of training tokens in f0 conditions of $200Hz$ and $300Hz$.

Condition	High pitch distribution	Low pitch distribution
$200Hz$	Gauss (200, 20)	Gauss (100, 8)
$300Hz$	Gauss (300, 20)	Gauss (150, 15)

4.2.3. Participants

Participants included two language groups: one of native Mandarin speakers ($n = 30,\ 18F,\ mean\ age = 20.43,\ range = 18.45 - 22.75$), and the other of native English speakers ($n = 31,\ 22F,\ mean\ age = 20.26,\ range = 18 - 24.5$). Participants had not learned a second language or dialect before age seven, and were recruited from the UCSD Psychology Subject Pool and received undergraduate course credit, or flyers on campus and received $15 compensation for their participation. No hearing or language disorders were reported. Three additional English speakers were tested and excluded due to ineligibility, either because of age or speaking another tone language.

All subjects heard the 40 training tokens of modal tones and 380 test tokens of resynthesized period-doubled tones during the experiment. However, a between-subjects design was used to test different f0 conditions such that within the group of Mandarin or English speakers, half of the participants heard the tokens based on an f0 of $200Hz$ and the other half heard those based on $300Hz$. The main reason for this was to avoid having long experimental sessions and thus to better hold participants' attention, and more importantly, to help them form structured perceptual categories during familiarization and training phases without unwanted interferences from a different set of frequencies.

In addition, participants filled out a questionnaire about their music experiences. If they had more than seven years of music experiences including either playing instruments or vocal performance, and still actively engage in music, they were categorized as "musician" for the purpose of the current study. This criterion was selected to ensure that this group of participants have sufficient music experience to be able to compare with the rest of the participants for the music effect. Similar criterion may be found in music perception studies (e.g., trained pianists in Manning et al., 2020). Fifteen participants were thus counted as musicians (Mandarin: 10; English: 5). Five participants self-reported as having absolutely no music experience (Mandarin: 3; English: 2). One Mandarin listener and one English listener also self-reported perfect pitch. 28 participants who had around five years of music experience but were not on an active status were classified as having "some" music experience (Mandarin: 15; English: 13), and 14 other participants who had less than five years of experience were categorized into having "a bit" music experience (Mandarin: 2; English: 12). Though, 'a bit' was collapsed into 'some' as one group in the statistical analysis.

4.2.4. Analysis

Statistical analyses were done in R (R Core Team, 2022) with mixed models from the lme4 package (Bates et al., 2015). The graphs were plotted using the ggplot2 package (Wickham et al., 2016). The reaction time (RT) in milliseconds was log-transformed and normalized by subject using z-scores. RTs having a z-score larger than 2.5 were excluded ($\sim$ 1.5% of the data).

Logistic and linear mixed-effects models were used to predict binomial categorization choice, indicated by whether a participant's chose ↑ ('up') or ↓ ('down'), and log transformed RT, respectively, given the independent variables *manipulation type* (no, amplitude, frequency, combined modulation), *f0 condition* (200Hz, 300Hz), their interactions, *language* (Mandarin, English), *music* (yes, no, some) and random intercepts of *subject* and *repetition* of the stimuli (first, second). Nested analysis of variance (ANOVA) model comparisons was used to compute the

overall significance of the effects of interest. Within the effective mixed-effects models, *lmerTest* (Kuznetsova et al., 2017) was used to assess the significance of individual levels within the fixed effects. The baselines used for within-factor comparison were modal (unmodulated), $200Hz$, and 'up' responses. In other words, R will automatically compute the coefficients of the fixed factors for the modulated conditions, $300Hz$, and 'down' responses, as compared to the baseline.

To investigate the consistency across participants' choices for a given stimulus, 'up' responses were coded as '0' and 'down' responses as '1' to be able to calculate mean and standard deviation of their choices. More consistent choices are expected if the mean ratings are around 0 or 1, and less consistent choices should have ratings approximating 0.5.

4.2.5. Results: categorization choice

Omnibus model The initial maximum structure is shown in equation (4.1).

$$choice \sim modulation\_type * f0 + language + music + (1|subject) + (1|rep) \tag{4.1}$$

However, this rendered convergence problems; using nested model comparisons with and without the factor of *language*, or *music* relative to equation (4.1), respectively, I found no main effect of language $[\chi^2(1) = 0.0006, p = .98]$, or music $[\chi^2(1) = 1.3, p = .52]$. Given that neither language nor music contributed to the variance of the response variable, in the omnibus model, the maximum parsimonious structure is given by equation (4.2). I will test potential interactions between language or music and the f0 condition in subset models separated by modulation types. Figure 4.3 plots the overall proportion of ↓ responses varied by modulation types. Figures 4.4 and 4.5 then plot the proportion of 'down' responses as a function of the varying degrees of amplitude and/or frequency modulation separately in subsets of modulation types. A linear trend can be observed in both modulation types; however, in the combined modulation, the proportion of 'down' responses followed the degrees of frequency modulation, which overrode the pattern

of amplitude modulation. Further, a different linear trend is shown by the higher f0 at $300Hz$ in amplitude-modulated tokens for both listener groups, which corresponds to the increased proportion of 'down' responses, compared to $200Hz$ tokens (Figure 4.4). This interaction effect is not found in frequency-modulated tokens on different stimulus f0s, but a trend of increased 'down' responses can be observed in Mandarin listeners at smaller frequency modulation degrees (<8; when the longer cycle is less than 1.7 times of the shorter one) for tokens of $300Hz$ (Figure 4.5).

$$choice \sim modulation\_type * f0 + (1|subject) + (1|rep) \tag{4.2}$$

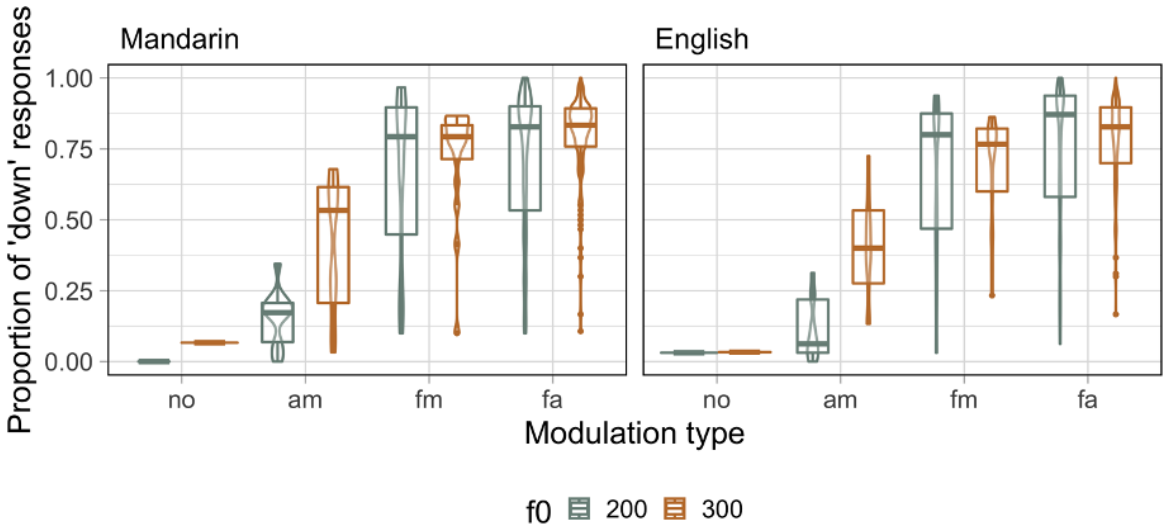
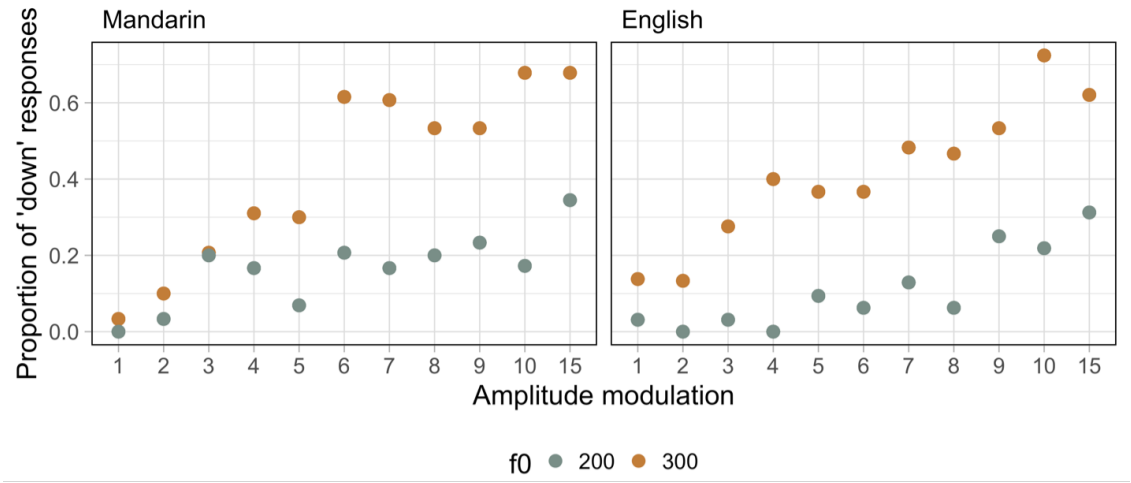


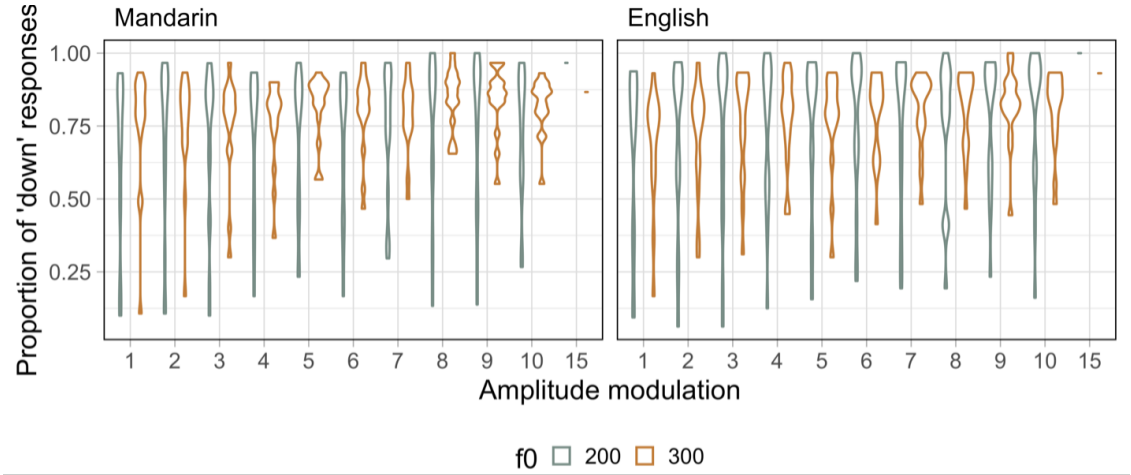
Figure 4.3: Proportion of 'down' responses as a function of modulation types. no: unmodulated; am: amplitude modulation; fm: frequency modulation; fa: combined frequency and amplitude modulation.

Nested model comparisons show that participants' choice varied as a function of modulation type [$\chi^2(3) = 1786.61, p < .001$], f0 [$\chi^2(1) = 6.44, p < .05$], and their interactions [$\chi^2(3) =$

$55.24, p < .001$]. Within the modulation types, all types had significant effects on the 'down' responses: frequency modulation had a stronger effect ($\beta = 5.00, p < .001$) than amplitude modulation ($\beta = 2.21, p < .05$), whereas the combined frequency and amplitude modulation had the strongest effect ($\beta = 5.32, p < .001$), showing an additive effect. Pairwise comparisons confirmed this order of effect strength: combined modulation $>$ frequency modulation $>$ amplitude modulation $>$ no modulation. The positive coefficients indicate that modulation biased listeners to choose more 'down' responses when categorizing the novel period-doubled tones in a novel language. Though the main effect of f0 condition and the interaction between f0 and modulation were significant, within-factor levels did not differ from the baseline ($200Hz$, unmodulated). However, there was an interaction effect of amplitude-modulated tokens at $300Hz$ when compared to the other types of modulation on the proportion of low tone responses ($p < .001$).

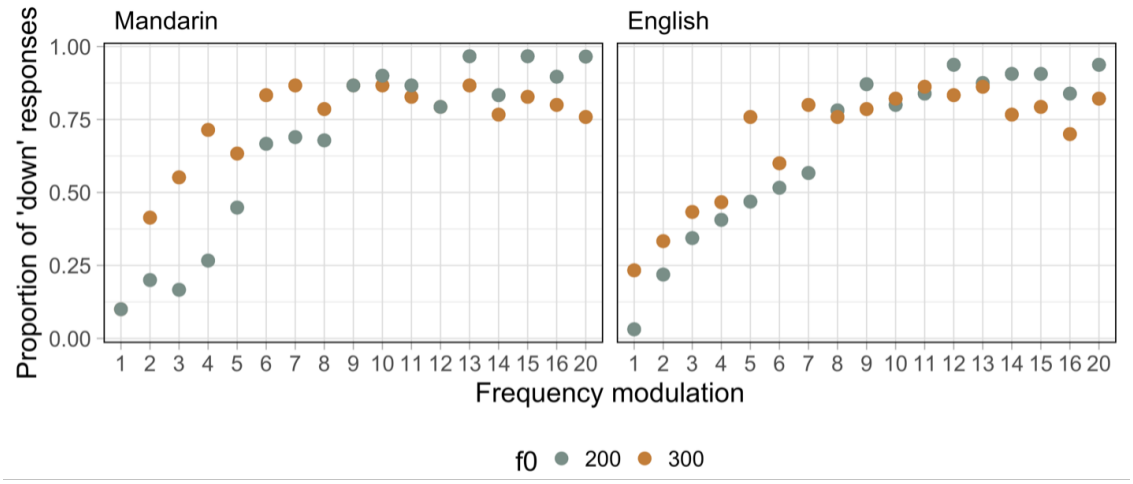


(a)

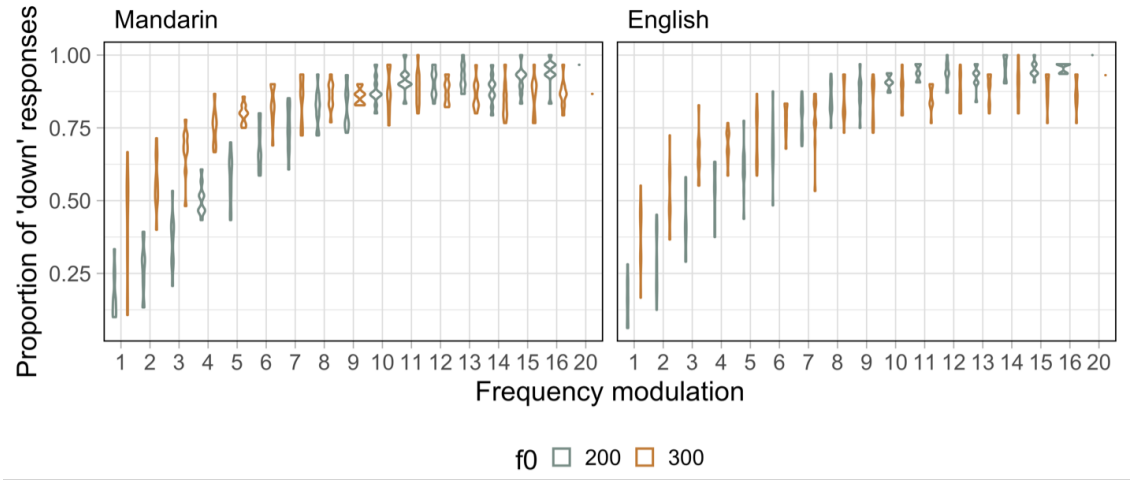


(b)

Figure 4.4: Proportion of 'down' responses as a function of the varying degrees of amplitude modulation in tokens of pure amplitude modulation (a) and combined modulation (b). The extreme anchor points at the modulation degree of 15 only have four trials per participant, thus showing fewer data points.



(a)



(b)

Figure 4.5: Proportion of 'down' responses as a function of the varying degrees of frequency modulation in tokens of pure frequency modulation (a) and combined modulation (b). The extreme anchor points at the modulation degree of 20 only have four trials per participant, thus showing fewer data points.

Amplitude modulation Within the amplitude modulation subset, the maximum parsimonious structure is shown in equation (4.3). Here, I model the responses as a function of the continuous variable *modulation degree* manipulated by the amplitude ratios (which resulted in 11 amplitude-modulated tokens without repetition), *f0*, their interactions, *language*, and *music* experiences. Given that purely amplitude-modulated tokens are fewer (1296 data points), the random intercept of repetition was omitted to achieve better model convergence.

$$choice \sim modulation\_degree * f0 + language + music + (1|subject) \tag{4.3}$$

Again, nested model comparisons showed that only modulation degree $[\chi^2(1) = 148.62, p < .001]$ and f0 $[\chi^2(1) = 31.97, p < .001]$ were significant, whereas other factors were not. In the detailed model output, stronger degrees of amplitude modulation ($\beta = 0.21, p < .001$), and a higher f0 at $300Hz$ ($\beta = 1.89, p < .001$) biased listeners to choose more 'down' responses.

Frequency modulation The maximum parsimonious structure remains the same as equation (4.3) within the frequency modulation subset. There were 17 frequency modulation degrees and 2006 data points in total. Nested model comparisons showed that only modulation degree $[\chi^2(1) = 452.99, p < .001]$ and its interaction with f0 $[\chi^2(1) = 46.12, p < .001]$ were significant, whereas other predictors were not. The intercept of the frequency modulation model had a higher estimate ($\beta = -1.64, p < .01$) than that of the amplitude modulation model ($\beta = -3.09, p < .001$), meaning overall fewer 'up' and more 'down' responses were observed here. As for within-factor levels, stronger degrees of frequency modulation ($\beta = 0.36, p < .001$), and a higher f0 at $300Hz$ ($\beta = 1.82, p < .001$) biased listeners to choose more 'down' responses, even though no main effect of f0 was found. In addition, a higher f0 at $300Hz$ with the same modulation degrees showed a small effect of biasing participants to choose fewer 'down' responses than the $200Hz$ ($\beta = -0.19, p < .001$).

Combined amplitude and frequency modulation For the subset of combined amplitude and frequency modulation (which forms the majority of the data at 19039 data points), similar to the omnibus model, I started the analysis with the maximum structure as in equation (4.4). However, due to convergence issues, the maximum parsimonious structure is given in equation (4.5) after

nested model selection showing that language or music did not account for significant variance of the response variable.

$$choice \sim amplitude\_degree * frequency\_degree * f0 + language + music + (1|subject) + (1|rep) \tag{4.4}$$

$$choice \sim amplitude\_degree * frequency\_degree * f0 + (1|subject) + (1|rep) \tag{4.5}$$

In equation (4.5), the categorization choice varied as a function of fixed factors: amplitude modulation degree [$\chi^2(1) = 129.40, p < .001$], frequency modulation degree [$\chi^2(1) = 3726.10, p < .001$], f0 [$\chi^2(1) = 5.2, p < .05$], the two-way interactions between two types of modulation degrees [$\chi^2(1) = 20.21, p < .001$], frequency modulation degree and f0 [$\chi^2(1) = 267.42, p < .001$], and marginally between amplitude modulation degree and f0 [$\chi^2(1) = 3.82, p = .05$]; the three-way interaction among two modulation degrees and f0 was significant [$\chi^2(1) = 5.79, p < .05$] as well. The model output confirmed the ANOVA results: stronger degrees of frequency ($\beta = 0.38, p < .001$) and amplitude modulation ($\beta = 0.090, p < .001$), a higher f0 ($\beta = 1.35, p < .001$) increased the 'down' responses; interestingly, the $300Hz$ tokens biased fewer 'down' responses in frequency-modulated tokens ($\beta = -0.13, p < .001$) but more 'down' responses in amplitude-modulated tokens ($\beta = 0.085, p < .01$); besides, the three-way interaction showed a small effect that stronger degrees of both modulations in $300Hz$ tokens led to fewer 'down' choices ($\beta = -0.008, p < .05$).

Ambiguity in pitch perception Table 4.3 shows the top 10 conditions where the standard deviation was smallest and largest, respectively. A mix of both f0s of the stimuli was observed, though more tokens of $200Hz$ appeared among the less ambiguous stimuli to be categorized. A relatively clear pattern is found whereby, when both amplitude and frequency modulation degrees were at extremes, the pitch was identified more consistently across participants and repetitions,

suggesting a more stable pitch. On the other hand, when both types of modulation were not at the extremes, especially when frequency modulation was relatively smaller, participants' perception was less consistent and both were selected, suggesting a more ambiguous pitch.

Table 4.3: Perceived pitch in least ambiguous (top) and most ambiguous (bottom) period-doubled tokens based on standard deviation of response. Unmodulated is indicated by 0 in steps; extreme values are 20 for frequency modulation (*fm*) and 15 for amplitude modulation (*am*). Response was coded as either '0' for 'up' or '1' for 'down'.

Stimulus f0 (*Hz*)	Fm step	Am step	Mean response	SD response
200	0	0	0.016	0.127
200	0	1	0.016	0.127
200	0	2	0.016	0.127
200	11	8	0.984	0.127
200	20	15	0.984	0.127
200	14	3	0.984	0.128
200	16	8	0.984	0.128
300	14	9	0.983	0.129
200	14	6	0.968	0.178
200	16	9	0.968	0.178
300	1	9	0.500	0.505
300	3	1	0.509	0.505
200	3	4	0.500	0.504
300	2	4	0.491	0.504
200	4	8	0.500	0.504
300	0	8	0.508	0.504
200	4	4	0.508	0.504
300	2	5	0.517	0.504
300	3	0	0.483	0.504
200	5	1	0.483	0.504

4.2.6. Results: reaction time

The most parsimonious models used to predict log-transformed reaction time (*log.rt*) in the omnibus model, amplitude modulation subset, frequency modulation subset, and combined amplitude and frequency modulation are shown in (4.6-4.9).

$$log.rt \sim modulation\_type * f0 + language + music + (1|subject) + (1|rep) \quad (4.6)$$

$$log.rt \sim amplitude\_degree * f0 + language + music + (1|subject) \quad (4.7)$$

$$log.rt \sim frequency\_degree * f0 + language + music + (1|subject) + (1|rep) \quad (4.8)$$

$$log.rt \sim amplitude\_degree * frequency\_degree \times f0 + (1|subject) + (1|rep) \quad (4.9)$$

Reaction time largely did not show differences based on the predictors. In the omnibus model, only the main effect of f0 [$F(1) = 3.15, p = .081$] and the interaction between modulation type of f0 [$F(3) = 2.38, p = 0.068$] were marginally significant. None of the within-factor levels were significant, however. In the amplitude modulation subset, none of the fixed factors or within-factor levels were significant. In the frequency modulation subset, only the main effect of f0 was marginally significant [$F(1) = 3.15, p = .081$]. Stronger degrees of frequency modulation led to a marginally quicker response ($\beta = -0.00097, p = .066$); so did a higher f0 at $300Hz$, only this effect reached significance ($\beta = -0.034, p < .05$). In the combined modulation subset, the main effect of f0 was marginally significant [$F(1) = 3.35, p = .072$], and the interaction between degrees of frequency modulation and f0 was significant [$F(1) = 17.91, p < .001$]. When the f0 was at $300Hz$, participants were faster to respond ($\beta = -0.038, p < .01$), but were slightly slower to respond for frequency-modulated tokens with $300Hz$ ($\beta = 0.0013, p < .05$).

4.2.7. Interim discussion

The results largely confirm our hypotheses: first, for all the modulation types, the higher the modulation degree, the more ↓ responses there were. Second, frequency modulation leads to more responses of the ↓ tone (relative to amplitude modulation), and also causes the rising trend of the proportion of 'low tone' responses to be steeper. Pitch perception during period doubling varies by degree of the modulation, such that even at low frequency modulation degrees (around

5), nearly 50% of the tokens were categorized as ↓. And only when the amplitude modulation extent is at least 6 or 7 (note the larger step in amplitude modulation compared to the frequency one), the proportion of ↓ responses is comparable to that with frequency modulation. When the tokens have combined modulation from both types, frequency modulation drives the trend of perceiving more ↓ tone responses, not amplitude modulation. These are consistent with previous findings by Sun and Xu (2002) and Bergan and Titze (2001).

Third, the two groups of speakers of a tonal language (Mandarin) versus non-tonal language (English) do not perceive period doubling differently. The responses and patterns were largely in agreement for both language groups, contrary to predictions. This suggests that pitch perception during period doubling is not language-specific, at least for Mandarin versus English speakers. Also contrary to predictions, music experience does not play a role in shaping the perception of these tones. Lastly, while previous studies only investigate either pure amplitude or frequency modulation, this study contributes to the perception of period doubling by adding the combined modulation, which is commonly found in speakers' production. An interesting interaction is observed here – when f0 is higher ($300Hz$), especially in amplitude-modulated tokens, a larger proportion of tokens were perceived as the ↓ tone. This finding is consistent with the results by Sun and Xu (2002) where in a pitch matching task participants tended to choose a lower pitch with tokens of $220Hz$ than $140Hz$ when the modulation extent reached 30% and beyond. However, Bergan and Titze (2001) had a different finding that a lower f0, rather than a higher one, had an earlier identification boundary at half of the original f0 for both types of modulation. Yet, in the current study, the multiple two-way or three-way interactions, and the interactional effects of a higher f0 on different modulation types, do not always agree in terms of the same direction. The different findings may also be attributed to the different setups of the current experiment versus the previous one. For example, Bergan and Titze (2001) used an ABX forced alternative task and participants chose whether the modulated token X is more similar to the higher f0 A or lower f0 B. Participants could be attending to the bottom-up psychoacoustic details of the particular token and overtly comparing it to one of the two frequencies, rather than categorizing linguistic tones

upon hearing the stimulus. In future work we should probe further, how different base f0s influence pitch and tone perception during period doubling, especially when interacting with different types of modulation.

In sum, with an artificial language learning paradigm, the findings of the current study are not limited to the range of psychoacoustic applications of period doubling, but also have linguistic implications. Based on the results, it is predicted that period doubling signals low tones in languages regardless of frequency or amplitude modulation, even when the f0 is high. This lines up with findings from Mandarin tone production (Huang, 2022; Chapter 3) where period doubling is far more frequently employed than vocal fry across all four different tones including high, low, rising and falling contours. Moreover, if a high tone occurs in the utterance-final position that likely induces period doubling (Chapter 3), the tonal distinction may be blurred by the presence of period doubling because it would signal a low tone instead.

4.3. Experiment 2: Pitch shadowing of resynthesized period-doubled tones

4.3.1. Stimuli

The stimuli were the same as Experiment 1.

4.3.2. Participants and procedure

The participants were the same as Experiment 1. However, due to equipment failure or participants' error during the experiment, seven recordings were not usable; thus, only 54 files were segmented, annotated, and processed (Mandarin: 25; English: 29). Immediately after the perception experiment, participants were asked to participate in the shadowing experiment. They were recorded using a desk-mounted Oktava MK-519 Microphone, with the SHURE SRH440 or AKG K-55 stereo headphones on. They were asked to produce the tones heard three times by

imitating the pitch and voice of the stimuli. They were also allowed to play the audio as many times as needed and were encouraged to practice as many times as needed before they were ready to produce the sounds. To mark the intended productions by participants, they were told to press a clicker before starting to produce the real targets. Prior to the experiment, they were instructed to produce tone sweeps to assess their vocal range when comparing to the pitch production of period doubling. I adopted a combination of the unprompted and prompted procedure described in Keating and Kuo (2012). Participants read through instructions of how to produce tone sweeps and heard examples produced and recorded by the author before proceeding to produce on their own. The instructions were as follows:

1) Gradually raise the pitch of your voice to the highest pitch until you feel your voice breaks;

2) Gradually lower to the lowest pitch until your voice breaks;

3) Repeat the raising and lowering again, ending in the lowest pitch.

At the end of this experiment, participants were asked to fill out a post-experiment questionnaire asking what they thought of the purpose of both experiments in the study, if they discovered any recurring patterns, and what strategies they have been using if any. Most of them noted pitch perception, overtones, and the increasingly salient lower frequency component.

4.3.3. Analysis

The recordings were segmented to recover the productions of [a] in PRAAT. All three repetitions were segmented, and their f0s and acoustic characteristics were extracted with a window size of 25 milliseconds using PRAAT cross-correlation algorithm embedded in VoiceSauce (Shue et al., 2011), which outputs a value at every millisecond via interpolation. Because the stimuli only contain level pitches, and f0 dynamics is irrelevant upon initial checking of the sustained shadowing productions, I extracted the average f0, max, and min f0 measurements over the three repeated to-

kens for a given condition. For each condition, the standard deviation across the repetitions were calculated to infer the consistency of shadowing during period doubling. According to our hypothesis, if the pitch is ambiguous to imitate (i.e., when participants might hear two competing pitches for a token), they should exhibit greater intra-token variability during shadowing; thus, a larger standard deviation across repetitions is expected to signal pitch ambiguity of period doubling in a certain condition based on its modulation extent. Each individual's vocal range was assessed by extracting the maximum and minimum f0s in their tone sweeps and coded as *max.range* and *min.range* as covariates included in the statistical models for predicting mean/max/min f0s. This was to control for the individuals' different vocal ranges while investigating the effects of type and degree of modulation on their shadowing productions.

Voice quality correlates such as H1*–H2* (based on PRAAT f0), harmonics-to-noise ratio (HNR), subharmonic-to-harmonic ratio (SHR), and strength of excitation (SoE) were extracted using VoiceSauce. H1*–H2* is one of the most common spectral tilt measures that calculates the difference between the first and second harmonics, with formants and bandwidths corrected across tokens. In general, the lower the H1*–H2*, the higher the degree of glottal constriction. Also see Garellek (2019) for a detailed overview of these acoustic measures and the implications for their use. H1–H2 is also used in creaky voice detectors developed by Drugman et al. (2020). HNR measures the amount of noise relative to the harmonics; a lower HNR value indicates a noisier quality. Here, I discuss the HNR calculated based on a frequency window of $0 - 500Hz$. HNR was found to be useful in predicting creaky voice in low falling tones of Mandarin, Cantonese, and White Hmong (Kim et al., 2020), and is also widely used in assessing voice disorders. SHR is a measure proposed by Sun (2002) to capture alternate cycles in speech signal, quantifying the subharmonic to harmonic ratio through spectrum shifting. Typically, the higher the SHR, the stronger the subharmonics present in the speech signal. SoE is an energy measure to quantify the strength of excitation around the instant of glottal closure (Murty and Yegnanarayana, 2008), and was found to signal creaky voice (Garellek et al., 2021), whereby lower values are associated with more constriction, and thus a creakier quality.

Here, I discuss different measurements of f0 and acoustic correlates using fixed factors *modulation type* (no: unmodulated, am: amplitude modulation, fm: frequency modulation, fa: frequency and amplitude modulation) and *f0* (200Hz, 300Hz) and their interactions. *Language* and *music* were also included, even though they were found to be not significant for the perception experiment in Section 4.2.5. *Subject* was included as a random intercept. For voice quality models, imitated f0 is included as a covariate to account for variations brought about by f0 given the correlation between f0 and acoustic measures. Nested ANOVA model comparisons were used to assess each fixed factor and *lmerTest* to evaluate the significance and effect size of the within-factor levels for each linear mixed-effects model. Further, I probe whether the productions are correlated with the perceptual results in Experiment 4.2; in other words, whether f0 or voice quality measures of the produced period-doubled tokens can be predicted by their prior categorical perceptual responses.

Data were trimmed to remove outliers determined by log-transformed f0 values that are larger than 2.5 standard deviations from the mean (1% of the original data). Following Yuan and Liberman (2014), for better interpretation, the shadowed f0 was converted to semitones according to the equation in (4.10). To normalize for individual variation, the f0 base was speaker dependent – chosen from 5th percentile of all f0 values of each speaker. The outliers of voice quality correlates were also removed if greater or less than 2.5 standard deviations from the mean. The maximal model structures for shadowed f0 and voice quality correlates are shown in (4.11-4.13).

$$Semitone = 12 \times \log_2(\frac{f0}{f0\_base}) \tag{4.10}$$

- $Mean/max/min.f0(semitone) \sim max/min.range(semitone) + modulation\_type * f0 + language + music + (1|subject)$ (4.11)
- $Mean.f0(semitone) \sim max/min.range(semitone) + modulation\_degree * f0 + language + music + (1|subject)$ (4.12)
- $Mean.z.H1H2/HNR/SHR/SoE \sim mean.z.log.f0 + modulation\_type * f0 + language + music + (1|subject)$ (4.13)

4.3.4. Results: shadowing

Imitation of pitch Nested model comparisons showed that participants' max and min vocal ranges were significant for mean f0 [$max : F(1) = 14.50, p < .001; min : F(1) = 8.16, p < .01$] and max f0 [$max : F(1) = 12.85, p < .001; min : F(1) = 7.32, p < .01$], but only min range was significant for min f0 [$min : F(1) = 11.32, p < .01$]. This is expected given that speakers may reach a higher or lower f0 by approaching the upper or lower limit of their vocal range. Modulation type was also significant in predicting mean f0 [$F(3) = 176.18, p < .001$], max f0 [$F(3) = 115.16, p < .001$], and min f0 [$F(3) = 70.58, p < .001$]. The interaction between modulation type and f0 condition was significant in predicting all the f0 measures ($p < .001$), whereas f0 was only significant in min f0 [$F(1) = 6.45, p < .05$], and marginally so for max f0 ($p = .073$). Language or music background did not play a role in predicting the imitated f0s upon hearing the period-doubled stimuli, as in the perception study for categorizing period-doubled tones.

When max and min vocal ranges were significant, they had a positive effect on the different f0 variables, as expected. For example, a person with a mean f0 of $120Hz$ might shadow $300Hz$ by producing $150Hz$ (i.e., an octave lower than the target, but a frequency that is within their normal speaking range). The fixed effects of interest, different modulation types including amplitude, frequency, or combined modulation had negative effects on the different f0 measures. Though the interaction term contributed to account for model variance, and in tokens of $300Hz$, different

signs were seen for amplitude modulation compared to the other types, no particular within-factor levels were significant. A detailed model summary is in Table 4.4. Other effects not reported were not significant in any of the f0 models.

Table 4.4: Significant fixed effects in predicting mean, max, and min imitated f0. ***: $< .001$; **: $< .01$; *: $< .05$. (am: amplitude modulation, fm: frequency modulation, fa: frequency and amplitude modulation).

Fixed effects	Mean f0		Max f0		Min f0	
	Estimate	*P* value	Estimate	*P* value	Estimate	*P* value
(Intercept)	2.82	0.33	4.26	0.06	4.37	0.20
Max.range	0.32	$< .001$ ***	0.082	0.201	0.36	$< .001$ ***
Min.range	0.12	0.006 **	0.108	0.002 **	0.14	0.009 **
Typeam	-1.46	$< .001$ ***	-0.80	0.226	-1.22	0.009 **
Typefm	-3.91	$< .001$ ***	-3.32	$< .001$ ***	-3.49	$< .001$ ***
Typefa	-4.28	$< .001$ ***	-3.59	$< .001$ ***	-3.80	$< .001$ ***
Typeam:f0300	-0.84	0.20	-0.59	0.54	-0.96	0.17
Typefm:f0300	1.11	0.084	1.29	0.18	0.86	0.22
Typefa:f0300	1.13	0.073	1.29	0.17	1.026	0.13

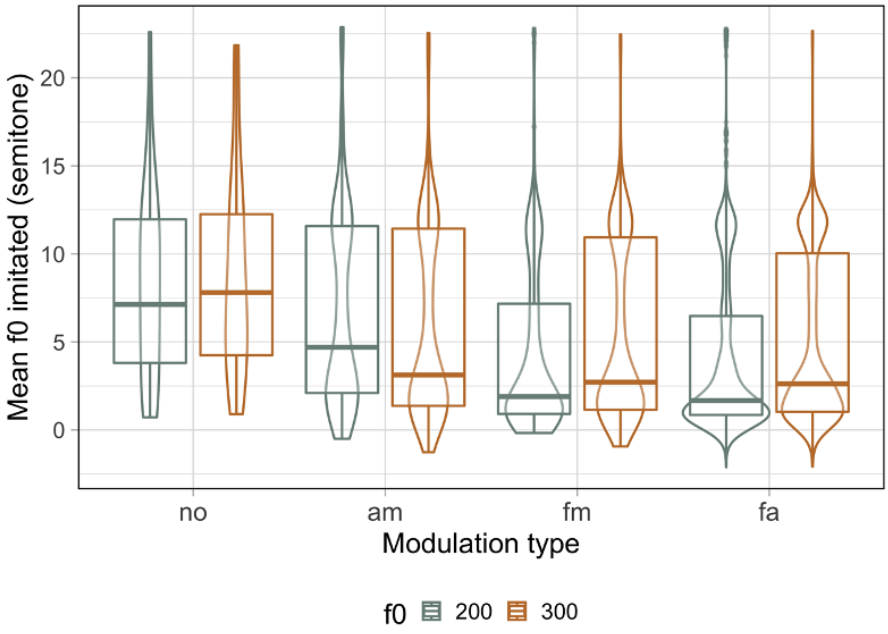


Figure 4.6: Imitated mean f0 (semitone) varied by modulation types. no: unmodulated; am: amplitude modulation; fm: frequency modulation; fa: combined frequency and amplitude modulation.

Figure 4.6 shows the mean f0 conditioned by the various modulation types. Two generalizations were observed. First, in both f0 conditions, amplitude and frequency modulations were associated with lower imitated f0, and combined modulation had an additive contribution to this main effect. Second, with the main lowering effect of modulation, imitation of amplitude-modulated tokens of $300Hz$ tended to behave differently from $200Hz$. The intrinsic differences between 200 and $300Hz$ reversed such that imitated f0 in this condition tended to be even lower than the $200Hz$ tokens, especially when compared with other (un)modulated tokens.

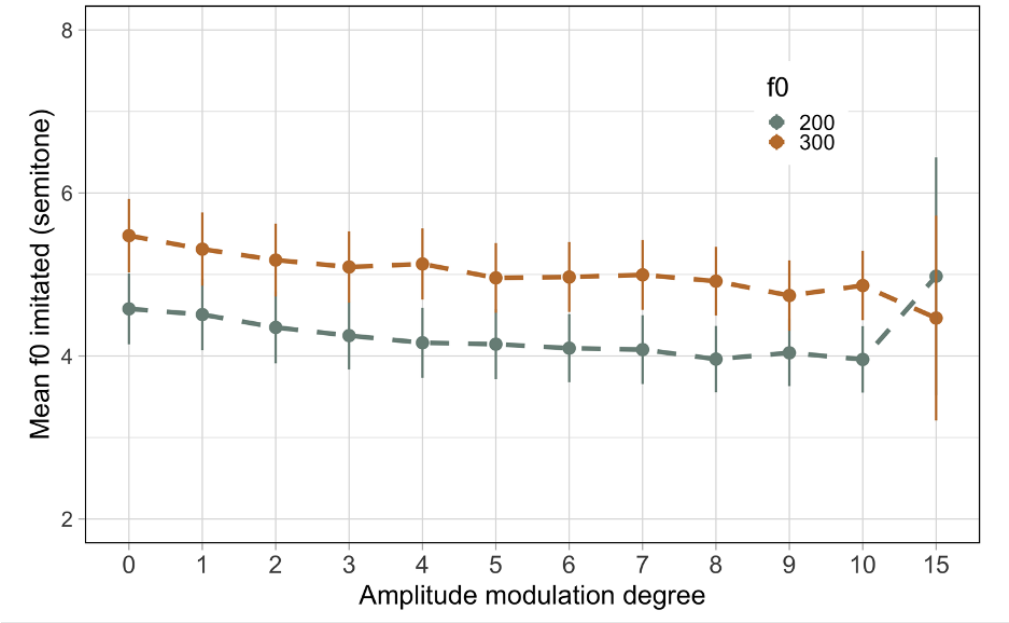
Further, in subset models of only tokens of amplitude, frequency, or combined modulation, results showed that imitated mean f0 decreased as a function of the increasing steps of modulation degrees. In general, the effect size of frequency modulation was found stronger than that of amplitude modulation across all subset models and the frequency-modulated tokens of $300Hz$ led to higher mean f0, which are consistent with the effect of modulation degree on the categorization results of 'down' responses in perception. Table 4.5 summarizes the significant results.

Table 4.5: Significant fixed effects in predicting mean imitated f0 imitated in subsets of modulation type. ***: $< .001$; *: $< .05$.

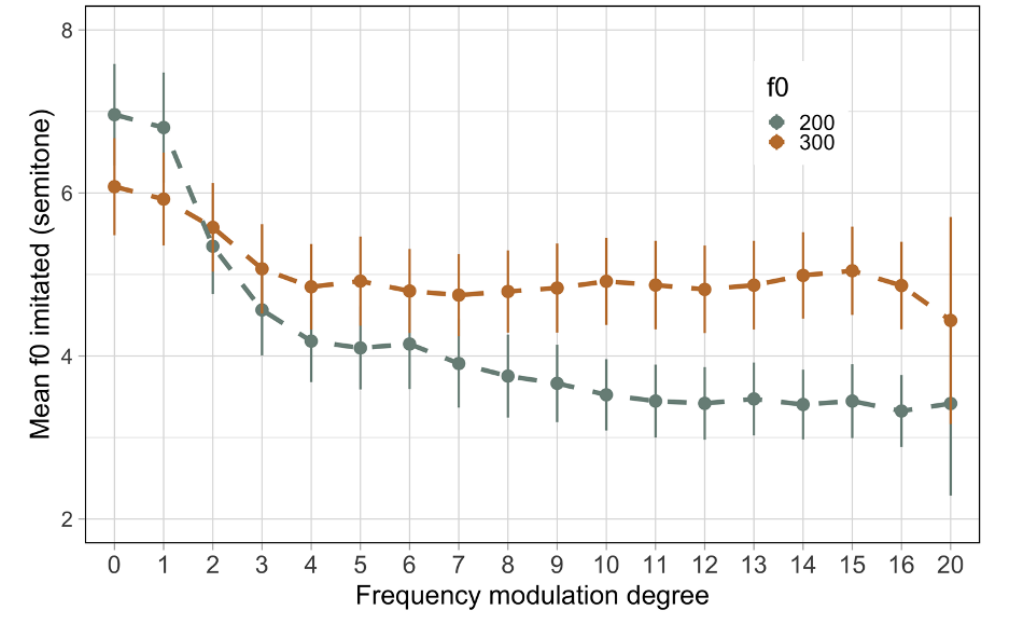
Modulation type	Fixed effects	Estimate	SE	T value	P value
Amplitude	am_degree	-0.14	0.03	-5.36	$<.001$ ***
Frequency	fm_degree	-0.19	0.02	-9.29	$<.001$ ***
Amplitude + frequency	am_degree	-0.10	0.01	-6.54	$<.001$ ***
	fm_degree	-0.19	0.01	-16.85	$<.001$ ***
	fm_degree:f0300	0.13	0.01	10.55	$<.001$ ***
	am_degree:fm_degree	0.0078	0.00	5.12	$<.001$ ***
	am_degree:fm_degree:f0300	-0.0027	0.00	-2.01	0.04 *

Two plots in Figure 4.7 illustrate the lowering effect using all data points conditioned by two sets of modulation degrees, respectively. The decreasing trend in imitated f0 according to amplitude modulation steps was less steep and spanned across fewer semitones than frequency modulation, though with larger variation at the strongest degree 15, possibly with fewer data points. Moreover, lowering trends of imitated f0 had different slopes according to frequency

modulation steps between tokens of 200 and 300Hz. Imitated f0 of 200Hz tokens had a sudden drop around the second degree and its range of variation is larger than that of 300Hz tokens.



(a)



(b)

Figure 4.7: Imitated mean f0 (semitone) varied by amplitude (a) or frequency (b) modulation degrees. 95% confidence intervals are plotted around the mean of a given condition across all participants.

Ambiguity in pitch shadowing Table 4.6 shows the top 10 conditions where the standard deviation of imitated f0 was smallest and largest, respectively. The fact that the majority of tokens had an f0 of 200Hz suggests that this lower f0 was both easier and harder than 300Hz for participants to process and then imitate – its pitch either being less ambiguous or very ambiguous. Compared to the perceptual results in Table 4.3, which showed that extreme frequency modulations led to less ambiguous perception and modulations in the middle led to more ambiguity, the pattern here is slightly different. Overall, when the frequency modulation extent was higher (the longer period is at least 1.7 times the shorter one), the pitch was less ambiguous for the purposes of imitation. But when the frequency modulation was reaching the lowest values, even if amplitude modulation was at the extreme, participants' productions were more variable and suggest more ambiguity.

Table 4.6: Imitated f0 (semitone) in ten least ambiguous (top) and ten most ambiguous (bottom) period-doubled tokens based on standard deviation of f0.

Stimulus f0 (Hz)	Fm step	Am step	Mean f0 (st)	SD f0 (st)
200	12	1	2.97	3.24
200	16	1	3.00	3.35
200	8	2	3.23	3.35
200	13	2	3.06	3.40
200	8	8	3.36	3.40
200	14	1	3.38	3.42
200	10	6	3.28	3.42
200	7	7	3.26	3.44
200	11	8	2.86	3.50
200	16	4	3.38	3.58
200	1	5	6.35	6.70
200	1	7	7.12	6.51
200	1	9	6.42	6.24
200	0	8	6.09	6.19
200	1	2	7.49	6.10
200	2	2	5.94	6.09
200	0	5	7.20	6.08
200	0	15	6.58	6.05
300	0	3	6.76	6.02
300	0	1	8.09	5.97

Imitation of voice quality Besides different f0 measures, four acoustic correlates of voice quality were investigated using the model in equation 4.11. A summary of results is in Table 4.7. The covariate log-transformed f0 (z-score) was significant in all the voice quality models ($p < .001$). For H1*–H2*, stimulus f0 was significant [$F(1) = 8.70, p < .01$], and modulation type was marginally so [$F(3) = 2.50, p = .057$]; within-factor levels showed that H1*–H2* was higher in tokens of $300Hz$ than $200Hz$ ($\beta = 0.044, p < .01$), as expected for a higher f0. For HNR, only the main effect of modulation type was significant [$F(3) = 6.38, p < .001$]. The overall intercept for unmodulated $200Hz$ tokens was negative ($\beta = -0.19, p < .01$). Contrary to expectation of roughness sounding of period-doubled tokens, all modulation types had a positive effect towards HNR: amplitude ($\beta = 0.16, p < .001$), frequency ($\beta = 0.21, p < .001$), combined ($\beta = 0.20, p < .001$) modulation, meaning that modulation led to a less noisy voicing during imitation. For SHR, no effect besides the f0 covariate was found. For SoE, surprisingly, music experience was significant [$F(2) = 3.97, p < .05$], with a small effect showing that musicians in general had a lower SoE than subjects who have not had music experience ($\beta = -0.17, p < .05$). This suggests that musicians produced the tokens with weaker voicing, likely due to increased glottal constriction. It is also possible that they are vocally more efficient or have learned to control their voicing intensity.

Table 4.7: Significant fixed effects in predicting imitated voice quality correlates. ***: $< .001$; **: $< .01$; *: $< .05$. (am: amplitude modulation, fm: frequency modulation, fa: frequency and amplitude modulation)

	Fixed effects	Estimate	SE	T value	*P* value
H1*–H2*	f0: 300 vs. 200	0.04	0.02	2.95	0.005 **
HNR	(Intercept)	-0.19	0.06	-3.10	0.002 **
	Typeam	0.16	0.06	2.83	0.005 **
	Typefm	0.21	0.06	3.69	$<$.001 ***
	Typefa	0.20	0.06	3.65	$<$.001 ***
SoE	Music: yes vs. no	-0.17	0.07	-2.36	0.022 *

In sum, modulation type only had some effects on HNR by raising its value to be less noisy compared to the modeled overall low intercept, and the stimulus f0 at $300Hz$ also raised

the spectral tilt in the expected direction of a higher f0, and music experience led to a weaker voicing, as reflected in lowered SoE. SHR was not affected by modulation type or stimulus f0. No generalizations can be drawn based on the current results regarding the effect of modulation on imitated voice quality. It is possible that participants focused more on imitating f0 than anything else.

4.3.5. Results: perception and shadowing

Using a series of perception and production experiments, one of the goals of the present study was to investigate the potential relation between perception and production of period doubling. Recall that participants had performed the perception task first, which was then followed by the shadowing task. I added a categorical variable *choice*, the response variable in Experiment 4.2, and its interactional terms with the existing fixed effects *modulation type* and *f0*, to predict the mean f0 and voice quality correlates of the produced tokens in the revised shadowing models, as shown in (4.14-4.15).

- $Mean.f0(semitone) \sim max/min.range(semitone) + modulation\_type * f0 * choice + language + music + (1|subject)$ (4.14)
- $Mean.z.H1H2/HNR/SHR/SoE \sim mean.z.log.f0 + modulation\_type * f0 * choice + language + music + (1|subject)$ (4.15)

Pitch matching Nested models showed that the three-way interaction among modulation type, f0, and perceptual choice was not significant; thus, the final model only included the fixed factors in question and the paired interactions among them. Here, I only discuss the effects related to the perceptual responses (choice, modulation_type:choice, f0:choice, and modulation_type:f0:choice), because other fixed effects have been discussed in the previous section.

The imitated f0 averaged across three repeated tokens produced by participants can be predicted using their own categorical response choice in perception. The main effect of choice [$F(1) = 571.83, p < .001$], and its interactions with modulation type [$F(3) = 5.62, p < .001$] and f0 [$F(1) = 113.44, p < .001$] were all significant. Specifically, when participants chose 'down' responses, their imitated f0 was also lower ($\beta = -3.44, p < .05$), showing a positive correlation between pitch perception and production. Further, among the 'down' choices, when the stimulus f0 was $300Hz$, the imitated f0 was slightly higher ($\beta = 0.081, p < .001$) than when the stimulus was $200Hz$. Figure 4.8 shows the relationship between imitated f0 in shadowing and the categorical responses in perception, such that 'down' responses drive the production to be lower, regardless of original stimulus f0. However, the amplitude-modulated tokens of $200Hz$ within 'down' responses seem to be an exception. In addition, a clear lowering trend according to modulation types is largely observed within 'up' responses for both stimulus f0s, except that amplitude modulated $300Hz$ tokens within 'up' responses had lower imitated f0.

Overall, the results demonstrate that perceptual responses well correlate with imitated f0 in shadowing, while amplitude modulation in different stimulus f0s had different interaction effects.

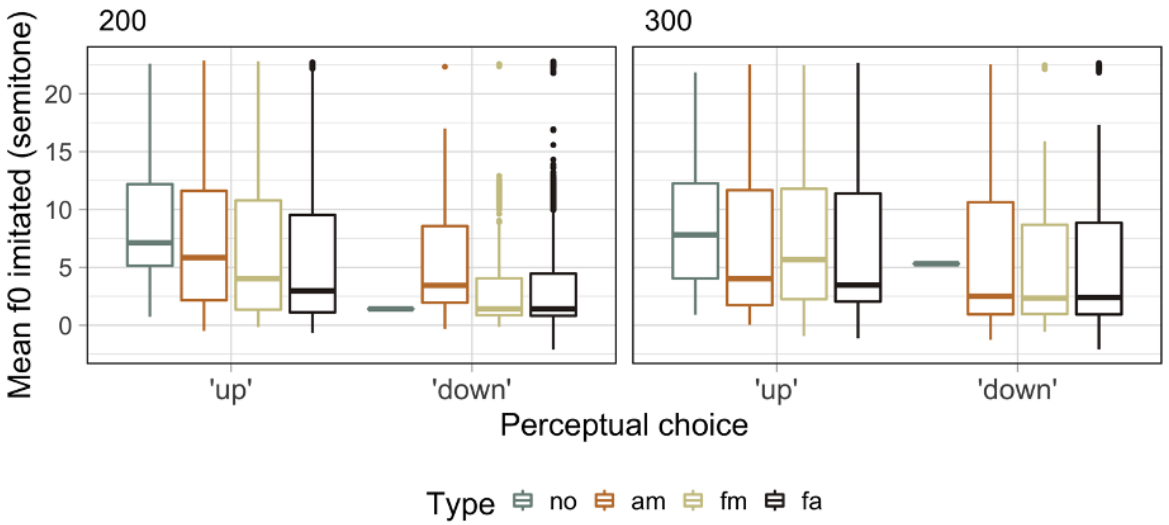


Figure 4.8: Imitated f0 across different f0 conditions in Experiment 2 as a function of perceptual responses in Experiment 1. There were only 49 unmodulated tokens ('no'), and those categorized as 'down' were very few. (no: unmodulated; am: amplitude modulation; fm: frequency modulation; fa: frequency and amplitude modulation).

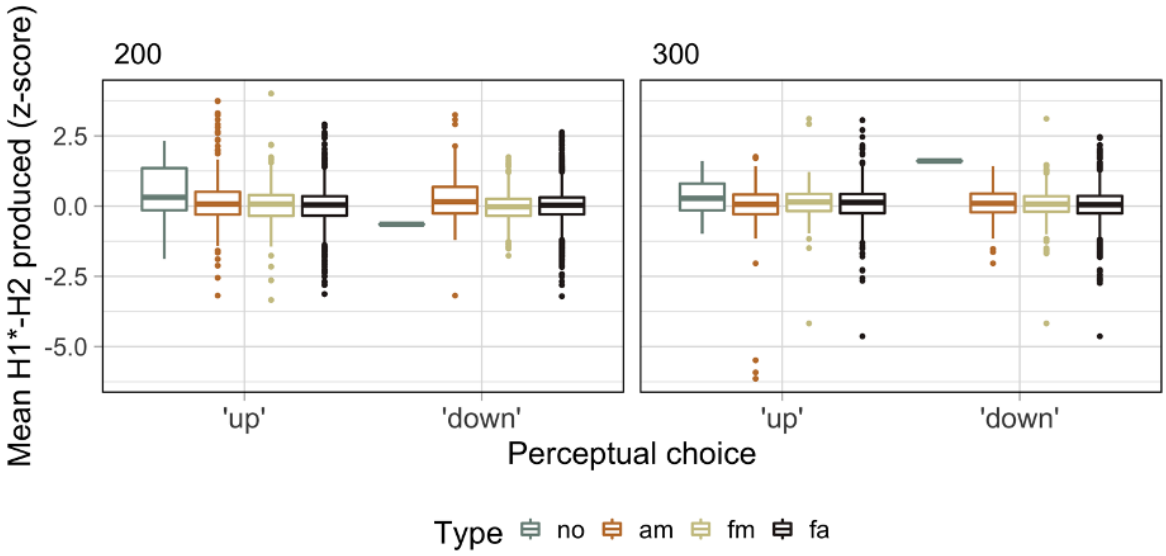
Voice matching Here I probe the relationship between the prior categorical perception and voice quality correlates in shadowed f0. For each of the models of the acoustic correlates, if the three-way interaction among choice, modulation type, and f0 were not significant, the models were then adjusted to only include two-way interactions involving choice and its main effect. The model comparisons output used to assess the perceptual factors is shown in Table 4.8.

As for H1*–H2*, the main effect of choice as well as its interaction with stimulus f0 and three-way interaction with modulation type and stimulus f0 were significant. When participants chose 'down' in modulated tokens of 300Hz in perception, their imitated H1*–H2* across different modulation types tended to be even lower ($\beta s < -1.75, p < .05$) than when they chose 'up', consistent with our previous findings in perception and imitated f0. In addition, in the revised model, amplitude modulation led to lower spectral tilt values than the unmodulated with a smaller

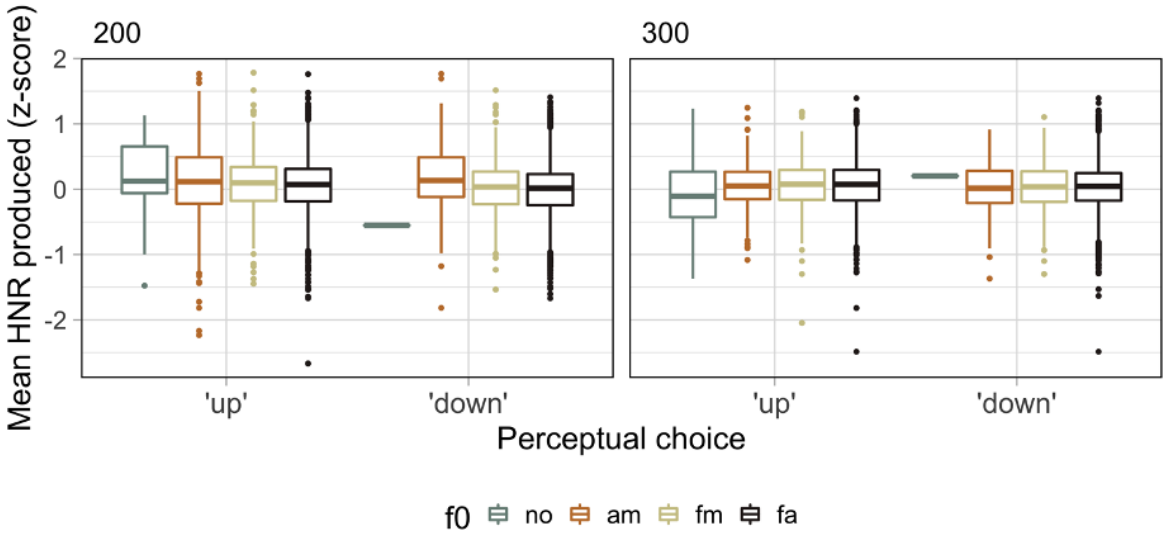
Table 4.8: Significance of independent and interactional predictors involving perceptual responses. ***: $< .001$; **: $< .01$; *: $< .05$.

	H1*–H2*	HNR	SHR	SoE
choice	$F(1) = 28.64$; $p < .001***$	$F(1) = 0.71$; $p = .40$	$F(1) = 6.85$; $p < .01**$	$F(1) = 26.44$; $p < .001***$
modulation_type:choice	$F(3) = 2.59$; $p = .051$	$F(3) = 0.11$; $p = .95$	$F(3) = 5.69$; $p < .001***$	$F(3) = 3.24$; $p < .05*$
f0:choice	$F(1) = 13.70$; $p < .001***$	$F(1) = 2.33$; $p = .13$	$F(1) = 1.39$; $p < .24$	$F(1) = 0.88$; $p < .35$
modulation_type:f0:choice	$F(3) = 2.75$; $p < .05*$	$F(3) = 1.32$; $p = .27$	$F(3) = 0.71$; $p = .55$	$F(3) = 1.50$; $p = .21$

main effect ($\beta = -0.16, p < .05$), which was then augmented by the interaction effects. The lowered spectral tilt suggests that the voice quality was constricted when imitating some modulated tokens. For HNR, no effects associated with perceptual choice were found. For SHR, the main effect of choice and its interaction with modulation type were significant. The main effect of choice towards imitated SHR was negative ($\beta = -1.47, p < .001$), meaning that a 'down' response is correlated with a smaller SHR. However, within 'down' responses, all the frequency and/or amplitude modulated tokens led to higher SHR, meaning stronger presence of subharmonics, than unmodulated ones ($\beta s > 1.43, p < .001$). Similarly, for SoE, the main effect of choice and its interaction effects with modulation type were significant. The main effect of choice towards imitated SoE was positive ($\beta = 1.01, p < .01$), meaning that a 'down' response is associated with a higher SoE. However, within 'down' responses, modulated tokens led to a lower SoE, meaning weaker voicing, than unmodulated ones ($\beta s < -1.03, p < .01$).

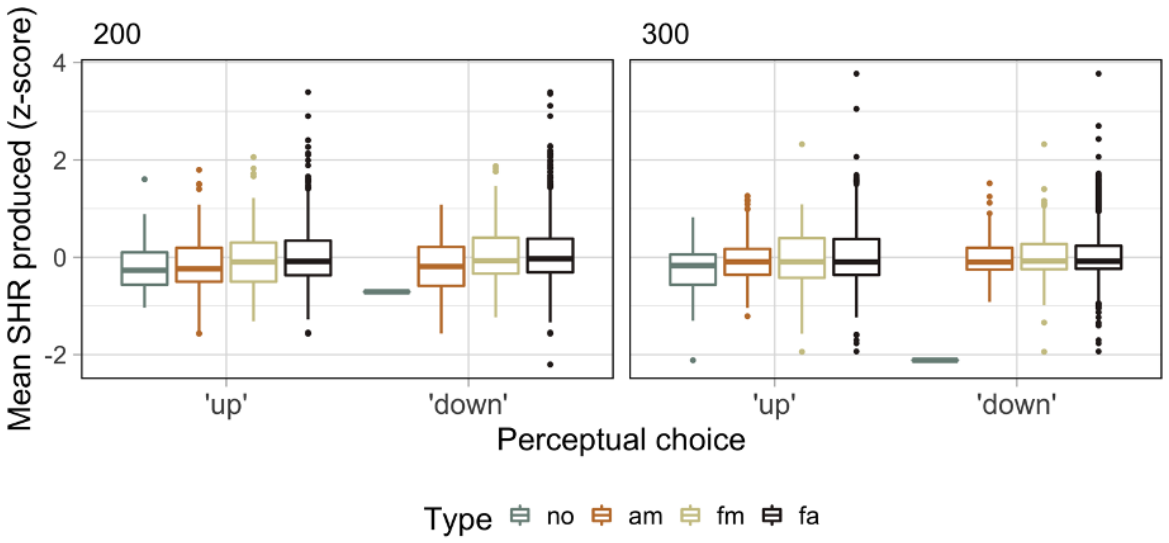


(a)

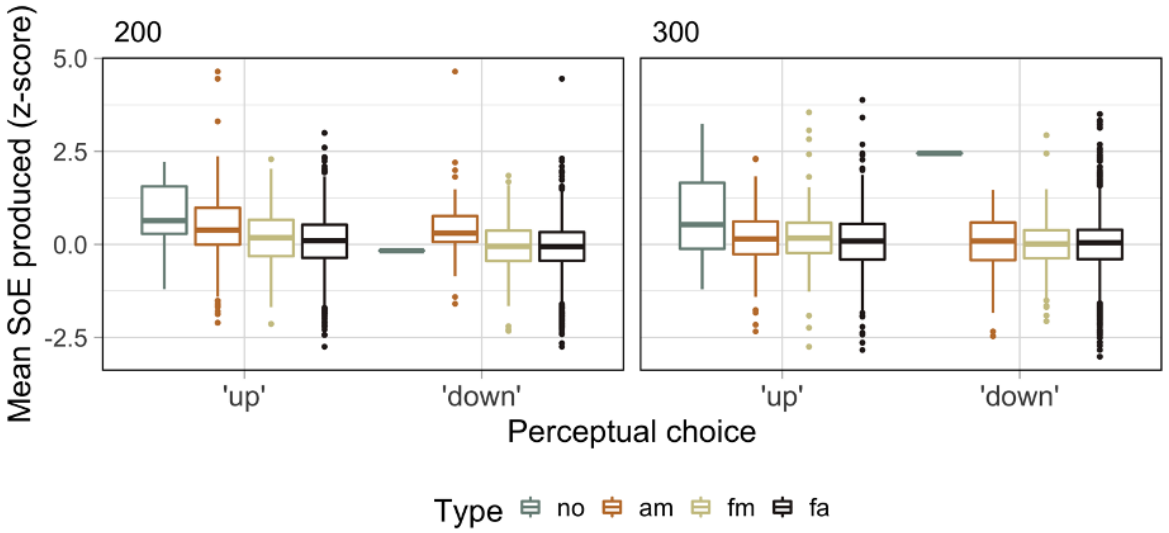


(b)

Figure 4.9: H1*–H2* (a), HNR (b), SHR (c), and SoE (d) in imitation across different f0 conditions in Experiment 2 vary as a function of perceptual responses and modulation type. There were only 49 unmodulated tokens ('no'), and those categorized as 'down' were very few.



(c)



(d)

Figure 4.9: H1*–H2* (a), HNR (b), SHR (c), and SoE (d) in imitation across different f0 conditions in Experiment 2 as a function of perceptual responses and modulation type. There were only 49 unmodulated tokens ('no'), and those categorized as 'down' were very few. (cont.)

Figure 4.9 plots the results of imitated voice quality correlates as a function of perceptual responses discussed above. For H1*–H2*, the values in modulated tokens of 300Hz among 'down' responses were found lower than 'up' responses, though the differences were small based on visualization. A more general lowering trend is observed when comparing modulated tokens to unmodulated tokens. Though choice was not predictive of HNR, differences according to modulation types within 'down' responses of 200Hz tokens and 'up' responses of 300Hz tokens seem to hold. For SHR and SoE, trends of raising or lowering can be seen across different modulation types when compared to unmodulated tokens, especially in tokens of 200Hz and/or 'up' responses. These trends are most borne out with the combined frequency and amplitude modulation – showing that it is mostly likely to be imitated with creakier voice quality. Altogether, these results of voice quality correlates suggest that participants may actively constrict their glottis with weaker (and sometimes noisier) voicing when imitating period-doubled voice. Interestingly, they even tended to increase the amplitudes/levels of subharmonic energy – possibly producing period doubling. Again, there were only 49 unmodulated tokens ('no'), and those categorized as 'down' were very few. The statistical results may be skewed by the few data of unmodulated tokens in 'down' responses to reach significance of an interaction effect. Thus, I advise caution when interpreting the imitation of voice quality, especially with the interaction between modulation type and perceptual choice.

4.3.6. Interim discussion

First, we found that participants produced f0s according to the type of modulation, even with an increasing effect size ranging from amplitude, frequency, to combined modulation. This linear trend was not only observed in mean f0 of their imitations, but also in extreme values (max/min f0). This reaffirms the increasingly stronger effect of modulation types (combined $>$ frequency $>$ amplitude), as reflected by pitch imitation.

Second, as for imitated voice quality correlates, their variances were well accounted for by imitated f0s, and fewer effects came from modulation. Nonetheless, trends were seen in lowering or raising the values of these correlates according to modulation types. For example, the findings that modulation types drove SHR to be higher were expected. Recall that SHR measures the strength of subharmonics relative to harmonics. A higher value of SHR indicates a stronger presence of subharmonics, which is hypothesized to be one of the typical characteristics of period doubling (Keating et al., 2015). Modulation types led to a lower SoE value in realizing pitch and voice of period-doubled tokens, which is also expected if the imitated voice had increased glottal constriction.

Third, perceptual responses were largely found to be predictive of f0 or other acoustic correlates in speakers' imitations. This implies a match between how participants heard the stimuli and how they imitate them in a novel tone language. The range of variation in imitated f0 was not large: <4 semitones in frequency modulation; <2 semitones in amplitude modulation. This likely stems from the fact that the lower pitch perceived from period doubling was too low for speakers to mimic, or speakers made use of an adequate range of semitone differences to signal non-linear perceptual changes according to different modulation types. Still, it is surprising that not only f0 but also voice quality shadowed (with trends) were predictable using their prior perceptual responses. This also shows that listeners/speakers can have a clear preference during perception of period doubling, and that they reproduce the corresponding pitch and voice they could detect from perception. Moreover, the variance across three repetitions (measured by standard deviation) of speakers' imitated f0s were also compared using different modulation types or stimulus f0s, and their results were not significant. This suggests that the range of variation of shadowed f0 was not conditioned by the particular modulation type or stimulus f0, in light of the (un)ambiguous conditions discussed in Section 4.3.4.

4.4. Discussion

4.4.1. Pitch perception and shadowing during period doubling

This study adopted an artificial language learning paradigm to investigate how Mandarin and English listeners perceive and process (pseudo-)linguistic tones realized by period doubling. In the perception experiment, participants were exposed to two novel categories implicitly separated by pitch ranges, and they were forced to categorize period-doubled tones into one of the two categories. In the shadowing experiment, participants were asked to imitate the pitch and voice they heard with the same stimuli in the perception experiment. In both cases, we observe a general trend that the perceived pitch or shadowed f0 is lower when the stimuli are modulated to have correlates of period doubling, and when those modulations increase in degree. Perceived pitch, as well as shadowed f0, shifted down from the original higher fundamental frequency to the f0 corresponding to a lower-frequency subharmonic and/or to the longer period defined by an aggregate of two adjacent pulses. This suggests that listeners tend to identify a lower pitch during period doubling, which was especially found for frequency modulations or when the modulation degree (of both frequency and amplitude) reaches a certain threshold. This threshold differs depending on the modulation type. For example, compared to frequency modulation, amplitude modulation only has around 70% of the tokens identified as low tones even when its modulation degree reaches the extreme (the stronger cycle is 4 times louder than the weaker one). Yet the extreme degree of frequency modulation biases listeners to hear nearly 100% of the tokens as low tones (the longer cycle is 3 times of the shorter one). Listeners are probably more sensitive to changes in period than amplitude of glottal pulses when detecting periodicity of speech signals, or the changes in period and frequency exert a stronger influence on the percept of pitch by the human ear. This may be related to findings that listeners tend to be influenced by changes in the frequency rather than the time domain. For example, temporal noise measures like jitter and

shimmer are not perceptually relevant independently of spectral HNR (Kreiman and Gerratt, 2005; Garellek, 2019).

Apart from the clearer cases where a high or low pitch is identifiable, bitonal percepts are prevalent, which are reflected by the increased variability across subject judgment as well as imitation. When both modulation degrees, especially frequency modulation, are at the upper extreme, pitch is not ambiguous. But the perceived or imitated pitch tends to be ambiguous when the degree of frequency modulation is low, such that the length of the longer cycle does not exceed 1.4 times of the shorter cycle (frequency modulation step ≤ 5 for most ambiguous period-doubled tokens) and when the amplitude modulation not at the extreme (see Tables 4.3 and 4.6). Though the specific conditions that exhibit variability and consistency across perception and imitation do not completely overlap, they largely speak to this generalization. One possibility for the differences between perception and imitation results regarding pitch ambiguity is that perceptual tasks were timed, and participants could only hear each token once in a trial block, whereas during the shadowing task they were allowed and in fact encouraged to play the token as many times as needed before articulating. This could to some extent reflect different timing and activation requirements in the cognitive processes: categorization was forced to happen more quickly, mirroring 'close shadowing' that is less affected by post-perceptual processes except that no production was involved, but the shadowing task could be deemed as 'distant shadowing', that the subjects are fully aware of their output and have taken effort to constrain spontaneous errors (Marslen-Wilson, 1985).

4.4.2. Voice quality imitation during period doubling

Period doubling is often heard as bitonal and rough sounding. It was hypothesized that when roughness accompanies (or even dominates) the tonal percept of period doubling besides pitch, subject will imitate roughness to match the perceived voice quality.

Several voice quality correlates were found tending to vary along with the types of modulation, implying that participants were able to imitate other phonetic cues to period doubling besides f0, involuntarily or deliberately. For example, frequency and combined modulation tended to lower spectral tilt, HNR, and SoE, but raise SHR, indicating that the imitated tones were produced constricted, noisier, quieter, but with more subharmonics, as in the actual tokens heard. These are typical characteristics found for period-doubled voice, as well as other creaky voice subtypes. In other words, speakers not only were able to imitate irregular voicing to match roughness sounding, but more specifically, they could be producing period doubling, or possibly use period doubling to realize roughness in the stimuli. However, we cannot be entirely sure of the precise type of creaky voice produced – whether listeners were able to target the specific subtype of creaky voice. A study designed to test imitation of different creaky voice subtypes would therefore be useful and generalizable.

Upon a closer examination of the imitated production of period-doubled voice, period doubling was indeed frequently observed at the edges of the tokens, connected to the findings in a previous acoustic analysis of period doubling that it may be a form of vocal instability (Chapter 3). Word-initial glottalization induced by the onsetless syllable [a] was also abundant. In an ongoing process, creaky voice is being manually coded by four native Chinese and English speakers including the author. The presence of creaky voice will be mapped to the amplitude and frequency manipulation degree in a given condition. For example, if at least one token out of the three tokens is coded as creaky, we associate the entire condition with potential creakiness. In future work, the perceptual coding of the presence of creak in imitated period-doubled voice will be used to compare with the acoustic correlates discussed here. Individual variation in terms of the extent of creakiness was notable such that some speakers use creak substantially whereas others rarely creak. In addition, besides creaky voice, sometimes nasality and breathy voice were perceived by raters, which calls to the need of a more in-depth and rigorous analysis to capture the imitated voice quality.

One thing to note about the shadowed production of period doubling is that, because vowel quality was possibly affected by the resynthesis in the conditions of frequency and combined modulation, the perceived vowel quality would change from the original vowel [a] to more back-sounding vowels like [ɑ] and [ɤ]. Several participants thus changed their productions to match the vocal tract characteristics, despite being instructed to focus only on aspects like pitch and voice quality. This is not surprising but still interesting, because it raises the question of how well people can imitate aspects of phonation including pitch, phonatory quality, and voicing intensity, etc. while ignoring changes in quality of the segment? In other words, this would involve an additional step to alter pitch and voice quality from a heard segment (e.g., [ɑ]) to another segment that they were instructed on (e.g., [a]) (c.f., Garellek et al., 2013; changes in sources spectral tilt led Hmong listeners to identify a higher vowel [ɔ] rather than [a]). Considering the vowel change as a result of the current resynthesis, an alternative is to use articulatory synthesis to generate the stimuli based on a voice source model (e.g, Liljencrants–Fant; Fant et al., 1985) with varying glottal modulation extent to manipulate the frequency and/or amplitude changes in adjacent cycles, and then apply vocal tract filtering to create a speech envelope. However, this approach assumes a less direct link from the glottal modulation degrees of a voice source model to human perception rather than from acoustic waveforms, which bypasses the intermediate step and so can make the interpretation challenging. In fact, both Bergan and Titze (2001) and Sun and Xu (2002) adopted the articulatory synthesis approach. Sun and Xu (2002) also converted their glottal modulation degrees to signal modulation degree for comparison and failed to do so for one of the conditions, partly due to the still opaque relationship in source-filter interaction. Nevertheless, a question for future studies is, how to predict the changes in perceived resonance as a function of changes in the source.

Lastly, the question of whether the shadowing of period-doubled voice was conscious or involuntary has implications for phonetic accommodation theories beyond pure imitation of period doubling. While higher and abstract structures such as phonological features are generally agreed to be encoded prior to the speech planning process, and even though studies have found that creaky

voice is often used to signal sociolinguistic meanings (Henton, 1989; Dallaston and Docherty, 2020; Podesva and Callier, 2015), researchers seldom discuss the stages in speech planning at which the voice enters, alters, and converges to meet the needs of daily communication.

4.4.3. Implications for use of period doubling in tone perception and production

An artificial language learning paradigm, rather than ABX discrimination or pitch matching (e.g., Bergan and Titze, 2001; Sun and Xu, 2002), was adopted in the current study. One of the reasons was to be able to generalize the pitch perception results beyond psychoacoustic findings, such that they may be interpretable with respect to perception and production of linguistic tone. In this setting, listeners are perceiving and imitating linguistic tones that are manifested in period-doubled voice. They were able to extract crucial information of pitch and voice in these tones to categorize them into high or low tones.

Based on the findings that period doubling leads to a low tone bias – regardless of the original f0 – we would predict that the presence of period doubling could be used to signal low tones in languages, even when the f0 of the original tone is high. It will also interfere with high-tone perception, at least with moderate to high modulation. We observed interesting interactions between the two f0 levels and modulation types. When f0 was at $200Hz$, amplitude modulation did not yet lead to a significant increase in the proportion of low tones heard ($< 25\%$) or decrease in pitch imitated. But when f0 was higher at $300Hz$, amplitude modulation led to a significantly larger proportion of low tones categorized $(40 \sim 50\%)$, and the imitated pitch was comparable to that of frequency and combined modulation.

Thus voicing with amplitude modulation can still signal a 'high' tone, even when the modulation is strong, and especially when the original f0 is *lower*. Both perception and shadowing results conform to this, though it seems counterintuitive. However, pitch and tone perception is relative, it is possible that if the original unmodulated tokens with an already low f0 ($200Hz$)

are categorized as the 'high' tone baseline for further comparison, the effect on the f0 induced by amplitude modulation would not be salient enough to signal pitch lowering for a 'low' tone category. Relatedly, Davidson (2020) found that listeners rate utterances with multiple pulsing as lower in pitch than modal utterances, but not for speakers with lower pitch. Additionally, the imitated f0 was not much lowered in all types of modulation when it reached below $200Hz$ for both original f0s. For $300Hz$, the difference between the imitated f0s from unmodulated to amplitude-modulated is more salient. For both stimulus f0s, the changes in f0s induced by frequency or combined modulation result in a more stabilized percept, though trends of potentially different frequency effects depending on the particular stimulus f0 can be observed in Mandarin listeners. However, the underlying reason behind the differences in pitch perception and imitation between amplitude and frequency modulations remains a puzzle. It appears that when changes in pitch are not driven by frequency modulation, listeners might ignore changes induced by amplitude modulation, in particular at lower frequencies.

Regarding the language and music effect, both English and Mandarin listeners, as well as musicians and nonmusicians, behaved similarly. This suggests that pitch perception during period doubling may not be language-specific and is likely not influenced by inherent knowledge of lexical tone or musical entrenchment. While Bergan and Titze (2001) found that vocalists rated pitch more accurately and consistently than non-vocalists, I suspect that the discrepancy between their and the present study may lie in the different experimental designs. They used pitch matching and discrimination tasks, but the present study used categorization in an artificial language setting which may be more complex than pure pitch perception. Finer-grained differences among pitch are probably ignored as long as period-doubled tokens can be classified into high versus low tone categories. Thus, the differences among music experiences are less relevant here. In the shadowing study, it may be worth sub-categorizing vocalist and non-vocalists among the musicians to probe any differences in vocal imitation; but still, if period-doubled tones were treated as linguistic categories, vocalists should not be expected to perform better than non-vocalists in terms of expressing linguistic meaning apart from vocal skills.

4.4.4. How does production align with perception during period doubling?

One of the central questions asked by using a shadowing study is whether we can establish a link between perception and production. In the current study, the perceptual choice significantly predicts the changes in pitch imitation: a token categorized as (chosen to be) 'high' corresponds to a higher shadowed f0; conversely, a 'low'-tone choice corresponds to a lower shadowed f0.

Besides pitch, voice quality correlates show an interesting connection between voice production and pitch perception of period doubling, as well as with presumed voice percept of period doubling besides pitch. Recall that spectral tilt, subharmonics, and voicing strength are not strongly affected by modulation when perceptual choice is not considered in the model. When perceptual factors are considered, H1*–H2* becomes lower in participants' productions as predicted by a low-tone choice in 300Hz tokens, implying a more constricted quality. SHR also becomes higher and SoE lower with low-tone responses in modulated tokens, implying 'multiply pulsed' with stronger subharmonics along the original harmonics and weaker voicing possibly due to stronger adduction. Though we cannot overgeneralize a holistic imitated voice to match period doubling from individual voice quality correlates, a trend exists in which the voice imitated in production maps onto the voice heard, as manifested by the tone categorization choice in perception.

4.5. Chapter summary

This chapter presents one study involving two experiments on perception and shadowing of period doubling. In the first experiment, I used an artificial language learning paradigm to investigate how people categorize tones of period-doubled voice in a novel language. The results demonstrated that the proportion of categorized 'low' tones increased with the presence of period-doubled modulations, as well as with increasing degrees of amplitude, frequency, and concurrent amplitude

and frequency modulations. Among them, frequency modulation had a stronger effect than amplitude modulation, and the concurrent modulation showed an additive effect over the frequency modulation alone. Interaction effects of amplitude modulation was found on different stimulus f0s such that tokens of 300Hz were perceived as 'low' tone more frequently than those of 200Hz. In the second experiment, I used a shadowing design to probe how people imitate period-doubled tones. The results indicated that listeners did adjust their f0s according to the modulation degrees; again, frequency modulation exerted greater influence on lowering one's imitated f0. Voice quality correlates were correlated with the presence of modulation, suggesting that participants were also shadowing the voice quality of stimuli. Importantly, the produced f0 and voice characteristics can be predicted by their previous perceptual choice of a stimulus' being representative of a low or high tone. This validates a perception-production link specifically for period-doubled voice. In addition, both studies show that the pitch of period doubling is ambiguous when frequency modulation is lighter and unambiguous when frequency modulation is at its extremes, and that amplitude modulation generally does not have a strong effect on perceived pitch. These findings contribute towards understanding of the role of creaky voice in lexical tone perception and production. It is predicted that languages that use period doubling to manifest linguistic tones typically result in a low tone percept for its users, and that speakers can produce such voice quality to convey meanings that are associated with low tones.

Chapter 5

General discussion

This dissertation investigates the defining features of period doubling, a subtype of creaky voice consisting of multiple pitches and used sub-phonemically in Mandarin tones, from perspectives of articulation, acoustics, and perception in three studies. By probing the phonetic details of period doubling, this work enriches the framework of creaky voice subtypes and explores the use of period doubling in language, especially how it contributes to tone production and perception. I find that, the alternation in pitch and voice quality in articulation and a lower spectral tilt differentiate period doubling from vocal fry and modal voice. These unique features then result in differential perception effects on pitch, voice quality, and tone, compared to modal voice and other irregular voice qualities.

This furthers our understanding of non-contrastive phonetic events, which may be reinterpreted as a part of the grammar instead of pure physiological phenomena, and especially contributes to linguistic meanings of tone and phonation categories. Investigating non-contrastive speech variation also provides insight into multidimensionality in sound categories and contributes towards understanding linguistic roles of seemingly redundant phonetic details. Moreover, because of the frequently-studied nature of period doubling in voice and singing literature, this dissertation lays a foundation for further comparisons of this voice between typical and pathological voice qualities, and similar or related voicing patterns in speech and singing. The following discussion centers around answering the overarching research questions raised in the introduction of this work.

5.1. What are the articulatory aspects of period doubling found in typical, non-disordered, speech?

Using electroglottography (EGG), Chapter 2 analyzed period doubling as found in Mandarin both synthetically, based on the meta-cycle that consists of a pair of alternating cycles, and analytically, based on the respective distributions of the alternating glottal cycles. On the one hand, a low 'fundamental' frequency that is half of the original one could be derived from the meta-cycle whose period usually doubles the original period. More variably-distributed glottal constriction measures (contact and speech quotients) were found for the meta-cycle in period doubling than vocal fry and modal voice due to the two 'sub-pulses' within the meta-cycle. The average degree of glottal constriction in period doubling is comparable to modal voice and lower than vocal fry. Thus, the articulation of this voice appears to be closer to modal voice with more balanced open and contact phases than vocal fry and other types of creaky voice that involve substantial glottal constriction. On the other hand, the meta-cycle in period doubling consists of two alternating pulses that have distinct pitches as well as voice qualities. That is, period doubling is articulated through pulses that not only alternate between amplitudes and/or periods (which was found in previous studies, e.g., Kreiman et al., 1993) but also often differ in pulse shapes indicating differences in voice quality. This characteristic is not found for vocal fry or modal voice.

Based on the empirical data, the differences in frequency and amplitude across the alternating strong and weak glottal cycles exhibit a ratio of 3:2 and 2:1, meaning a larger difference in amplitude than in frequency. The frequency ratio is consistent with a subharmonic vibratory pattern documented in Švec et al. (1996). Further, the differences in pulse shapes show alternating contact durations, pulse shape symmetry, and speed of vocal fold contact, suggesting different modes of articulation: more vs. less constriction, more vs. less balanced open and contact phases, and faster vs. slower contacting motion.

The articulatory findings have implications for understanding the range of possible voicing patterns produced by human vocal folds, including both irregular and quasi-periodic vibrations (such as the period-doubled pattern), to be incorporated in phonation models, for example, Hanson et al.'s (2001) general model of both modal and non-modal phonation, and Zhang's (2018) model of irregular voicing that account for vocal instability. Section 5.3 will further the discussion around vocal instability, a possible mechanism that drives the formation of period doubling.

The results also contribute to devising methods of synthesizing period doubling. It calls to the importance and a need of incorporating alternations in pulse shapes and voice quality in addition to those in amplitude and frequency. For example, as used in Huang (2020), the 'double pulsing' parameter in Klatt synthesizer only manipulates concurrent modulations in frequencies and amplitude, which may not be accurate. To mirror the articulatory patterns in period doubling, a better synthesis would need to take into account of the alternation in voice quality; though, this would pose challenges to synthesize acoustic measures such as spectral tilt (see Section 5.2).

What is special about the alternation in voice quality in addition to pitch is that it may contribute to the indeterminate percept in voice quality as well as pitch. Period doubling has been characterized by having an indeterminate pitch and rough quality, which was assumed to result (only) from multiple frequencies (Redi and Shattuck-Hufnagel, 2001; Yu, 2010; Keating et al., 2015). However, the present findings suggest that the indeterminate or concurrent percept can also be a result of the inherent alternating voice qualities, on top of distinct pitches. This has implications for the use of period doubling in languages with tone and phonation.

For instance, alternating more and less constricted qualities suggests a covariation between creakiness and breathiness. A voicing such as period doubling can thus allow the co-existence of two different qualities and may even interfere with the 'non-target' breathiness (as noted in Swedish by Hedelin and Huber, 1990). Such covariation is also seen in tone languages such as White Hmong (Keating et al., 2010), !Xóõ (Garellek, 2020), and Northern Vietnamese (Brunelle et al., 2010). It is possible that period doubling acts as a carrier to switch between qualities that

are consistent with creakier or breathier voice. This nevertheless raises, rather than answers, more questions about the interaction between voice quality and pitch, for example: how do listeners perceive and resolve the different interactions between multiple voice qualities and pitches, and why is roughness often perceived from indeterminate quality and pitch?

Further, the fact that period doubling does not exhibit degrees of constriction comparable to expectations of most types of creaky voice (except for non-constricted creak) calls for improvement to the taxonomy of creaky voice. This issue will be expanded and discussed in Section 5.5.

Apart from the EGG waveforms in the time domain, Chapter 2 also looked into spectral characteristics of period doubling. Spectra of the EGG signal can be a meaningful addition to those of acoustic signals for studying harmonics and subharmonics in the voice source. Indeed, in the frequency domain, the stronger presence of subharmonics is attested in period doubling compared to vocal fry and modal voice. Corresponding to the different kinds of period doubling in the waveform, various spectral patterns were found depending on the relative strength of the subharmonics to harmonics: stronger H1 (originated from subharmonics) or H2 (originated from original harmonics), alternation in the magnitude of harmonics and subharmonics, amplitude dips in certain frequency ranges, and groupings of frequencies. The different temporal and spectral profiles would lead to timbre differences within period doubling that contribute to pitch perception (de Cheveigné, 2005) and tone perception (Gandour, 1978). One line of future work should be devoted to establishing the relationship between the frequency/amplitude ratios in the waveform and the presence and strength of subharmonics in the spectrum, and thus determining different kinds of period doubling based on voice source spectra.

Last but not least, the articulation analysis answers part of the 'what' question of period doubling, but not necessarily the question of 'how is period doubling produced', because the product of period-doubled phonation is rather taken as granted for the current purpose of the study. The analysis of the acoustic output of EGG waveforms does not explain what physiological con-

figurations and vibratory patterns lead to period doubling. To clarify the underlying mechanisms of vibration during period doubling, air flow and voice source modeling studies could be useful by investigating factors such as subglottal pressure, aspiration, ventricular constriction, and various conditions that may favor period doubling. In addition, future studies using high-speed imaging techniques, for example, trans-nasal laryngoscopy, can help pin down the exact nature of phonatory activities underneath the seemingly intricate period-doubled pattern and capture individual and population variability in production. These will also enable a comparison of period doubling to other related voicings in pathological voice with vocal and/or ventricular folds asymmetry (e.g., diplophonia: Bae et al., 2019) and certain singing styles that involve coupling of vocal and ventricular folds vibration (e.g., Mongolian throat singing: Lindestad et al., 2001; Sardinian: Bailly et al., 2010; Tibetan chant: Fuks et al., 1998).

5.2. What are the acoustic characteristics of period doubling that distinguish it from other voicing types?

Using statistical models and machine learning methods, Chapter 3 reports the most important acoustic (and articulatory) features in differentiating period doubling from modal voice and vocal fry. Period doubling is readily acoustically identifiable using spectral tilt measures, especially H1*–H2* (* corrected for formants), which shows a three-way separation: period doubling < vocal fry < modal voice. Period doubling also has a mid H1*, and higher H2* and H2*–H4* compared to vocal fry and modal voice. SHR, however, did not emerge as a distinctive measure in statistical models except that both vocal fry and period doubling were found to have higher values than modal voice, as expected for creaky voice. Most of the statistical significance was established only for women as 73% of the tokens used in model comparisons were from the women subset.

Random forest models confirmed the importance of f0, spectral tilt, SoE, and HNR in classifying the three types of voicing; adding articulatory features from EGG helped improve the model performance, such that the duration between the ends of the opening phase and various contact quotient (CQ) measures contribute to the classification, with H1*–H2* remaining the most robust measure. The results of the logistic regression model with added articulatory features show that vocal fry can be characterized as having higher contact quotients using threshold or hybrid method, whereas period doubling is favored by quotient methods that incorporated derivatives of the EGG. It is also likely that CQ measurements during period doubling (and maybe other irregular voicing types) could benefit from tailored algorithms so that the specific characteristics of a particular voicing can be dealt with more straightforwardly (e.g., alternation in frequency, amplitude, and pulse shapes in period-doubled cycles). In addition, the comparison between using only the acoustic features and a combination of acoustic and articulatory features furthers our understanding of how these two aspects of production weigh in signaling a meaningful difference in voicing types. As a cautionary note, the EGG signal is only an approximation of the vibration of vocal folds, less direct than glottal airflow or area obtained through laryngoscopy or other invasive imaging techniques. Nonetheless, using machine learning approaches, Chapter 3 shows that the glottal constriction measures derived from EGG waveforms complement the acoustic findings and well distinguish period doubling from vocal fry and modal voice.

When comparing to the articulatory findings in Chapter 2, there appears to be a discrepancy regarding constriction as reflected by the different results of contact quotients and spectral tilt measures. Both CQ and H1*–H2* were found often correlated with glottal constriction (Childers et al., 1990; Holmberg et al., 1995a). The average CQ during period doubling was lower than vocal fry and similar to modal voice with a wider range of variation, indicating a lower degree of constriction overall (though, constricted 'vocal fry-like' glottal cycles are often seen in one of alternating smaller cycles). However, the lower H1*–H2* found in period doubling than vocal fry would suggest a higher degree of constriction. Not unexpectedly, I also did not find any strong correlations between CQ and H1–H2, similar to some other studies where only moderate

correlations exist (DiCanio, 2009; Kreiman et al., 2012). However, a lower CQ and a lower H1*–H2* are not necessarily incompatible when we closely examine the patterns of period doubling. During period doubling, the presumable fundamental frequency that is one octave lower than the original f0 is originated from the subharmonic, and the original f0 becomes the second harmonic. The original f0 remains stronger on the spectrum, leading to a stronger H2 which induces a lower H1–H2. The fact that period doubling does not have a lower H1 than vocal fry is additional evidence that, in period doubling, a lower H1–H2 (as a form of spectral tilt in the lower frequency region) may not be caused by increased glottal constriction. Thus, this seemingly contradiction speaks to the same characteristic that defines period doubling.

Now considering the alternation found in glottal constriction measures, which will not be realistic for acoustic measures such as H1–H2. The calculation of CQ is done on a cycle-by-cycle basis, whereas the spectral measures depend on f0, whose estimate needs to be calculated over multiple periods. Thus, the alternating voice quality cannot be replicated in the spectral domain of the acoustic signal, because measures that depend on a window spanning multiple periods would effectively wash out the alternation. If we were to incorporate this voice quality alternation in speech synthesis, then perhaps articulatory synthesis rather than acoustic synthesis is more feasible.

One line of future work I propose is motivated by the concurrent changes or mismatch between voice source (manifested by EGG) and vocal filter. I ask, how do the glottal changes reflected in the EGG waveform connect to changes in the audio output? Machine learning approaches can be used to evaluate the relationship between the dynamics of the source and the filter, by extracting features of EGG pulses that form a selection criterion of period doubling to be tested on the corresponding audio signals. This is to study how well the EGG source can predict the occurrence of period doubling from the audio output. The direction of the prediction can also be reversed – using the audio output to predict the voice source of period doubling. For example, the (f0-related) formant changes in the speech signal brought about by different vocalic segments

and absent in the EGG signal (and source) can also affect the lower harmonics such as H1 and H2.

These findings in Chapter 3 help inform research in more technical and applied fields such as automatic speech recognition and speech processing by taking into consideration of the various acoustic as well as articulatory features. In particular, refining the features of different types of non-modal voicing can help improve the effectiveness of pitch and/or creaky voice detectors and naturalness of voice synthesizers. For example, a recently developed creaky voice detector by Drugman et al. (2014) used H2–H1 and f0 during creak as acoustic features for creaky voice characterization, though without reference to the specific types of creak. They have drawn insights from categories of irregular voicing in Redi and Shattuck-Hufnagel 2001, which included documentation of both period doubling and vocal fry. Based on the present results, other spectral measures such as H2*–H4* and energy and noise measures could be useful additions to the existing algorithms.

5.3. Where does period doubling occur linguistically? Specifically, does it occur in specific tonal and phrasal environments?

To understand why period doubling is formed and used frequently, I presented *post-hoc* analysis of prosodic distributions of period doubling and vocal fry in Chapter 3. Period doubling is more prevalent (occurred more than 3 times often) than vocal fry anywhere in the utterance and more frequently found in women than men ($\sim 73\%$ of the period-doubled utterances). Why is period doubling so prevalent? Yu (2010) argued that vocal fry is contingent on low f0, but period doubling is not contingent on any particular f0. Perhaps because of the non-dependence on low f0, period doubling can occur across all tones without restrictions and is preferred by female voice. Still, sentence-medial tones with a lower pitch (Mandarin Tones 2, 3) than higher pitch (Mandarin Tones 1, 4) are more frequently associated with period doubling. Moreover, period

doubling and vocal fry have different phrasal distributions. Period doubling is more likely to occur at utterance edges (utterance-initial and utterance-final positions), and its occurrences gradually increase towards the end of an utterance. For example, in the Mandarin corpus, scripted utterances always end in a final high tone (which would be an unfavorable position for creak), yet the final segment was often realized as period-doubled. In contrast, vocal fry is rarely found utterance-initially and mostly found in the post-focal position immediately after the stimulus, likely driven by the post-focal compression of f0 ranges (Xu, 2011).

The present findings that period doubling occurs most often at utterance edges speak to forms of vocal instability induced by the beginning and end of phonation. Consequently, the linear increasing trend of occurrences of period doubling towards utterance-final positions could be a byproduct of downdrift (declination) as f0 progressively lowers towards the utterance edge without necessary constriction. In contrast, vocal fry is typically triggered by a low and compressed f0 range with more probable constriction. The fact that the occurrences of vocal fry are more restricted and mostly occur in the penultimate position suggests that vocal fry could signal a stronger linguistic role such as marking a weak prosodic element post-focally.

What is the linguistic role of period doubling? If it is a non-targeted byproduct of declination and a form of unstable voicing where the subglottal pressure is low, period doubling may be realized involuntarily due to articulatory and/or respiratory factors. If it is controlled for realizing creak and/or low pitch, we may expect period doubling to have a different linguistic function from vocal fry. The current results do not reveal the answer of whether period doubling is the target of any speech, but suggest that this voicing might be tied to utterance-level phenomena, not tonal ones. If confirmatory studies were to follow up on this question, we would hypothesize that to the extent that period doubling is controlled in speech, it may be driven by phrasal factors, rather than constriction or pitch alone.

Interestingly, Slifka (2006) found a type of non-constricted creak at the end of an utterance, which is usually produced with short intervals of adduction followed by longer intervals

of abduction and/or with incomplete closure of the vocal folds, unlike vocal fry. Period doubling thus may be comparable to this similar phonation-ending gesture driven by variations in the subglottal pressure as utterances unfold. The trend of f0 declination that accompanies the utterance-level articulatory gestures (Yuan and Liberman, 2014) can also be attested by different kinds of period doubling seen in different phrasal positions. For instance, at the end of the utterance, tokens of period doubling with both frequency modulation and fry-like pulses are more likely to occur than tokens with amplitude modulation. At the beginning of a tone or an utterance, tokens with amplitude modulation often occur but the stronger and weaker pulses do not have considerable differences in amplitude or shape. The findings of Chapter 4 provide further support: utterance-initial f0 is higher (corresponding to a more frequent percept of high-tone during amplitude modulation) than utterance-final f0 (corresponding to a more frequent low-tone during frequency modulation).

What might be the sociolinguistic functions of period doubling and vocal fry? Both as subtypes of creak, they are often used to establish aspects of identities, speaker groups, and convey various social purposes (see a review in Davidson 2021). Are they merely two different voicing marking the same sub-phonemic voice quality and thus bear the same linguistic meaning? The results suggest they appear to have different roles in the linguistic signal because of their distributions. It is less likely that either voicing is redundant; however, it does not rule out the possibility that speakers can trade between period doubling and vocal fry to realize creak for uses that tie to tone, prosody, or pragmatics. Given the asymmetric gender distributions, women might be generally using period doubling more than men because producing period doubling would be easier than vocal fry which requires maintaining articulating low f0 with a constricted glottis. For example, women might be using period doubling as a less effortful means to access creaky voice and lower pitch which were found to be associated with socially desired impressions (Davidson, 2021; Borkowska and Pawlowski, 2011; Yuasa, 2010). Follow-up studies need to test under what communicative contexts and linguistic environments period doubling or vocal fry may be preferably used by speakers.

One implication for theories and models of phonation types gathered from the findings of the distributional properties of period doubling is that the demarcation between modal and non-modal phonation may be more blurred than previously proposed (e.g., by Gordon, 2001). The examples of period doubling found in the corpus demonstrate the "effortless" transition from regular to irregular voicing, especially when approaching the end of the utterance. Thus, I recommend the following empirical considerations for speech production studies. Data elicitation of both segmental and suprasegmental units should be conducted in a carrier phrase rather than in isolation to avoid utterance edges which will lead to conflation of modal and non-modal phonation. For example, the realization of period doubling in Mandarin Tone 3 or other tones utterance-finally will result in similar kinds of articulation and voicing instability, thus an obscured underlying voicing differences in these lexical tones. When trying to study aspects of phonation other than period doubling and factors other than phrasal positions, these prosodic environments should be avoided. Comparison of phonation types should generally be done in such a way that controls for positions in utterance.

Two lines of future work merit attention. First, based on the distributions of period doubling in certain tones and phrasal positions, future studies may test the role of period doubling in perceiving tone and phrasing. The questions to be asked are, which tones are heard more often with period doubling? Can period doubling signal phrase-finality? One ongoing project is to test if and how period doubling and vocal fry signal tones with a low pitch. Following the visual sort-and-rate paradigm proposed by Granqvist (2003), I am currently conducting a study on Mandarin tone perception to test how listeners sort and rate tones according to their naturalness and typicality, using naturalistic tokens with modal voice, vocal fry, or different types of period doubling. The expectation is that, if period doubling or vocal fry is predictable for certain low tones, these low tones associated with kinds of creaky voice would be rated more natural than their modal counterparts. Alternatively, if the indeterminate percept of period doubling is interfering, the period-doubled tones may be sorted towards the least representative tones.

Moreover, research has found that creaky voice biases listeners to disambiguate phrase versus compound in ambiguous prosodic structures (Crowhurst, 2018), or cues sentence boundaries in natural conversations (Kreiman, 1982). A follow-up experiment can therefore use resynthesized period doubling especially at the end of a prosodic phrase to test whether it signals phrase finality and affect perceptual grouping of phrasal components.

Second, in addition to the *post-hoc* analysis of the occurrences of period doubling in the Mandarin scripted corpus, it is worth investigating those in spontaneous speech, and in non-tonal languages and populations. In future studies, for example, I can use creaky voice detector in other corpora, such as the English CMU Arctic corpus, which has both EGG and acoustic data.

5.4. How do listeners perceive pitch, voice quality, and tone during period doubling?

Using an artificial language paradigm, I tested perception and imitation of linguistic tones with resynthesized period doubling in Chapter 4. The perceptual results indicate that both Mandarin and English listeners regardless of music background hear the pitch of period-doubled tones based on the strength of frequency modulation more than amplitude modulation. In general, a low-tone bias emerges among both groups of listeners. When frequency modulation is at the extreme degree, pitch is not ambiguous and heard as lower. When frequency modulation is weak and amplitude modulation is at most moderate, pitch is often heard as ambiguous between high and low tones. Amplitude modulation has a weaker effect on lowering the pitch perception; for example, the proportion of a high-tone percept is largely equal to a low-tone percept, even when the stronger cycle is four times louder than the weak cycle. In the shadowing study, listeners are able to imitate the period-doubled tones not only by adjusting f0, but by also modulating their voice quality. Again, frequency modulation leads to a larger pitch range produced across almost four semitones, whereas amplitude modulation only induces a pitch range of fewer than two semitones.

Besides pitch changes, lowering or raising trends of voice quality correlates in the direction of creakiness such as spectral tilt, noise, energy, and subharmonics measures accompany the types of modulation, especially when considering perceptual response. In particular, the trends of imitated creaky voice in the form of period doubling are shown in the combined frequency and amplitude modulation. In addition, by correlating the perceptual categorization results to the imitated f0 and acoustic correlates in the imitation, I show that there is a link between perception and production during period doubling.

One implication for pitch perception models lies in the differential effects between frequency and amplitude modulation. How do we define a cycle to form the temporal basis of pitch perception? Modifying levels of period induces a stronger percept of the lower frequency component than modifying levels of amplitude. This may be related to findings that listeners tend to be influenced by changes in the frequency rather than time domains. For example, temporal noise measures like jitter and shimmer are not perceptually relevant independently of spectral HNR (Kreiman and Gerratt, 2005; Garellek, 2019). Though Bergan and Titze (2001) and Sun and Xu (2002) have already found similar stronger effects of frequency modulation, the interaction effects between these types of modulation and stimulus f0 are novel: a higher f0 ($300Hz$) with amplitude modulation can lead to a lower percept than a lower f0 ($200Hz$); a lower f0 ($200Hz$) with frequency modulation induces larger changes in imitated f0 than a higher f0 ($300Hz$). Future work is welcome to replicate and expand the current findings crossing different f0s with the modulation types. When it comes to tone perception, the current results suggest that speakers could use period doubling to signal low tones in a tone language when needed, even when in the range of an original high tone; or, use it to expand the range of an original low tone. Drawing from the results of distribution in Section 5.2, it is predicted that the presence of period doubling especially found to be more frequent at utterance edges will interfere with high-tone perception, at least with moderate-high modulation.

Following up on the perceptual results, one puzzle to be entertained is how we may interpret the bitonal percept of the ambiguous pitch in period-doubled tokens especially with low-to-moderate modulation degrees. Investigating contextual variation of such voice will help clarify the properties of certain pitch ambiguity. In future work, I can test the role of these ambiguous pitches and how they may be settled by embedding period-doubled tokens in different acoustic contexts. Chambers et al. (2017) has shown that prior acoustic context almost fully determines the perceived direction of frequency in 'Shepard tones' (pure tones with octave relationships) which contain an equal likelihood of upward or downward pitch shift. For example, if we were to employ a similar experiment re Chambers et al. by placing period-doubled tones in different surrounding pitch contexts, several possibilities will be borne out. Listeners may perceive a discontinuity in pitch due to the rough quality of period doubling; they may perceive a pitch drop after a high probe tone due to the prevalent low-tone bias; or they could perceive a fully smooth tone determined by the prior pitch context, resembling a Shepard tone.

In Chapter 4, period doubling was manipulated in the temporal domain by creating alternating cycles that differ in amplitude and/or frequency. One unanswered question in this dissertation concerns the relationship between the indeterminate pitch percepts and the two sets of harmonics and subharmonics in the spectral domain. Alternative synthesis that manipulates the relative strength of harmonics (e.g., H1 and H2, the spectral tilt) in the source spectrum of period doubling, combined with inverse filtering would be useful in future work. In addition, alternating cycles during period doubling often differ in their pulse shapes to reflect different voice qualities (Chapter 2; Section 5.1). The resynthesized stimuli based on manipulations on the acoustic output do not include such variation. Thus, studies could make use of articulatory models based on source-filter theories to directly manipulate various phonatory aspects that mirror the patterns of period doubling. This way, we can better capture its acoustics and articulatory characteristics. An implication for phonation perception follow from this alternative approach will be that the period doubling as a subtype of creaky voice may interfere with other voice qualities such

as breathiness because the alternating constricted and non-constricted qualities in glottal pulses implies covariation between harshness and breathiness.

Related to the discussion in Section 5.3 as to whether the production of period doubling is a byproduct of vocal instability out of phrasing considerations, the shadowing of such voice may also be involuntary or conscious. If imitated creaky voice were produced by a conscious effort, implications for phonetic accommodation theories exist beyond pure imitation of period doubling. If speakers intentionally imitate or produce period-doubled voice, when during speech planning does the imitation occur? Is it part of the speech planning process? It is generally assumed that discrete units such as phonological features are encoded prior to the speech planning stage and continuous phonetic gestures are calculated after the planning (Keyser and Stevens, 2006). The discussion on speech planning process in past literature has been focused on segmental and prosodic structures like gestures, syllables, and pauses (Krivokapić, 2014; Cholin and Levelt, 2009; Tilsen et al., 2013). It is an open question as to whether voice quality is imitated like other shorter-term linguistic units like gestures (or segments, or higher structures) or as part of broad (long-term) voice quality possibly based on communication needs, with considerable sociolinguistic bearings (Henton, 1989; Dallaston and Docherty, 2020; Podesva and Callier, 2015); we should clarify at which stage voice quality enters into speech production.

5.5. Implications for voice and linguistic theories and beyond

Taken together, how does period doubling inform theories and frameworks of voice and phonation? This dissertation starts with an assumption that period doubling is defined based on the existing framework of creaky voice subtypes proposed by Keating et al. (2015). However, after a close examination of the defining characteristics of period doubling, the properties in articulation and acoustics do not fit in neatly with the taxonomy according to the expectations of canonical creaky voice. For example, each of the representative acoustic attributes: constricted voicing, irregular f0, and low f0, would be sufficient for generating a creaky percept (Garellek, 2019).

Chapter 2 discussed how period doubling might not be consistent with either attribute for the following reasons: no considerable degrees of glottal constriction, no irregularity in the voicing periods besides alternation, and not necessarily low f0. It is certainly different from prototypical creaky voice, also unlike vocal fry, aperiodic voicing, and tense voice. Even if period doubling still belongs to the 'multiply-pulsed' voice, the low spectral tilt does not represent glottal constriction, but is rather caused by a more prominent second harmonic from the original higher f0. The closest creak subtype may be the non-constricted creak; however, low f0 cannot be stipulated directly from period doubling because of the presence of multiple frequencies. A percept of low pitch and rough quality as shown by the perception study appears to be the one and only dimension that would place period doubling together with the other subtypes of creak.

The results suggest that period doubling is intimately linked to low pitch (and possibly constriction), with its alternation in pitches and voice qualities, rather than inherently low f0 or increased constriction found in other creaky voices. Does *low pitch* work for other creak subtypes? Subtypes with low f0 should all be correlated with a low pitch perception (e.g., vocal fry: Kuang and Liberman, 2016), except for tense voice which is only recently found can have non-high f0 (Dawson et al., 2022). Aperiodic voice is mainly captured by noise without a perceptible pitch and would agree on the roughness dimension with period doubling. If we were to add a new dimension, *low pitch*, to the taxonomy, that will likely solve the problem raised by the dissertation. The only consideration is that the existing framework only makes use of acoustic properties whereas pitch is a perceptual entity. In that regard, acoustic and perceptual dimensions may need to be combined to capture voice quality as a psychoacoustic phenomenon.

The results also blur the distinction between previously strictly-defined modal vs. non-modal voice qualities along a continuum of glottal opening degrees such that creaky voice is expected to stay the most constricted and breathy voice has to be most open (e.g., Gordon, 2001; Laver, 1980; Esling et al., 2019). The presence of alternation between two modes in period doubling, the fact that this voicing shares acoustic characteristics of both vocal fry and modal

voice, and the mechanism of unstable voicing reflected by its occurrences towards utterance-final positions are all evidence showing obscurity of the acoustic differences between modal and non-modal phonation. We should revisit the question, what is creaky voice? The answer may not be how we measure its acoustic properties but rather, how we perceive the voice, and how we measure our percept of creakiness. It is thus proposed that *low pitch* may be the most important dimension of creaky voice, though more work needs to be conducted to figure out the psychoacoustic relationship between low pitch and roughness that essentially describes what we perceive as 'creak'.

The perceptual qualities of the human voice are of practical interest to speech and voice clinicians and singing scientists. For example, 'diplophonia' is a well-known marker of speech disorders, and the degree of roughness/harshness has been a critical scale in evaluating clinical voice pathology. Although roughness is one of the inherent components of pathological voice assessment (e.g., GRBAS scale, CAPE-V protocols; De Bodt et al., 1997; Kempster et al., 2009), knowledge about the acoustic attributes contributing to roughness and its occurrence in typical voices is lacking. One contribution from this dissertation has been to characterize f0 and voice quality from period doubling (both articulatorily and acoustically), and to determine what kinds or modulation degrees of period doubling are heard as ambiguous in pitch (thus likely sounding 'rough'). What is directly relevant is the application of the quantitative findings of period-doubled pulses that occurred in typical speech to assessing 'diplophonia' in voice disorders. By knowing more about how pitch ambiguity (and consequently roughness) can result from voicing irregularity, we can learn more about relating perceived roughness and voice production. Using the attested amplitude and/ frequency modulation degrees found in typical speech, as quantified by the empirical amplitude/frequency ratios, the findings can be used by clinicians to improve diagnoses of vocal pathology from non-disordered speech, and therapy strategies.

As voice is the 'carrier' of speech, it lays a solid foundation for talker identification and speech recognition. The lower-level acoustic phonetic properties involved in manifesting a lin-

guistic voice category are also used in constructing a talker's idiosyncratic traits of their voice identity, because the different uses of voice share the articulators – the larynx and vocal folds. Voice quality is manifested by the same set of acoustic features: ranges of frequencies and spectral profiles that contribute to the percept of pitch and timbre of the voice, intensity that contribute to loudness, etc. Thus, the linguistic use of different voicing types interacts with the general use of voice, especially in speaker normalization. Speakers may possess different default voice qualities, such as modal, creaky, or breathy. When these speakers make use of voice quality linguistically, we will relevel the baseline according to each speaker's default voice quality and our perception of linguistic voice would be shaped differently. We will also have expectations towards the degree of creakiness or breathiness being used by different speakers. Studying the effects of different talkers on perception of linguistic voice quality is thus meaningful to both general and linguistic uses of voice, and will contribute to an integral theory of voice.

This work also has implications for the interface between phonetic details and phonological categories regarding voice quality. Period doubling and vocal fry are both manifestations of sub-phonemic creaky voice that occur in Mandarin tones. As we switch how we view non-contrastive phonetic events (such as sub-phonemic voice quality) from physiological phenomena to part of the grammar, we rethink about the representation of speech sound categories. Exploring the fine phonetic details of subtypes of creaky voice furthers our understanding of their roles in speech production and perception. The current results predict these creak subtypes have different roles in the linguistic signal. The next step might be to probe what phonological function or linguistic importance is associated with a particular type of creaky voice when forming contrastive categories such as tone and phonation types.

5.6. Future directions

Where do we go beyond the phonetics of period doubling? In Sections 5.1-5.4, I lay out several future avenues following up from unanswered questions or remaining puzzles of the studies in

the dissertation. Here, I describe broader future directions to add to the existing explorations that may interface with other disciplines and connect to important theoretical frameworks.

5.6.1. Typology of creaky voice in world's languages

Part of the dissertation is built upon the framework of subtypes of creaky vice – the definitions of period doubling (a special case of 'multiply pulsed voice') and vocal fry followed the descriptions in Keating et al. 2015. This dissertation contributes to the framework of creaky voice subtypes, by adding a thorough description and analysis of period doubling (and vocal fry) with their different characteristics, in a tonal language where these two subtypes are found as allophonic features to tonal categories. This work expands Keating et al.'s description by demonstrating that period doubling is more than doubling of periods. Importantly, this voice blurs the 'robust' distinctions between modal and non-modal voice qualities, and challenges the generally assumed acoustic dimensions that suffice to capture a creak percept alone. In future work, researchers are encouraged to investigate the diversity of voicing types in other languages that exhibit a more complex relationship between tone, register, and phonation (Cantonese, Mpi: Silverman, 1997; Mazatec: Garellek and Keating, 2011; White Hmong: Esposito, 2012; Black Miao: Kuang, 2013; Zapotec: Pickett et al., 2010, among others; also see Esposito and Khan, 2020 for a cross-linguistic survey of phonation types). Both acoustic and articulatory investigations of richer samples of creaky voice identified from various languages will benefit the typological study of creaky voice.

5.6.2. Learning linguistic categories of voice quality

As Kreiman and Sidtis (2011) noted, a voice quality problem is a perception problem, ultimately. If a type of voicing is perceptually distinct and contributes to linguistic contrasts, it is linguistically meaningful. This dissertation has investigated perception of pitch and voice quality during period

doubling to further inform the role of period doubling in tone and phonation perception for future work.

Here, to approach the linguistic importance question in Section 5.5, I propose a perceptual study to test if linguistic functions of subtypes of creaky voice differ from each other in terms of building a perceptual or phonological category of phonation. Through a perceptual learning experiment, I ask the following research questions: what is the perceptual importance of subtypes of creaky voice (e.g., vocal fry, period doubling, and aperiodic voicing), and how would listeners of different language backgrounds select among the different acoustic parameters to form and acquire a categorical percept of creaky voice?

On the one hand, the proposed study aims to test the perceptual salience of different types of creaky voice in representing a linguistic contrast. On the other hand, the study probes the relationship among period doubling and other types of creaky voice: do they function together or individually? Period doubling often results in an indeterminate pitch percept and thus is a consistent reflector of creaky voice. However, other types of creaky voice such as vocal fry and aperiodic voice also exhibit roughness as a percept. Though studies have found that different kinds of non-modal phonation types can be perceptually distinct (Kreiman et al., 1993; Gerratt and Kreiman, 2001), it is unclear how they differ in signal linguistic categories and conveying linguistic meaning when interfacing with phonological and prosodic categories. The results will thus inform the dimensions of period doubling, vocal fry, and aperiodic voice in which listeners use to differentiate and form one or several 'creaky' categories.

5.7. Concluding remarks

This dissertation depicts the articulatory and acoustic traits, and investigates percept of pitch and voice during period doubling, a type of supraperiodic (neither periodic nor aperiodic) voicing (Gerratt and Kreiman, 2001), which frequently occurs in natural speech and is regarded as a sub-

type of creaky voice. Alternating glottal pulses that are distinct in pitch as well as voice qualities define the articulatory aspects of period doubling. Low spectral tilt and stronger second harmonic define the acoustic features of period doubling. Stronger degrees of frequency modulation in period doubling lead to a low tone bias in perception and roughness is (re)produced by speakers in imitation. Higher or lower fundamental frequencies of the original period-doubled voice interact with modulation effects.

Period doubling is established not only as a distinct phonetic category from other voicing types such as modal voice and vocal fry, but with different linguistic functions based on its unique articulatory and acoustic features, and its role in pitch and tone perception. The findings of period doubling provide insight into speech production, perception, and processing including articulatory and acoustic models, voice classifiers, perception of pitch, tone, and phonation, and phonetic convergence, among others.

APPENDIX

A1. Chapter 3

Table 3.A1: Word list used for elicitation of the production study in Huang (ur).

Sequence	Stimuli	Sequence	Stimuli
1-1-1	收割机 /ʂou1 kɤ1 tɕi1/ “harvester”	1-2-1	功德箱 /koŋ1 tɤ2 ɕjaŋ1/ “donation box”
1-3-1	八宝粥 /pa1 pau3 ʈʂou1/ “eight-ingredient porridge”	1-4-1	织布机 /ʈʂɻ̩1 pu4 tɕi1/ “sewing machine”
1-1-2	三八节 /san1 pa1 tɕje2/ “March 8th Festival (Women’s day)”	1-2-2	英格兰 /jiŋ1 kɤ2 lan2/ “England”
1-3-2	八股文 /pa1 ku3 wən2/ “eight-parted essay”	1-4-2	冤大头 /ɥɛn1 ta4 tʰou2/ “a person deceived on account of his generosity”
1-1-3	交杯酒 /tɕjau1 pei1 tɕjou3/ “cross-cupped wine”	1-2-3	八达岭 /pa1 ta2 liŋ3/ “Badaling (the Great Wall)”
1-3-3	东北虎 /toŋ1 pei3 xu3/ “Siberian tiger”	1-4-3	招待所 /ʈʂau1 tai4 swo3/ “hotel”
1-1-4	三八线 /san1 pa1 ɕjɛn4/ “38°N Line (Military Demarcation Line)”	1-2-4	孤独症 /ku1 tu2 ʈʂəŋ4/ “autism”
1-3-4	亲笔信 /tɕʰin1 pi3 ɕin4/ “handwritten letter”	1-4-4	兄弟会 /ɕjoŋ1 ti4 xwei4/ “fraternity”
2-1-1	男低音 /nan2 ti1 jin1/ “bass (male singer)”	2-2-1	白皮书 /pai2 pʰi2 ʂu1/ “White Paper”
2-3-1	平底锅 /pʰiŋ2 ti3 kwo1/ “fry pan”	2-4-1	红豆羹 /xoŋ2 tou4 kəŋ1/ “red bean soup”

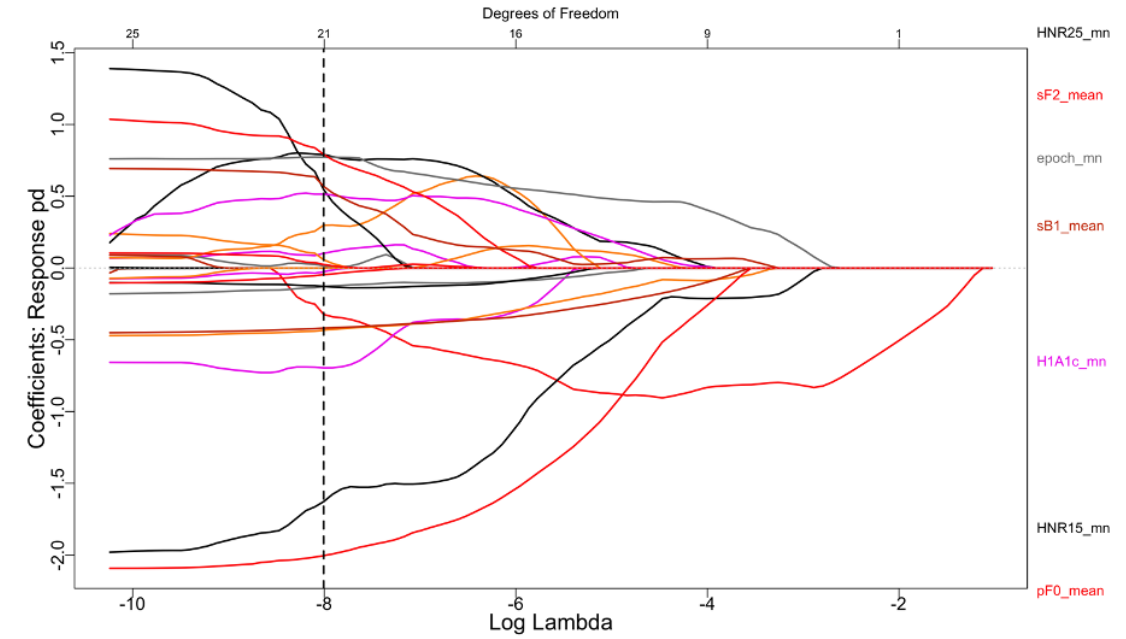
Table 3.A1: Word list used for elicitation of the production study in Huang (under review). (cont.)

Sequence	Stimuli	Sequence	Stimuli
	长歌行		齐白石
2-1-2	/ʈʂʰaŋ2 kɤ1 ɕiŋ2/	2-2-2	/tɕʰi2 pai2 ʂɻ̩2/
	“Changge Xing (poem title)”		“Baishi Qi (person name)”
	藏宝图		邮递员
2-3-2	/tsʰaŋ2 pau3 tʰu2/	2-4-2	/jou2 ti4 ɥɛn2/
	“treasure map”		“mailman”
	皮包骨		皮革厂
2-1-3	/pʰi2 pau1 ku3/	2-2-3	/pʰi2 kɤ2 ʈʂʰaŋ3/
	“all skin and bones”		“leather factory”
	劳改所		房地产
2-3-3	/lau2 kai3 swo3/	2-4-3	/faŋ2 ti4 ʈʂʰan3/
	“labor camp”		“real estate”
	荷包蛋		合格线
2-1-4	/xɤ2 pau1 tan4/	2-2-4	/xɤ2 kɤ2 ɕjɛn4/
	“poached egg”		“pass score”
	毛笔字		情报站
2-3-4	/mau2 pi3 tsɻ4/	2-4-4	/tɕiŋ2 pau4 ʈʂan4/
	“calligraphy written by brush”		“information station”
	女低音		脑白金
3-1-1	/ny3 ti1 jin1/	3-2-1	/nau3 pai2 tɕin1/
	“alto (female singer)”		“Naobaijin (supplement brand)”
	百宝箱		北大荒
3-3-1	/pai3 pau3 ɕjaŋ1/	3-4-1	/pei3 ta4 xwaŋ1/
	“treasure box”		“Beidahuang (region name)”
	短歌行		搞独裁
3-1-2	/tuan3 kɤ1 ɕiŋ2/	3-2-2	/kau3 tu2 tsʰai2/
	“Duange Xing (poem title)”		“to dictate”
	甲骨文		胆固醇
3-3-2	/tɕa3 ku3 wən2/	3-4-2	/tan3 ku4 ʈʂʰwən2/
	“Oracle bone script”		“cholesterol”
	剪刀手		小白马
3-1-3	/tɕjɛn3 tau1 ʂou3/	3-2-3	/ɕjau3 pai2 ma3/
	“V gesture”		“little white horse”
	老古董		党代表
3-3-3	/lau3 ku3 toŋ3/	3-4-3	/taŋ3 tai4 pjau3/
	“old-fashioned”		“party representative”
	等高线		洗涤剂
3-1-4	/təŋ3 kau1 ɕjɛn4/	3-2-4	/ɕi3 ti2 tɕi4/
	“contour line”		“detergent”

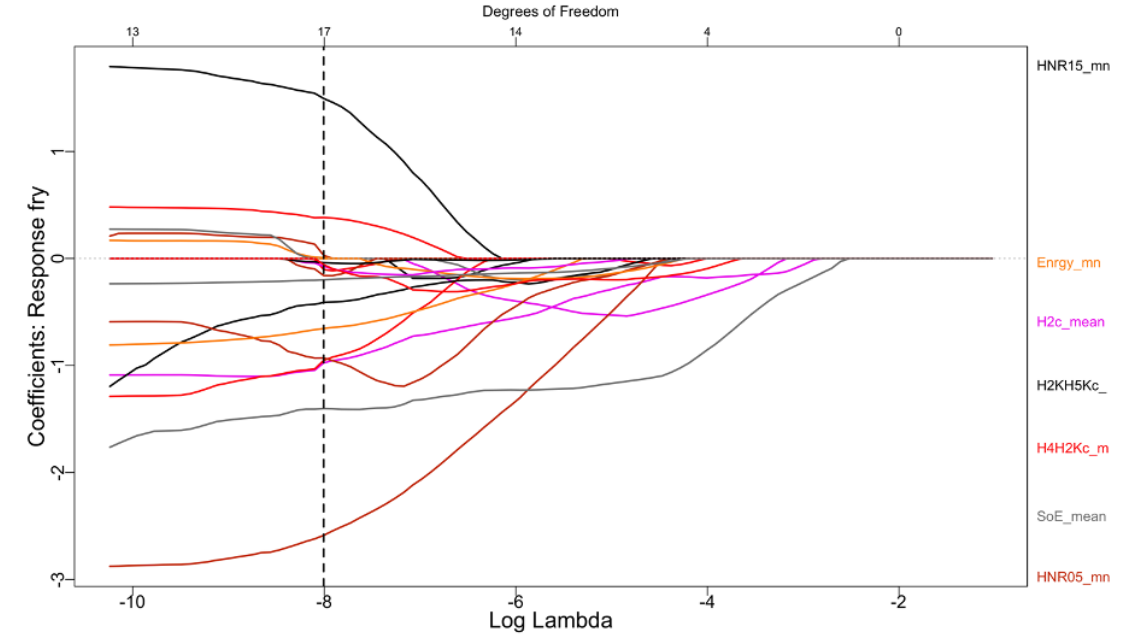
Table 3.A1: Word list used for elicitation of the production study in Huang (under review). (cont.)

Sequence	Stimuli	Sequence	Stimuli
3-3-4	简笔画 /tɕan3 pi3 xwa4/ “sketch”	3-4-4	土地税 /tʰu3 ti4 ʂwei4/ “property tax”
4-1-1	腊八粥 /la4 pa1 ʈʂou1/ “Laba rice porridge)”	4-2-1	复读机 /fu4 tu2 tɕi1/ “repeater”
4-3-1	赋比兴 /fu4 pi3 ɕiŋ1/ “exposition, comparison, and affective image (rhetorics)”	4-4-1	落地窗 /lwo4 ti4 ʈʂʰwaŋ1/ “French window”
4-1-2	腊八节 /la4 pa1 tɕje2/ “Laba Festival”	4-2-2	蛋白酶 /tan4 pai2 mei2/ “protease”
4-3-2	聚宝盆 /tɕy4 pau3 pʰən2/ “treasure bowl”	4-4-2	并蒂莲 /piŋ4 ti4 ljɛn2/ “one-stalked twin lotus flowers”
4-1-3	奥巴马 /au4 pa1 ma3/ “Obama”	4-2-3	信达雅 /ɕin4 ta2 ja3/ “faithfulness, expressiveness and elegance”
4-3-3	半导体 /pan4 tau3 tʰi3/ “semiconductor”	4-4-3	胖大海 /pʰaŋ4 ta4 xai3/ “Sterculia lychnophora”
4-1-4	大都会 /ta4 tu1 xwei4/ “metropolis”	4-2-4	大革命 /ta4 kɤ2 miŋ4/ “grand revolution”
4-3-4	对比度 /twei4 pi3 tu4/ “contrast”	4-4-4	放大镜 /faŋ4 ta4 tɕiŋ4/ “magnifier”

The results of the shrunk coefficients of the acoustic features, and the combined acoustic and articulatory features, during Lasso regularization are shown in Figures 3.A1 and 3.A2. These demonstrate the process of selecting variables and how the number and coefficients of the variables vary as a function of lambda for different voice categories.

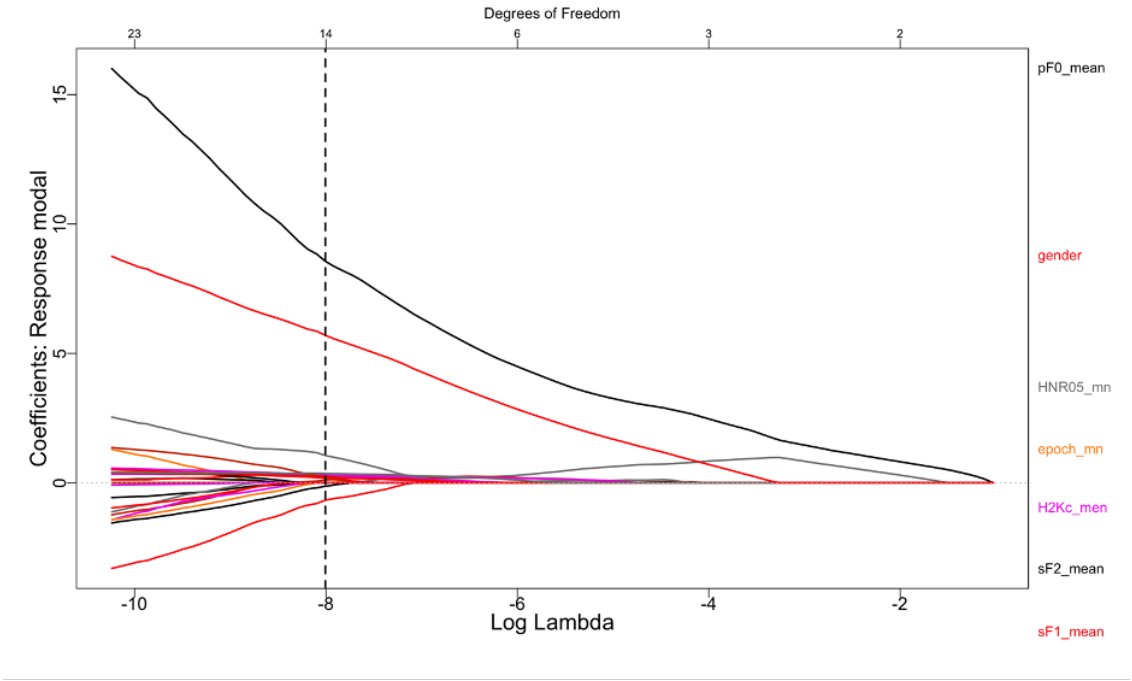


(a)



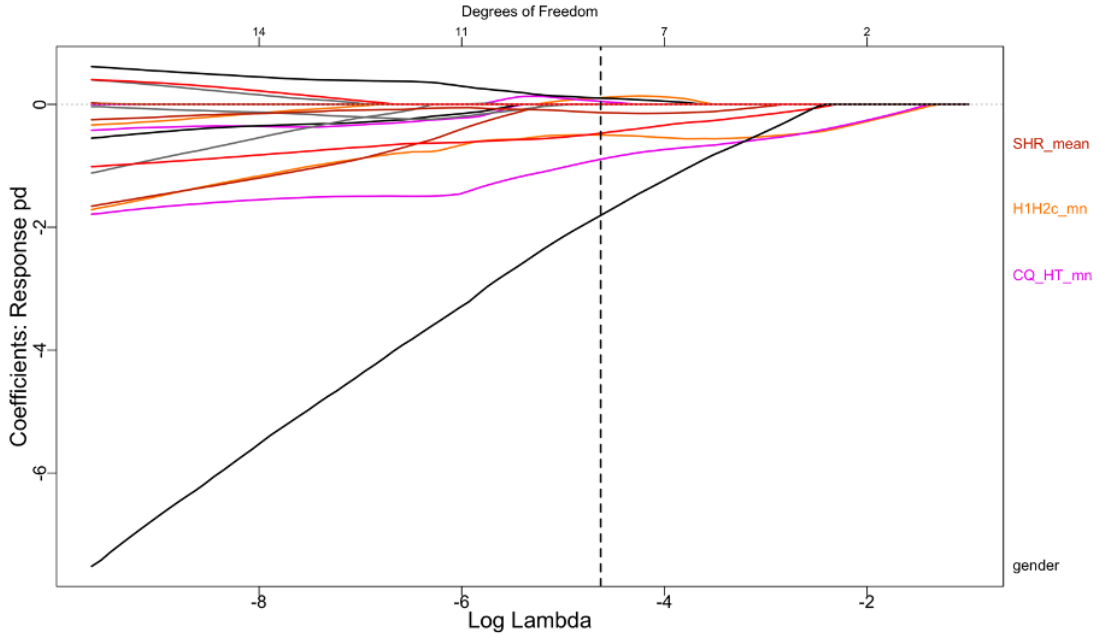
(b)

Figure 3.A1: Shrunk coefficients of acoustic features in classifying period doubling (a), vocal fry (b), and modal voice (c) using different lambda values during the Lasso selection process. The dashed vertical line shows the optimal lambda with the minimum classification error. The final seven predictors of biggest coefficients are labeled.

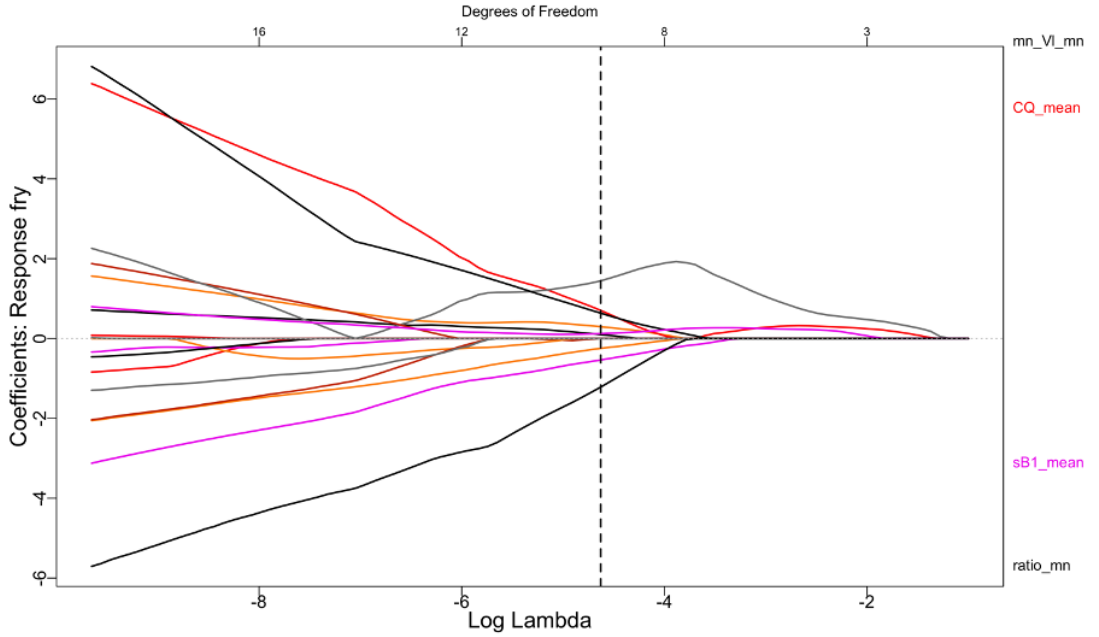


(c)

Figure 3.A1: Shrunk coefficients of acoustic features in classifying period doubling (a), vocal fry (b), and modal voice (c) using different lambda values during the Lasso selection process. The dashed vertical line shows the optimal lambda with the minimum classification error. The final seven predictors of biggest coefficients are labeled. (cont.)

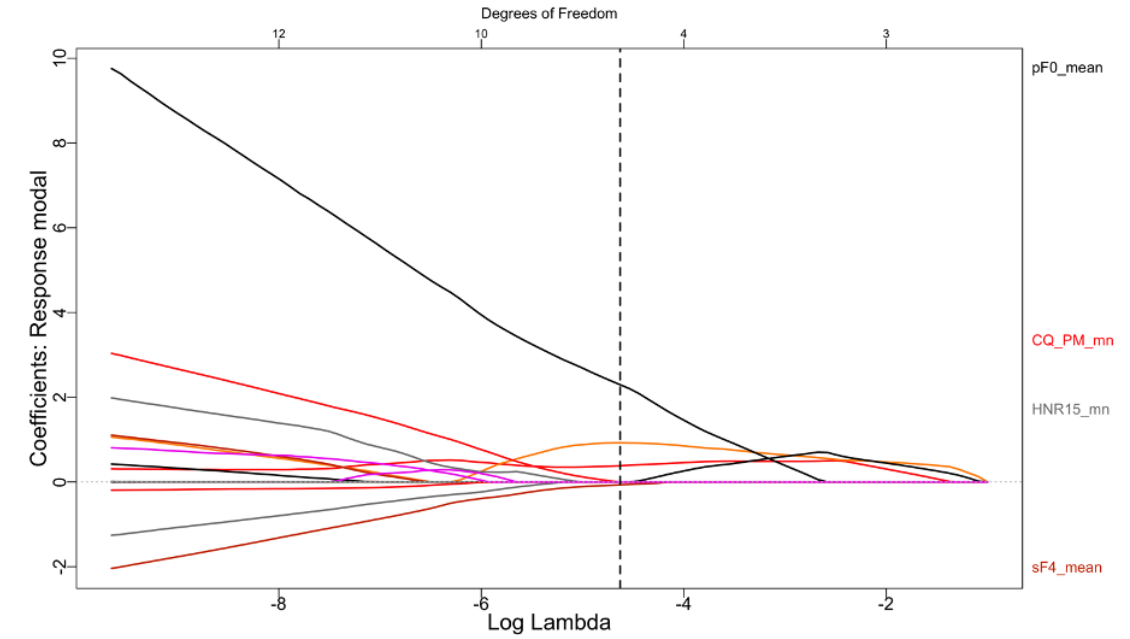


(a)



(b)

Figure 3.A2: Shrunk coefficients of acoustic and articulatory features in classifying period doubling (a), vocal fry (b), and modal voice (c) using different lambda values during the Lasso selection process. The dashed vertical line shows the optimal lambda with the minimum classification error. The final four predictors of biggest coefficients are labeled.



(c)

Figure 3.A2: Shrunk coefficients of acoustic and articulatory features in classifying period doubling (a), vocal fry (b), and modal voice (c) using different lambda values during the Lasso selection process. The dashed vertical line shows the optimal lambda with the minimum classification error. The final four predictors of biggest coefficients are labeled. (cont.)

BIBLIOGRAPHY

Aichinger, P. (2014). *Diplophonic Voice*. PhD thesis, Medical University of Vienna.

Aichinger, P., Hagmüller, M., Roesner, I., Bigenzahn, W., Schneider-Stickler, B., and Schoentgen, J. (2016). Diplophonia disturbs jitter and shimmer measurement. *Folia Phoniatrica et Logopaedica*, 68(1):22–28.

Aichinger, P., Pernkopf, F., and Schoentgen, J. (2019). Detection of extra pulses in synthesized glottal area waveforms of dysphonic voices. *Biomedical signal processing and control*, 50:158–167.

Aichinger, P., Roesner, I., Schneider-Stickler, B., Leonhard, M., Denk-Linnert, D.-M., Bigenzahn, W., Fuchs, A. K., Hagmüller, M., and Kubin, G. (2017). Towards objective voice assessment: The diplophonia diagram. *Journal of voice*, 31(2):253–e17.

Alexander, J. A., Bradlow, A. R., Ashley, R. D., and Wong, P. C. (2011). Music-melody perception in tone-language and non-tone-language speakers. *Psycholinguistic representation of tone.*

Awan, S. N., Giovinco, A., and Owens, J. (2012). Effects of vocal intensity and vowel type on cepstral analysis of voice. *Journal of Voice*, 26(5):670.e15–670.e20.

Bae, I.-H., Wang, S.-G., Kwon, S.-B., Kim, S.-T., Sung, E.-S., and Lee, J.-C. (2019). Clinical application of two-dimensional scanning digital kymography in discrimination of diplophonia. *Journal of Speech, Language, and Hearing Research*, 62(10):3643–3654.

Bailly, G. (2003). Close shadowing natural versus synthetic speech. *International Journal of Speech Technology*, 6(1):11–19.

Bailly, L., Henrich, N., and Pelorson, X. (2010). Vocal fold and ventricular fold vibration in period-doubling phonation: Physiological description and aerodynamic modeling. *The Journal of the Acoustical Society of America*, 127(5):3212–3222.

Baken, R. J. and Orlikoff, R. F. (2000). *Clinical measurement of speech and voice*. Cengage Learning.

Bateman, L. A. (2003). *Soprano, style and voice quality: Acoustic and laryngographic correlates*. PhD thesis, University of Victoria.

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Batliner, A., Burger, S., Johne, B., and Kießling, A. (1994). MüSLI: A classification scheme for laryngealizations. In *Proceedings of ESCA Workshop on Prosody*, pages 176–179.

Bele, I. V. (2006). The speaker's formant. *Journal of Voice*, 20(4):555–578.

Belotel-Grenié, A. and Grenié, M. (1994). Phonation types analysis in standard Chinese. In *ICSLP*.

Belotel-Grenié, A. and Grenié, M. (1997). Types de phonation et tons en chinois standard. *Cahiers de linguistique-Asie orientale*, 26(2):249–279.

Belotel-Grenié, A. and Grenié, M. (2004). The creaky voice phonation and the organisation of Chinese discourse. In *International symposium on tonal aspects of languages: With emphasis on tone languages*.

Bergan, C. C. and Titze, I. R. (2001). Perception of pitch and roughness in vocal signals with subharmonics. *Journal of Voice*, 15(2):165–175.

Bickley, C. (1982). Acoustic analysis and perception of breathy vowels. *Speech communication group working papers*, 1:71–81.

Bidelman, G. M., Gandour, J. T., and Krishnan, A. (2011). Musicians and tone-language speakers share enhanced brainstem encoding but not perceptual benefits for musical pitch. *Brain and cognition*, 77(1):1–10.

Boersma, P. and Weenink, D. (2022). Praat: doing phonetics by computer (version 5.1.13).

Borkowska, B. and Pawlowski, B. (2011). Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour*, 82(1):55–59.

Bowler, N. W. (1964). A fundamental frequency analysis of harsh vocal quality. *Communications Monographs*, 31(2):128–134.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, 44(1):108–132.

Brunelle, M., Nguyên, D. D., and Nguyên, K. H. (2010). A laryngographic and laryngoscopic study of Northern Vietnamese tones. *Phonetica*, 67(3):147–169.

Buhr, R. and Keating, P. (1977). Spectrographic effects of register shifts in speech production. *The Journal of the Acoustical Society of America*, 62(S1):S25–S25.

Chambers, C., Akram, S., Adam, V., Pelofi, C., Sahani, M., Shamma, S., and Pressnitzer, D. (2017). Prior context in audition informs binding and shapes simple features. *Nature communications*, 8(1):1–11.

Chao, Y. R. (1968). *A grammar of spoken Chinese*. University of California Berkeley Press.

Chen, Y., Robb, M. P., and Gilbert, H. R. (2002). Electroglottographic evaluation of gender and vowel effects during modal and vocal fry phonation. *Journal of Speech, Language, and Hearing Research*, 45(5):821–829.

Childers, D. G., Hicks, D., Moore, G., Eskenazi, L., and Lalwani, A. (1990). Electroglottography and vocal fold physiology. *Journal of Speech, Language, and Hearing Research*, 33(2):245–254.

Cholin, J. and Levelt, W. J. (2009). Effects of syllable preparation and syllable frequency in speech production: Further evidence for syllabic units at a post-lexical level. *Language and cognitive processes*, 24(5):662–684.

Coleman, R. O. (1971). Male and female voice quality and its relationship to vowel formant frequencies. *Journal of Speech and Hearing Research*, 14(3):565–577.

Creel, S. C., Weng, M., Fu, G., Heyman, G. D., and Lee, K. (2018). Speaking a tone language enhances musical pitch perception in 3–5-year-olds. *Developmental science*, 21(1):e12503.

Crowhurst, M. J. (2018). The joint influence of vowel duration and creak on the perception of internal phrase boundaries. *The Journal of the Acoustical Society of America*, 143(3):EL147–EL153.

Dallaston, K. and Docherty, G. (2020). The quantitative prevalence of creaky voice (vocal fry) in varieties of English: A systematic review of the literature. *PloS one*, 15(3):e0229960.

D'Amario, S. and Daffern, H. (2017). Using electrolaryngography and electroglottography to assess the singing voice: A systematic review. *Psychomusicology: Music, Mind, and Brain*, 27(4):229.

Davidson, L. (2019). The effects of pitch, gender, and prosodic context on the identification of creaky voice. *Phonetica*, 76(4):235–262.

Davidson, L. (2020). Contributions of modal and creaky voice to the perception of habitual pitch. *Language*, 96(1):e22–e37.

Davidson, L. (2021). The versatility of creaky phonation: Segmental, prosodic, and sociolinguistic uses in the world's languages. *WIREs Cognitive Science*, 12(3):e1547.

Davison, D. S. (1991). An acoustic study of so-called creaky voice in Tianjin Mandarin. *UCLA Working papers in Phonetics*, 78:50–57.

Dawson, H., Garellek, M., López-Francisco, O., and Amith, J. D. (2022). Tense voice without the high f0: The case of glottalized vowels in Zongozotla Totonac. *The Journal of the Acoustical Society of America*, 152(4):A286–A286.

De Bodt, M. S., Wuyts, F. L., Van de Heyning, P. H., and Croux, C. (1997). Test-retest study of the grbas scale: Influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice*, 11(1):74–80.

de Cheveigné, A. (2005). *Pitch Perception Models*, pages 169–233. Springer New York, New York, NY.

Dejonckere, P. H. and Lebacq, J. (1983). An analysis of the diplophonia phenomenon. *Speech Communication*, 2(1):47–56.

DiCanio, C. T. (2009). The phonetics of register in Takhian Thong Chong. *Journal of the International Phonetic Association*, 39(2):162–188.

DiCanio, C. T. (2012). Coarticulation between tone and glottal consonants in Itunyoso Trique. *Journal of Phonetics*, 40(1):162–176.

D'Imperio, M., Cavone, R., and Petrone, C. (2014). Phonetic and phonological imitation of intonation in two varieties of Italian. *Frontiers in psychology*, 5:1226.

Drugman, T., Kane, J., and Gobl, C. (2014). Data-driven detection and analysis of the patterns of creaky voice. *Computer Speech & Language*, 28(5):1233–1253.

Drugman, T., Kane, J., and Gobl, C. (2020). Data-driven detection and analysis of the patterns of creaky voice. *arXiv preprint arXiv:2006.00518*.

Echternach, M., Dippold, S., Sundberg, J., Arndt, S., Zander, M. F., and Richter, B. (2010). High-speed imaging and electroglottography measurements of the open quotient in untrained male voices' register transitions. *Journal of voice*, 24(6):644–650.

Eriksson, A. and Wretling, P. (1997). How flexible is the human voice?-A case study of mimicry. In *Fifth European Conference on Speech Communication and Technology*.

Esling, J. H. (1978a). The identification of features of voice quality in social groups. *Journal of the International Phonetic Association*, 8(1-2):18–23.

Esling, J. H. (1978b). Voice quality in Edinburgh - a sociolinguistic and phonetic study.

Esling, J. H., Dickson, B. C., and Snell, R. C. (1992). Analysis of phonation type using laryngographic techniques. In *ICSLP*, volume 92, pages 1107–1110.

Esling, J. H., Moisik, S., Benner, A., and Crevier-Buchman, L. (2019). *Voice Quality: The Laryngeal Articulator Model*. Number 1. Cambridge University Press.

Esposito, C. M. (2010). Variation in contrastive phonation in Santa Ana del Valle Zapotec. *Journal of the International Phonetic Association*, 40(2):181–198.

Esposito, C. M. (2012). An acoustic and electroglottographic study of White Hmong tone and phonation. *Journal of Phonetics*, 40(3):466–476.

Esposito, C. M. and Khan, S. u. D. (2012). Contrastive breathiness across consonants and vowels: A comparative study of Gujarati and White Hmong. *Journal of the International Phonetic Association*, 42(2):123–143.

Esposito, C. M. and Khan, S. u. D. (2020). The cross-linguistic patterns of phonation types. *Language and Linguistics Compass*, 14(12):e12392.

Fant, G., Liljencrants, J., Lin, Q.-G., et al. (1985). A four-parameter model of glottal flow. *Speech, Music and Hearing Quarterly Progress and Status Report*, 4(1985):1–13.

Ferrand, C. T. (2002). Harmonics-to-noise ratio: an index of vocal aging. *Journal of Voice*, 16(4):480–487.

Ferrer, C., De Bodt, M., Hernández-Díaz, M., Van de Heyning, P., and Maryn, Y. (2007). Properties of the cepstral peak prominence and its usefulness in vocal quality measurements. In *Models and analysis of vocal emissions for biomedical applications: 5th international workshop*, pages 93–96. Firenze University Press.

Fourcin, A. (1986). Electrolaryngographic assessment of vocal fold function. *Journal of Phonetics*, 14(3-4):435–442.

Fraile, R. and Godino-Llorente, J. I. (2014). Cepstral peak prominence: A comprehensive analysis. *Biomedical Signal Processing and Control*, 14:42–54.

Fraj, S., Schoentgen, J., and Grenez, F. (2012). Development and perceptual assessment of a synthesizer of disordered voices. *The Journal of the Acoustical Society of America*, 132(4):2603–2615.

Fuks, L., Hammarberg, B., and Sundberg, J. (1998). A self-sustained vocal-ventricular phonation mode: acoustical, aerodynamic and glottographic evidences. *KTH TMH-QPSR*, 3(1998):49–59.

Gandour, J. T. (1978). The perception of tone. In *Tone*, pages 41–76. Elsevier.

Garellek, M. (2014). Voice quality strengthening and glottalization. *Journal of Phonetics*, 45:106–113.

Garellek, M. (2019). The phonetics of voice. In *The Routledge handbook of phonetics*, pages 75–106. Routledge.

Garellek, M. (2020). Acoustic discriminability of the complex phonation system in !xóõ. *Phonetica*, 77(2):131–160.

Garellek, M. (2022). Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *Journal of Phonetics*, 94:101155.

Garellek, M., Chai, Y., Huang, Y., and Van Doren, M. (2021). Voicing of glottal consonants and non-modal vowels. *Journal of the International Phonetic Association*, pages 1–28.

Garellek, M. and Esposito, C. M. (2021). Phonetics of White Hmong vowel and tonal contrasts. *Journal of the International Phonetic Association*, pages 1–20.

Garellek, M. and Keating, P. (2011). The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *Journal of the International Phonetic Association*, 41(2):185–205.

Garellek, M., Keating, P., Esposito, C. M., and Kreiman, J. (2013). Voice quality and tone identification in White Hmong. *The Journal of the Acoustical Society of America*, 133(2):1078–1089.

Garellek, M., Ritchart, A., and Kuang, J. (2016). Breathy voice during nasality: A cross-linguistic study. *Journal of Phonetics*, 59:110–121.

Garellek, M. and Seyfarth, S. (2016). Acoustic differences between English /t/ glottalization and phrasal creak. In *Interspeech*, pages 1054–1058.

Garnier, M., Lamalle, L., and Sato, M. (2013). Neural correlates of phonetic convergence and speech imitation. *Frontiers in psychology*, 4:600.

Gaskill, C. S., Awan, J. A., Watts, C. R., and Awan, S. N. (2017). Acoustic and perceptual classification of within-sample normal, intermittently dysphonic, and consistently dysphonic voice types. *Journal of Voice*, 31(2):218–228.

Gerratt, B. R. and Kreiman, J. (2001). Toward a taxonomy of nonmodal phonation. *Journal of Phonetics*, 29(4):365–381.

Gerratt, B. R., Precoda, K., Hanson, D. G., and Berke, G. S. (1988). Source characteristics of diplophonia. *The Journal of the Acoustical Society of America*, 83(S1):S66–S66.

Gessinger, I., Raveh, E., Steiner, I., and Möbius, B. (2021). Phonetic accommodation to natural and synthetic voices: Behavior of groups and individuals in speech shadowing. *Speech Communication*, 127:43–63.

Gibson, T. A. (2017). The role of lexical stress on the use of vocal fry in young adult female speakers. *Journal of Voice*, 31(1):62–66.

Gillespie, A. I., Dastolfo, C., Magid, N., and Gartner-Schmidt, J. (2014). Acoustic analysis of four common voice diagnoses: moving toward disorder-specific assessment. *Journal of Voice*, 28(5):582–588.

Gobl, C., , and Chasaide, A. N. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40(1):189–212.

Gobl, C. (1989). A preliminary study of acoustic voice quality correlates. *STL-QPSR*, 4:9–21.

Gordon, M. (2001). Linguistic aspects of voice quality with special reference to Athabaskan. In *Proceedings of the 2001 Athabaskan languages conference*, pages 163–178. Citeseer.

Granqvist, S. (2003). The visual sort and rate method for perceptual evaluation in listening tests. *Logopedics Phoniatrics Vocology*, 28(3):109–116.

Granqvist, S., Hertegård, S., Larsson, H., and Sundberg, J. (2003). Simultaneous analysis of vocal fold vibration and transglottal airflow: Exploring a new experimental setup. *Journal of Voice*, 17(3):319–330.

Guion, S. G., Post, M. W., and Payne, D. L. (2004). Phonetic correlates of tongue root vowel contrasts in Maa. *Journal of Phonetics*, 32(4):517–542.

Guzmán, M., Castro, C., Madrid, S., Olavarria, C., Leiva, M., Muñoz, D., Jaramillo, E., and Laukkanen, A.-M. (2016). Air pressure and contact quotient measures during different semioccluded postures in subjects with different voice conditions. *Journal of voice*, 30(6):759–e1.

Hanson, D. G., Gerratt, B. R., and Ward, P. H. (1983). Glottographic measurement of vocal dysfunction: a preliminary report. *Annals of Otology, Rhinology & Laryngology*, 92(5):413–420.

Hanson, H. M., Stevens, K. N., Kuo, H.-K. J., Chen, M. Y., and Slifka, J. (2001). Towards models of phonation. *Journal of Phonetics*, 29(4):451–480.

Hedelin, P. and Huber, D. (1990). Pitch period determination of aperiodic speech signals. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 361–364. IEEE.

Henrich, N., d'Alessandro, C., Doval, B., and Castellengo, M. (2005). Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency. *The Journal of the Acoustical Society of America*, 117(3):1417–1430.

Henrich, N., Lortat-Jacob, B., Castellengo, M., Bailly, L., and Pelorson, X. (2006). Period-doubling occurrences in singing: The 'bassu' case in traditional Sardinian 'A Tenore' singing. *Proc. ICVPB*.

Henton, C. G. (1989). Sociophonetic aspects of creaky voice. *The Journal of the Acoustical Society of America*, 86(S1):S26–S26.

Herbst, C. T. (2004). Evaluation of various methods to calculate the EGG contact quotient. *Department of Speech, Music and Hearing, KTH, Stockholm, Sweden* (*diploma thesis in music acoustics*).

Herbst, C. T. (2020). Electroglottography–an update. *Journal of Voice*, 34(4):503–526.

Herbst, C. T., Schutte, H. K., Bowling, D. L., and Svec, J. G. (2017). Comparing chalk with cheese—the egg contact quotient is only a limited surrogate of the closed quotient. *Journal of Voice*, 31(4):401–409.

Herzel, H. (1993). Bifurcations and chaos in voice signals. *Applied Mechanics Reviews*, 46(7):399–413.

Herzel, H., Berry, D., Titze, I. R., and Saleh, M. (1994). Analysis of vocal disorders with methods from nonlinear dynamics. *Journal of Speech, Language, and Hearing Research*, 37(5):1008–1019.

Herzel, H. and Reuter, R. (1996). Biphonation in voice signals. *AIP Conference Proceedings*, 375(1):644–657.

Herzel, H. and Reuter, R. (1997). Whistle register and biphonation in a child's voice. *Folia phoniatrica et logopaedica*, 49(5):216–224.

Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research*, 37(4):769–778.

Hollien, H. and Michel, J. F. (1968). Vocal fry as a phonational register. *Journal of Speech and Hearing Research*, 11(3):600–604.

Holmberg, E. B., Hillman, R. E., Perkell, J. S., Guiod, P. C., and Goldman, S. L. (1995a). Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice. *Journal of Speech, Language, and Hearing Research*, 38(6):1212–1223.

Holmberg, E. B., Hillman, R. E., Perkell, J. S., Guiod, P. C., and Goldman, S. L. (1995b). Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice. *Journal of Speech, Language, and Hearing Research*, 38(6):1212–1223.

Howard, D. M. (1995). Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers. *Journal of Voice*, 9(2):163–172.

Hsieh, K.-T., Dong, D.-H., and Wang, L.-Y. (2013). A preliminary study of applying shadowing technique to English intonation instruction. *Taiwan Journal of Linguistics*, 11(2).

Huang, Y. (2020). Different attributes of creaky voice distinctly affect Mandarin tonal perception. *The Journal of the Acoustical Society of America*, 147(3):1441–1458.

Huang, Y. (2022). Articulatory properties of period-doubled voice in Mandarin. *Proc. Speech Prosody 2022*, pages 545–549.

Huang, Y. (u.r.). F0 and voice quality of coarticulated Mandarin tones. *Language and Speech*.

Huang, Y., Athanasopoulou, A., and Vogel, I. (2018). The effect of focus on creaky phonation in Mandarin Chinese tones. *University of Pennsylvania Working Papers in Linguistics*, 24(1):12.

Imaizumi, S. and Gauffin, J. (1992). Acoustic and perceptual modelling of the voice quality caused by fundamental frequency perturbation. In *ICSLP*.

Irons, S. T. and Alexander, J. E. (2016). Vocal fry in realistic speech: Acoustic characteristics and perceptions of vocal fry in spontaneously produced and read speech. *The Journal of the Acoustical Society of America*, 140(4):3397–3397.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Kania, R. E., Hans, S., Hartl, D. M., Clement, P., Crevier-Buchman, L., and Brasnu, D. F. (2004). Variability of electroglottographic glottal closed quotients: necessity of standardization to obtain normative values. *Archives of Otolaryngology–Head & Neck Surgery*, 130(3):349–352.

Keating, P. (1980). Patterns of fundamental frequency and vocal registers. *Infant communication: cry and early speech*, page 209.

Keating, P., Esposito, C., Garellek, M., Khan, S. u. D., and Kuang, J. (2010). Phonation contrasts across languages. In *Poster presented at the 12th Conference on Laboratory Phonology*.

Keating, P., Garellek, M., and Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice. In *ICPhS*, volume 2015, pages 2–7.

Keating, P. and Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *The Journal of the Acoustical Society of America*, 132(2):1050–1060.

Kelman, A. (1981). Vibratory pattern of the vocal folds. *Folia Phoniatrica et Logopaedica*, 33(2):73–99.

Kelterer, A. and Schuppler, B. (2020). Phonation type contrasts and tone in Chichimec. *The Journal of the Acoustical Society of America*, 147(4):3043–3059.

Kempster, G. B., Gerratt, B. R., Abbott, K. V., Barkmeier-Kraemer, J., and Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*, 18(2):124–132.

Keyser, S. J. and Stevens, K. N. (2006). Enhancement and overlap in the speech chain. *Language*, 82(1):33–63.

Khan, S. u. D., Becker, K., and Zimman, L. (2015). The acoustics of perceived creaky voice in American English. In *170th Meeting of the Acoustical Society of America*.

Kim, S., Matachana, C., Nyman, A., and Yu, K. M. (2020). Creak in the phonetic space of low tones in Beijing Mandarin, Cantonese, and White Hmong. In *10th International Conference on Speech Prosody 2020, Tokyo*, pages 523–527.

Kirby, J. P. (2014). Incipient tonogenesis in Phnom Penh Khmer: Acoustic and perceptual studies. *Journal of Phonetics*, 43:69–85.

KIRITANI, S. (1995). Vocal fold vibrations associated with involuntary voice changes in certain pathological cases. *Vocal Fold Physiology, Voice Quality Control*, pages 269–281.

Kiritani, S., Hirose, H., and Imagawa, H. (1993). High-speed digital image analysis of vocal cord vibration in diplophonia. *Speech communication*, 13(1-2):23–32.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*, 67(3):971–995.

Klatt, D. H. and Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, 87(2):820–857.

Kramer, E., Linder, R., and Schönweiler, R. (2013). A study of subharmonics in connected speech material. *Journal of Voice*, 27(1):29–38.

Kreiman, J. (1982). Perception of sentence and paragraph boundaries in natural conversation. *Journal of Phonetics*, 10(2):163–175.

Kreiman, J., Antoñanzas-Barroso, N., and Gerratt, B. R. (2010). Integrated software for analysis and synthesis of voice quality. *Behavior research methods*, 42(4):1030–1041.

Kreiman, J. and Gerratt, B. R. (2005). Perception of aperiodicity in pathological voice. *The Journal of the Acoustical Society of America*, 117(4):2201–2211.

Kreiman, J., Gerratt, B. R., Precoda, K., and Berke, G. S. (1993). Perception of supraperiodic voices. *The Journal of the Acoustical Society of America*, 93(4):2337–2337.

Kreiman, J., Shue, Y.-L., Chen, G., Iseli, M., Gerratt, B. R., Neubauer, J., and Alwan, A. (2012). Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. *The Journal of the Acoustical Society of America*, 132(4):2625–2632.

Kreiman, J. and Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons.

Krivokapić, J. (2014). Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1658):20130397.

Kuang, J. (2013). The tonal space of contrastive five level tones. *Phonetica*, 70(1-2):1–23.

Kuang, J. (2017). Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *The Journal of the Acoustical Society of America*, 142(3):1693–1706.

Kuang, J. (2018). The influence of tonal categories and prosodic boundaries on the creakiness in Mandarin. *The Journal of the Acoustical Society of America*, 143(6):EL509–EL515.

Kuang, J. and Keating, P. (2014). Glottal articulations in tense vs. lax phonation contrasts. *Journal of the Acoustical Society of America*, 136(5):2784–2797.

Kuang, J. and Liberman, M. (2016). The effect of vocal fry on pitch perception. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pages 5260–5264.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. (2017). lmertest package: tests in linear mixed effects models. *Journal of statistical software*, 82:1–26.

Lambert, S. (1992). Shadowing. *Meta*, 37(2):263–273.

Laver, J. (1980). The phonetic description of voice quality. *Cambridge Studies in Linguistics London*, 31:1–186.

Leonard, R. J., Ringel, R., Horii, Y., and Daniloff, R. (1988). Vocal shadowing in singers and nonsingers. *Journal of Speech, Language, and Hearing Research*, 31(1):54–61.

Leong, K., Hawkshaw, M. J., Dentchev, D., Gupta, R., Lurie, D., and Sataloff, R. T. (2013). Reliability of objective voice measures of normal speaking voices. *Journal of Voice*, 27(2):170–176.

Levitt, A. and Lucas, M. (2018). Effects of four voice qualities and formant dispersion on perception of a female voice. *Psychology of Language and Communication*, 22(1):394–416.

Li, S., Gu, W., Liu, L., and Tang, P. (2020). The role of voice quality in Mandarin sarcastic speech: An acoustic and electroglottographic study. *Journal of Speech, Language, and Hearing Research*, 63(8):2578–2588.

Lindestad, P.-Å., Södersten, M., Merker, B., and Granqvist, S. (2001). Voice source characteristics in Mongolian "throat singing" studied with high-speed imaging technique, acoustic spectra, and inverse filtering. *Journal of Voice*, 15(1):78–85.

Local, J. K., Kelly, J., and Wells, W. H. (1986). Towards a phonology of conversation: turn-taking in Tyneside English. *Journal of Linguistics*, 22(2):411–437.

Ma, E. P.-M. and Love, A. L. (2010). Electroglottographic evaluation of age and gender effects during sustained phonation and connected speech. *Journal of voice*, 24(2):146–152.

Manning, F. C., Siminoski, A., and Schutz, M. (2020). Exploring the effects of effectors: Finger synchronization aids rhythm perception similarly in both pianists and non-pianists. *Music Perception*, 37(3):196–207.

Marasek, K. (1996). Glottal correlates of the word stress and the tense/lax opposition in German. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1573–1576. IEEE.

Marslen-Wilson, W. D. (1985). Speech shadowing and speech comprehension. *Speech communication*, 4(1-3):55–73.

Martin, P. (2012). Automatic detection of voice creak. In *Speech Prosody 2012*.

Mazaudon, M. and Michaud, A. (2008). Tonal contrasts and initial consonants: a case study of Tamang, a 'missing link' in tonogenesis. *Phonetica*, 65(4):231–256.

Mazo, M. (1995). Emotion and expression: Temporal data on voice quality in Russian lament. In *The 8th Vocal Fold Physiology Conf., Kurume, Japan, 1995*.

McClaskey, C. M. (2016). *Factors affecting relative pitch perception*. PhD thesis, UC Irvine.

Mende, W., Herzel, H., and Wermke, K. (1990). Bifurcations and chaos in newborn infant cries. *Physics Letters A*, 145(8-9):418–424.

Michaud, A. (2004). Final consonants and glottalization: new perspectives from Hanoi Vietnamese. *Phonetica*, 61(2-3):119–146.

Michel, J. F. (1968). Fundamental frequency investigation of vocal fry and harshness. *Journal of Speech and Hearing Research*, 11(3):590–594.

Mizuta, M., Abe, C., Taguchi, E., Takeue, T., Tamaki, H., and Haji, T. (2020). Validation of cepstral acoustic analysis for normal and pathological voice in the Japanese language. *Journal of Voice*.

Monsen, R. B. (1979). Acoustic qualities of phonation in young hearing-impaired children. *Journal of Speech, Language, and Hearing Research*, 22(2):270–288.

Moore, P. and Von Leden, H. (1958). Dynamic variations of the vibratory pattern in the normal larynx. *Folia Phoniatrica et Logopaedica*, 10(4):205–238.

Mooshammer, C. (2010). Acoustic and laryngographic measures of the laryngeal reflexes of linguistic prominence and vocal effort in german. *The Journal of the Acoustical Society of America*, 127(2):1047–1058.

Murphy, P. J. (2000). Spectral characterization of jitter, shimmer, and additive noise in synthetically generated voice signals. *The Journal of the Acoustical Society of America*, 107(2):978–988.

Murty, K. S. R. and Yegnanarayana, B. (2008). Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1602–1613.

Omori, K., Kojima, H., Kakani, R., Slavit, D. H., and Blaugrund, S. M. (1997). Acoustic characteristics of rough voice: Subharmonics. *Journal of voice*, 11(1):40–47.

Peirce, J. W. (2007). Psychopy—psychophysics software in python. *Journal of neuroscience methods*, 162(1-2):8–13.

Pickett, V. B., Villalobos, M. V., and Marlett, S. A. (2010). Isthmus (Juchitán) Zapotec. *Journal of the International Phonetic Association*, 40(3):365–372.

Pierce, L. and Silbiger, H. R. (1972). Use of shadowing in speech quality evaluation. *The Journal of the Acoustical Society of America*, 51(1A):121–121.

Pittam, J. (1987). Listeners' evaluations of voice quality in Australian english speakers. *Language and Speech*, 30(2):99–113.

Plack, C. J. and Oxenham, A. J. (2005). The psychophysics of pitch. In *Pitch*, pages 7–55. Springer.

Podesva, R. J. and Callier, P. (2015). Voice quality and identity. *Annual Review of Applied Linguistics*, 35:173–194.

Qi, Y., Hillman, R. E., and Milstein, C. (1999). The estimation of signal-to-noise ratio in continuous speech for disordered voices. *The Journal of the Acoustical Society of America*, 105(4):2532–2535.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Redi, L. and Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *Journal of Phonetics*, 29(4):407–429.

Robb, M. P. and Saxman, J. H. (1988). Acoustic observations in young children's non-cry vocalizations. *The Journal of the Acoustical Society of America*, 83(5):1876–1882.

Rothenberg, M. and Mahshie, J. J. (1988). Monitoring vocal fold abduction through vocal fold contact area. *Journal of Speech, Language, and Hearing Research*, 31(3):338–351.

Sakakibara, K.-I. (2003). Production mechanism of voice quality in singing. *Journal of the Phonetic Society of Japan*, 7(3):27–39.

Sands, S. and Lubera, A. (2017). Vowel quality and sound context of vocal fry. *NSURJ*, page 61.

Sapienza, C. M., Stathopoulos, E. T., and Dromey, C. (1991). Changes in open and speed quotient values as a function of measurement criteria. *The Journal of the Acoustical Society of America*, 89(4B):1935–1935.

Sapienza, C. M., Stathopoulos, E. T., and Dromey, C. (1998). Approximations of open quotient and speed quotient from glottal airflow and EGG waveforms: effects of measurement criteria and sound pressure level. *Journal of Voice*, 12(1):31–43.

Sataloff, R. T., Heman-Ackah, Y. D., Simpson, L. L., Park, J.-B., Zwislewski, A., Sokolow, C., and Mandel, S. (2002). Botulinum toxin type b for treatment of spasmodic dysphonia: a case report. *Journal of Voice*, 16(3):422–424.

Schreibweiss-Merin, D. and Terrio, L. M. (1986). Acoustic analysis of diplophonia: A case study. *Perceptual and motor skills*, 63(2):755–765.

Seyfarth, S. and Garellek, M. (2018). Plosive voicing acoustics and voice quality in Yerevan Armenian. *Journal of Phonetics*, 71:425–450.

Shimizu, M. and Dantsuji, M. (2000). A new proposal of laryngeal features for the tonal system of Vietnamese. In *Sixth International Conference on Spoken Language Processing*.

Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (2011). Voicesauce: A program for voice analysis. In *Proceedings of the ICPhS XVII*, pages 1846–1849.

Silverman, D. (1997). Laryngeal complexity in Otomanguean vowels. *Phonology*, 14(2):235–261.

Sirviö, P. and Michelsson, K. (1976). Sound-spectrographic cry analysis of normal and abnormal newborn infants. *Folia phoniatrica et logopaedica*, 28(3):161–173.

Sjölander, K. (2004). The snack sound toolkit [computer program]. *Retrieved February*.

Skuk, V. G. and Schweinberger, S. R. (2014). Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender. *Journal of Speech, Language, and Hearing Research*, 57(1):285–296.

Slavit, D. H. and McCaffrey, T. V. (1995). Open slope quotient: a new glottographic parameter. *Journal of Voice*, 9(1):86–94.

Slifka, J. (2006). Some physiological correlates to regular and irregular phonation at the end of an utterance. *Journal of Voice*, 20(2):171–186.

Sun, X. (2002). Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In *2002 IEEE international conference on acoustics, speech, and signal processing*, volume 1, pages I–333. IEEE.

Sun, X. and Xu, Y. (2002). Perceived pitch of synthesized voice with alternate cycles. *Journal of Voice*, 16(4):443–459.

Švec, J. G., Schutte, H. K., and Miller, D. G. (1996). A subharmonic vibratory pattern in normal vocal folds. *Journal of Speech, Language, and Hearing Research*, 39(1):135–143.

Szkiełkowska, A., Krasnodębska, P., Miaśkiewicz, B., and Skarżyński, H. (2018). Electroglottography in the diagnosis of functional dysphonia. *European Archives of Oto-rhino-laryngology*, 275(10):2523–2528.

Tehrani, H. (2009). EGGWorks: A program for automated analysis of EGG signals. *Cited on*, page 171.

Tigges, M., Mergell, P., Herzel, H., Wittenberg, T., and Eysholdt, U. (1997). Observation and modelling of glottal biphonation. *Acta Acustica united with Acustica*, 83(4):707–714.

Tilsen, S., Burgess, D., and Lantz, E. (2013). Imitation of intonational gestures: a preliminary report. *Cornell Work. Pap. Phon. Phonol*, pages 1–17.

Timcke, R., von Leden, H., and Moore, P. (1959). Laryngeal vibrations: Measurements of the glottic wave: Part II—Physiologic variations. *AMA archives of otolaryngology*, 69(4):438–444.

Titze, I. R. (1994). Fluctuations and perturbations in vocal output. *Principles of voice production*, pages 209–306.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Van der Woerd, B., Wu, M., Parsa, V., Doyle, P. C., and Fung, K. (2020). Evaluation of acoustic analyses of voice in nonoptimized conditions. *Journal of Speech, Language, and Hearing Research*, 63(12):3991–3999.

van Mersbergen, M., Lyons, P., and Riegler, D. (2017). Vocal responses in heighted states of arousal. *Journal of Voice*, 31(1):127.e13–127.e19.

Wang, F. (2015). Variations of laryngeal features in Jianchuan Bai. *Journal of Chinese Linguistics*, 43(1):434–452.

Ward, P. H., Sanders, J. W., Goldman, R., and Moore, G. P. (1969). Lxvii diplophonia. *Annals of Otology, Rhinology & Laryngology*, 78(4):771–777.

Waters, S., Kanber, E., Lavan, N., Belyk, M., Carey, D., Cartei, V., Lally, C., Miquel, M., and McGettigan, C. (2021). Singers show enhanced performance and neural representation of vocal imitation. *Philosophical Transactions of the Royal Society B*, 376(1840):20200399.

Wayland, R. and Jongman, A. (2003). Acoustic correlates of breathy and clear vowels: The case of Khmer. *Journal of Phonetics*, 31(2):181–201.

Wendahl, R. (1962). A photophonelographic analysis of hoarse voice quality. In *Proceedings of the 4th International Congress of Phonetic Sciences. The Hague*, pages 307–10.

Wickham, H., Chang, W., and Wickham, M. H. (2016). Package 'ggplot2'. *Create elegant data visualisations using the grammar of graphics. Version*, 2(1):1–189.

Wightman, F. L. and Green, D. M. (1974). The perception of pitch: The pitch of a sound wave is closely related to its frequency or periodicity—but the exact nature of that relation remains a mystery. *American Scientist*, 62(2):208–215.

Wolfe, V. and Martin, D. (1997). Acoustic correlates of dysphonia: type and severity. *Journal of Communication Disorders*, 30(5):403–416.

Xu, A. and Lee, A. (2018). Perception of vocal attractiveness by Mandarin native listeners. In *Proceedings of the International Conference on Speech Prosody*, pages 344–348.

Xu, Y. (2011). Post-focus compression: Cross-linguistic distribution and historical origin. In *ICPhS*, volume 2011, pages 152–155.

Yiu, E. M., Murdoch, B., Hird, K., and Lau, P. (2002). Perception of synthesized voice quality in connected speech by Cantonese speakers. *The Journal of the Acoustical Society of America*, 112(3):1091–1101.

Yokonishi, H., Imagawa, H., Sakakibara, K.-I., Yamauchi, A., Nito, T., Yamasoba, T., and Tayama, N. (2016). Relationship of various open quotients with acoustic property, phonation types, fundamental frequency, and intensity. *Journal of Voice*, 30(2):145–157.

Yoshinaga, I. and Kong, J. (2012). Laryngeal vibratory behavior in traditional Noh singing. *Tsinghua Science and Technology*, 17(1):94–103.

Yost, W. A. (2009). Pitch perception. *Attention, Perception, & Psychophysics*, 71(8):1701–1715.

Yu, K. M. (2010). Laryngealization and features for Chinese tonal recognition. In *Eleventh Annual Conference of the International Speech Communication Association*.

Yuan, J. and Liberman, M. (2014). F0 declination in English and Mandarin broadcast news speech. *Speech Communication*, 65:67–74.

Yuasa, I. P. (2010). Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech*, 85(3):315–337.

Yumoto, E., Gould, W. J., and Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *The Journal of the Acoustical Society of America*, 71(6):1544–1550.

Zhang, Z. (2018). Vocal instabilities in a three-dimensional body-cover phonation model. *The Journal of the Acoustical Society of America*, 144(3):1216–1230.

Zheng, X. (2006). Voice quality variation with tone and focus in Mandarin. In *Tonal Aspects of Languages*.