The spectral dynamics of voiceless sibilant fricatives in English and Japanese

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree
Doctor of Philosophy
in the Graduate School of The Ohio State University

By
Patrick F. Reidy

Graduate Program in Linguistics
The Ohio State University
2015

Dissertation Committee:
Professor Mary E. Beckman, Advisor
Professor Micha Elsner
Professor Eric Fosler-Lussier
Professor Eric Healy

Abstract

Voiceless sibilant fricatives, such as the consonant sounds at the beginning of the English words *sea* and *she*, are articulated by forming a narrow constriction between the tongue and the palate, which directs a turbulent jet of air toward the incisors downstream. Thus, the production of these sounds involves the movement of a number of articulators, including the tongue, jaw, and lips; however, the principal method for analyzing the acoustics of sibilant fricatives has been to extract a single "steady state" interval from near its temporal midpoint, and estimate spectral properties of this interval. Consequently, temporal variation in the spectral properties of sibilant fricatives has not been systematically studied.

This dissertation investigated the temporal variation of a single spectral property, referred to as *peak ERB$_N$-number*, that denotes the most prominent psychoacoustic frequency. The dynamic aspects of peak ERB$_N$-number trajectories were analyzed with fifth-order polynomial time growth curve models. A series of analyses revealed a number of novel findings.

First, a comparison of English- and Japanese-speaking adults indicated that the language-internal contrast between English /s/ and /ʃ/ and between Japanese /s/ and /ɕ/ is indicated by dynamic properties of peak ERB$_N$-number across the time course of the sibilants. Furthermore, a cross-linguistic comparison of the adults' productions of English /s/ and Japanese /s/ indicated that the peak ERB$_N$-number of a sibilant follows a language-specific trajectory.

Next, the development of the sibilant fricative contrast in native English- and Japanese-acquiring children, between the ages of two and five years, was investigated in terms of peak ERB$_N$-number trajectory. The English-acquiring children were found to contrast /s/ and /ʃ/ in terms of similar aspects of peak ERB$_N$-number trajectory as the English-speaking adults did—namely, average peak ERB_N-number and the concavity of its trajectory. Moreover, the extent to which the children differentiated the consonants in terms of these properties increased with age. The analysis of the Japanese-acquiring children was complicated by an apparent developmental regression in the five-year-olds.

Third, effects of vowel context on peak ERB_N-number trajectory were examined in the English-speaking adults' and the English-acquiring children's productions. Both age groups exhibited effects of vowel frontness on both sibilants: a following front vowel raised peak ERB_N-number across the full duration of these sibilants. Similarly, both age groups exhibited effects of vowel rounding on /s/: a following rounded vowel lowered peak ERB_N-number across the time course of /s/. The adults exhibited effects of vowel height on the dynamic aspects of peak ERB_N-number trajectory for both sibilants. Such effects were not found in the children. Furthermore, the extent of the children's vowel-context effects differentially weakened (vowel rounding and frontness) and strengthened (vowel height) developmentally.

Finally, the development of sibilant contrast was investigated in a cohort of cochlear-implanted English-acquiring children. The sibilant contrast, in terms of average peak ERB_N-number and curvature of its trajectory, tended to improve with prolonged hearing experience, but no association was found with age at implantation. The children with cochlear implants were compared to children with normal hearing, who were matched on hearing age. The implanted children seemed to differentiate the onset and offset of frication in a more adultlike fashion than did the children with normal hearing. Furthermore, no group-related differences in the extent of sibilant contrast were found, suggesting that the children with cochlear implants differentiated the sibilants as well as the children with normal hearing.

Acknowledgements

It is an odd thing, reading just one name on the title page of this dissertation, knowing, as I do, that if I have done anything here, it has been with the support of a great number of people.

I would first like to thank my advisor, Dr. Mary Beckman, for her patience, encouragement, and guidance during my tenure as a graduate student at Ohio State and during the completion of my dissertation research. I have learned so much from her that attempting to enumerate is hopeless.

I would also like to thank my committee members Dr. Micha Elsner and Dr. Eric Fosler-Lussier, for discussions of the statistical models used herein; and Dr. Eric Healy, for encouraging me to think about speech production data in terms of what is detected by the auditory system. My development as a researcher has also benefited from the mentorship of Dr. Jan Edwards and Dr. Ben Munson; together with Dr. Beckman, they have embodied the ethic that research is a social activity, and that nothing worth doing is done alone.

I am greatly indebted to, and sincerely thank, all the experimenters and researchers who helped prepare, collect, and annotate the speech production data that were analyzed in this dissertation: Laura Slocum and Anne Hoffmann, for recruiting and recording the English-speaking adults and normal-hearing children; Sarah Schellinger and Emilie Sweet, for transcribing the productions from the English-speaking adults and normal-hearing children; Kiyoko Yoneyama, Junko Davis, and Yuki Sunawatari, for recruiting and recording the Japanese-speaking adults and normal-hearing children; Junko Davis and Hiroko Morioka, for transcribing the productions from Japanese-speaking adults and normal-hearing children; Fangfang Li, Chanelle Mays, Oxana Skorniakova, Asimina Syrika, Julie Johnson, and Jeff Holliday, for tagging the fricative events in the English and Japanese portions of the

Paidologos corpus; Ann Todd and Emilie Sweet, for recruiting and recording the cochlear-implanted children; and Rebecca Hatch, for transcribing and tagging the fricative events for the productions from the cochlear-implanted children.

During my time as a graduate student at Ohio State, I benefitted immeasurably from knowing Abby Walker, Andrew Plummer, and Jeff Holliday, whose friendship, humor, and curiosity were at times my best reason to go into Oxley.

Lastly, I thank my wife, Callan, for her unflagging love and friendship. It has been more than I could have asked for, and more than I deserve.

This research was supported by NIH Grant R01DC02932 to Jan Edwards and Mary E. Beckman; NIH Grant R01DC008365 to Ruth Y. Litovsky; an Ohio State Univeristy Center for Cognitive Science seed grant to Mickhail Belkin, Mary E. Beckman, and Eric Fosler-Lussier; and a Raymond H. Stetson Scholarship (Acoustical Society of America) to myself.

Vita

April, 1985	Born in Houston, Texas
May, 2007	B.A. with High Honors in Linguistics, University of Texas at Austin
April, 2012	Married C. C. Searcy in Austin, Texas

Publications

Reidy, P. F., Beckman, M. E., Litovsky, R. Y., and Edwards, J. 2015. The acquisition of English sibilant fricatives by children with bilateral cochlear implants. To appear in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS).*

Nicholson, H., Munson, B., Reidy, P. F., and Edwards, J. 2015. Effects of age and vocabulary size on production accuracy and acoustic differentiation of young children's sibilant fricatives. To appear in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS).*

Holliday, J. J., Reidy, P. F., Beckman, M. E., and Edwards, J. 2015. Quantifying the robustness of the English sibilant fricative contrast in children. In press *Journal of Speech, Language, & Hearing Research.*

Reidy, P. F. and Beckman, M. E. 2015. Vowel context effects on the spectral dynamics of English and Japanese sibilant fricatives. *Journal of the Acoustical Society of America*, 137(4):2381.

Reidy, P. F. 2015. A comparison of spectral estimation methods for the analysis of sibilant fricatives. *Journal of the Acoustical Society of America–Express Letters*, 137(4):EL248–EL254.

Reidy, P. F. 2014. Moving targets and unsteady states: "Shifting" productions of sibilant fricatives by young children. *Journal of the Acoustical Society of America*, 136(4):2262.

Reidy, P. F. and Beckman, M. E. 2014. Steady as /ʃi/ goes: The kinematics of sibilant fricatives in English and Japanese. *Journal of the Acoustical Society of America*, 135(4):2355.

Reidy, P. F. 2013. The (null) effect of spectral estimator on the estimation of spectral moments. *Journal of the Acoustical Society of America*, 134(5):4328.

Reidy, P. F. 2013. An introduction to random processes for the spectral analysis of speech data. In Beckman, M. E., Lesho, M., Tonhauser, J., and Tsui, T.-H. (Eds.), *Ohio State Working Papers in Linguistics* (No. 60), 67–116.

Reidy, P. F. and Beckman, M. E. 2012. The effect of spectral estimator on common spectral measures for sibilant fricatives. *Proceedings of InterSpeech-2012*, 1516–1519.

Fields of Study

Major field: Linguistics

Minor fields: Acoustic phonetics, language acquisition, speech production

Table of Contents

Abstract . . . ii

Acknowledgements . . . iv

Vita . . . vi

List of Figures . . . xi

List of Tables . . . xiii

1 Introduction . . . 1
- 1.1 Articulatory kinematics, but spectral statics? . . . 2
- 1.2 Phonological development . . . 5
- 1.3 Acoustic versus electrical hearing . . . 7
- 1.4 Organization of the dissertation . . . 9

2 Spectral estimation . . . 12
- 2.1 Spectral representation in a statistical setting . . . 12
- 2.2 Spectral estimation . . . 13
  - 2.2.1 The periodogram and its discontents . . . 15
  - 2.2.2 Methods for reducing the variance of the periodogram . . . 17
- 2.3 The auditory model . . . 21
  - 2.3.1 Two perspectives of the auditory system . . . 21
  - 2.3.2 A computational model of the auditory system . . . 26
  - 2.3.3 The auditory model as a series of smoothing windows . . . 29
  - 2.3.4 Spectral dynamics from multiple ‘glimpses’ . . . 29
- 2.4 Conclusion . . . 30

3 The spectral dynamics of sibilant contrast . . . 31
- 3.1 Introduction . . . 31
  - 3.1.1 Articulatory postures of sibilant fricatives . . . 31
  - 3.1.2 Static spectral properties of sibilant fricatives . . . 34
  - 3.1.3 Articulatory kinematics and spectral dynamics of English /s/ . . . 36
  - 3.1.4 Purpose and hypotheses . . . 37
- 3.2 Method . . . 39
  - 3.2.1 Participants . . . 39
  - 3.2.2 Materials . . . 39

3.2.3 Elicitation and recording procedure 41
3.2.4 Annotation of fricative events 42
3.2.5 Peak ERB$_N$-number trajectories 43
3.2.6 Growth-curve analysis of peak ERB$_N$-number trajectories 44
3.3 Results 45
3.3.1 Contrast between English /s/ and /ʃ/ 45
3.3.2 Contrast between Japanese /s/ and /ɕ/ 49
3.3.3 Cross-linguistic comparison of English and Japanese /s/ 53
3.4 Discussion 56
3.4.1 Potential artifacts of time normalization 57
3.4.2 Articulatory kinematics 58
3.4.3 Language specificity of gestural coproduction 61

4 The development of contrastive spectral dynamics in sibilant fricatives 63
4.1 Introduction 63
4.1.1 Ages of acquisition of intelligible sibilants, cross-linguistically 63
4.1.2 Language-specific differentiation and tuning of sibilant fricatives 64
4.1.3 Purpose and hypotheses 67
4.2 Method 69
4.2.1 Participants 69
4.2.2 Materials 71
4.2.3 Elicitation and recording procedure 71
4.2.4 Segmentation, transcription, and annotation of target productions 71
4.2.5 Peak ERB$_N$-number trajectories 74
4.2.6 Community-norm consonant-contrast models 74
4.3 Results 75
4.3.1 The development of sibilant contrast in English-acquiring children 75
4.3.2 The development of sibilant contrast in Japanese-acquiring children 81
4.4 Discussion 86
4.4.1 English-acquiring children 86
4.4.2 Japanese-acquiring children 88

5 Vowel-context effects on spectral dynamics 91
5.1 Introduction 91
5.1.1 Coarticulatory effects on sibilants 92
5.1.2 Acoustic vowel-context effects 95
5.1.3 Contextual effects as a window into phonological organization 96
5.1.4 Purpose and hypotheses 97
5.2 Method 98
5.2.1 Vowel-context models 99
5.3 Results 100
5.3.1 Community-norm vowel-context effects on English /s/ 100
5.3.2 The development of vowel-context effects on English /s/ 102
5.3.3 Community-norm vowel-context effects on English /ʃ/ 106

5.3.4 The development of vowel-context effects on English /ʃ/ 108
5.4 Discussion 110

6 Sibilant contrast by children with cochlear implants 114
6.1 Introduction 114
6.1.1 Sibilant production deficits in pediatric CI users 114
6.1.2 Purpose of analyses 116
6.2 Method 117
6.2.1 Participants 117
6.2.2 Materials 119
6.2.3 Elicitation and recording procedure 119
6.2.4 Transcription and annotation of fricative events 119
6.2.5 Normal-hearing age matches 120
6.3 Results 121
6.3.1 Hearing experience and sibilant contrast in pediatric CI users 121
6.3.2 Comparison with hearing-age matched normal-hearing peers 123
6.4 Discussion 127
6.4.1 The development of sibilant differentiation in pediatric CI users . . . 127
6.4.2 Group differences in peak-ERB$_N$ trajectory shape, but not consonant contrast 129

7 Conclusion 132
7.1 Summary and implications of empirical findings 132
7.1.1 Language-specificity in sibilant productions 132
7.1.2 Development of sibilant differentiation 135
7.1.3 Development of sibilant-vowel coproduction 137
7.2 Limitations of the study 139
7.2.1 Prosodic position of the elicited sibilant fricatives 140
7.2.2 Subdivision of a sibilant into closure, maintenance, and opening movements 140

References 142

List of Figures

2.1 A realization of the waveform of /s/ 14

2.2 A periodogram estimate of the spectrum of /s/ 17

2.3 Discrete prolate spheroidal sequences of orders $k = 0$ through $k = 7$ 19

2.4 A multitaper spectral estimate of /s/ 20

2.5 Frequency responses of fourth-order gammatone filters at different center frequencies 25

2.6 The excitation pattern output by the auditory model in response to /s/ . . 28

3.1 Orthogonal polynomial functions used in growth curve models 45

3.2 Predicted peak ERB$_N$-number trajectories for the English-speaking adults' consonant-contrast model 48

3.3 Predicted peak ERB$_N$-number trajectories for the Japanese-speaking adults' consonant-contrast model 52

3.4 Predicted peak ERB$_N$-number trajectories for the cross-linguistic comparison model for /s/ 56

4.1 The English-acquiring children's individual consonant effects plotted against age 78

4.2 Fixed-effects predictions of the consonant-contrast models fitted to the productions of the English-acquiring children. 79

4.3 The Japanese-acquiring children's individual consonant effects, plotted against age 84

4.4 Fixed-effects predictions of the consonant-contrast models fitted to the productions of the Japanese-acquiring children. 85

5.1 Predicted peak ERB$_N$-number trajectories for the English-speaking adults' vowel-context models. 102

5.2 The English-acquiring children's individual rounding and frontness effects on the intercept, for /s/, plotted against age 103

5.3 The English-acquiring children's individual height effects on cubic and quintic time, for /s/, plotted against age 105

5.4 The English-acquiring children's individual vowel effects for /ʃ/, plotted against age 109

6.1 The implant histories of the children with bilateral cochlear implants, who participated in the repetition task 118

6.2 Predicted peak ERB$_N$-number trajectories from the English-acquiring pediatric CI users' consonant-contrast model 123

6.3 The pediatric CI users' individual consonant effects 124

6.4 Peak ERB$_N$-number trajectories for the pediatric cochlear implant users and their hearing-age matched peers, predicted by the fixed effects of the fitted consonant-contrast model. Data means are plotted as points. 127

List of Tables

3.1 The /s/- and /ʃ/-initial target words used in the English word-repetition task 40

3.2 The /s/- and /ɕ/-initial target words used in the Japanese word-repetition task 41

3.3 Fixed-effects structure of the consonant-contrast model fitted to the adult English speakers' sibilant productions 46

3.4 Fixed-effects coefficients of the consonant-contrast model fitted to the adult English speakers' sibilant productions 47

3.5 Fixed-effects structure of the consonant-contrast model fitted to the adult Japanese speakers' sibilant productions 50

3.6 Fixed-effects coefficients of the consonant-contrast model fitted to the adult Japanese speakers' sibilant productions 51

3.7 Fixed-effects structure of the cross-linguistic comparison model fitted to the adult English and Japanese speakers' productions of /s/ 54

3.8 Fixed-effects coefficients of the cross-linguistic comparison model fitted to the adult English and Japanese speakers' productions of /s/ 55

4.1 Demographic summary of the English-acquiring children 70

4.2 Demographic summary of the Japanese-acquiring children 71

4.3 The distribution of analyzable productions from the English-acquiring children 73

4.4 The distribution of analyzable productions from the Japanese-acquiring children 74

4.5 Fixed-effects coefficients for the consonant-contrast models fitted to the English-acquiring children's sibilant productions 76

4.6 Fixed-effects coefficients for the consonant-contrast models fitted to the Japanese-acquiring children's sibilant productions 82

5.1 Fixed effects structure of the vowel-context model for English-speaking adults' peak ERB$_N$-number trajectories for /s/. 100

5.2 Fixed-effects coefficients of the vowel-context model fitted to the English-speaking adults' productions of /s/ 101

5.3 Fixed effects structure of the vowel-context model for English-speaking adults' peak ERB$_N$-number trajectories for /ʃ/. 106

5.4 Fixed-effects coefficients of the vowel-context model fitted to the English-speaking adults' productions of /ʃ/ 107

6.1 Fixed-effects coefficients of the consonant-contrast model fitted to the sibilant productions of the English-acquiring children who have cochlear implants . 122

6.2 Fixed-effects structure of the consonant-contrast model fitted to the productions of the pediatric cochlear implant users and their hearing-age matched peers with normal hearing 125

6.3 Fixed-effects coefficients of the hearing-status model comparing the sibilant productions of English-acquiring children with normal hearing and those with cochlear implants 126

Chapter 1

Introduction

Voiceless sibilant fricatives denote a class of speech sounds that includes English /s/ and /ʃ/, which allow speakers to minimally differentiate word pairs like *sue–shoe* (/su/–/ʃu/); and Japanese /s/ and /ɕ/, which minimally differentiate the words *sumi* ‘charcoal’ (/sumi/) and *shumi* ‘hobby’ (/ɕumi/). These sounds are articulated by forming a narrow constriction, bounded by the tongue and the palate, in the oral cavity. Noise is generated when air flowing through the constriction becomes turbulent and impinges on the upper and lower incisors downstream. These noise source mechanisms differentiate the voiceless sibilants from voiced sibilants, such as English /z/ (*zoo*), which also involve vocal fold vibration; from stops, such as English /t/ (*two*), and affricates, such as Japanese /ts/ (*tsumi* ‘sin’), which involve the build up and sudden release of air pressure due to an occlusion of the vocal tract; and weak fricatives, such as Japanese /ɸ/ (*fumi* ‘letter’), in which the turbulent airflow exiting the oral constriction does not collide with a downstream obstacle. Similarly, This dissertation presents methods for representing the time-varying spectral patterns of voiceless sibilant fricatives, which indicate the different patterns of coordination among articulatory movements and postural targets involved in the production of these complex sounds. Additionally, the development, in children, of the differentiation of articulatory gestures for sibilants, and of the coordination of these gestures with those of a following vowel, is investigated in terms of the dynamic spectral properties of these consonants.

1.1 Articulatory kinematics, but spectral statics?

Articulatory studies of children and adults have found that their principal speech organs undergo continual motion during the production of sibilant fricatives. In particular, this is true of the tongue dorsum, body, and blade for adults and preadolescents (Zharkova et al., 2014). Additionally, the height of the jaw, lower lip, and tongue tip have been observed to vary with time in adults productions of /s/ (Iskarous et al., 2011). Finally, both horizontal and vertical movement of the tongue tip and body have been recorded in children as young as five years old (Katz and Bharadwaj, 2001).

Despite this overwhelming evidence for variation in the position of the articulators over the course of the production, temporal variation in the acoustic features of sibilant fricatives has received very little attention in the literature. Indeed, the quantal theory of speech (Stevens, 1972, 1989) submits that acoustic properties of a speech sound are not likely to vary much across its duration. This theory stems from the realization that the mapping between articulator position and acoustic output is nonlinear; thus, there will be regions in the articulatory space where a variation will yield only an insignificant acoustic change, and there will be other articulatory regions where a comparable variation will yield sizable acoustic changes. The quantal theory posits: first, that the sounds of a language tend to involve articulatory regions where variation is not likely to affect the output acoustics, ensuring that during production the inevitable variation in the articulators' positions over time will not engender noticeable acoustic differences; and second, that two contrastive sounds of a language will involve two such regions of stability that are separated by an unstable region of articulatory space, whereby crossing that unstable region yields large differences in the acoustics of the two sounds.

Perkell et al. (1979, 2004) argue that English /s/ and /ʃ/ stand in such a quantal relationship by virtue of differential contact between the underside of the tongue and the lower incisors. Specifically, whereas /s/ is articulated with the tongue blade postured to make a continuous floor for a cavity between the constriction and the teeth, /ʃ/ is articulated in such a way that the supralingual front cavity floor is broken by the opening into a sublingual cavity of air that separates the tongue from the lower incisors. Due to the presence of this sublingual cavity, the total size of the cavity in front of the constriction is much greater for /ʃ/ than for /s/, causing the resonances of the former to occur at much lower frequencies.

Behrens and Blumstein (1988) analyzed temporal changes, across the duration of /s/ and /ʃ/, in the frequency of the most prominent spectral peak, which represents the resonant frequency of the front cavity. Peak frequency values were computed from the beginning, middle, and final 15 ms of each fricative, revealing that the peak frequency was higher for /s/ (3.8–8.5 kHz) than for /ʃ/ (2.3–7 kHz). Furthermore, the authors noted that the spectral "patterns appeared to be maintained across the three time windows," which led them to suggest that any characterization of /s/ and /ʃ/ "based on spectral properties can probably be derived from ... a static configuration of the frication noise itself ... irrespective of where the frication noise is measured" (pp. 297–298).

Here, care is needed in interpreting exactly what is maintained across the duration of the fricatives. One interpretation is that the relative distribution of /s/ and /ʃ/'s peak frequency is the same regardless of where in the time course of the sibilant it is computed. Under this interpretation, a single measure of peak frequency is sufficient to characterize the difference between /s/ and /ʃ/ because the relative peak frequency is higher for /s/ than for /ʃ/ at any interval. A stronger interpretation is that the peak frequency of each sibilant is reasonably constant across time. Under this second interpretation, a single measure of peak frequency would be sufficient for characterizing both /s/ and /ʃ/, in their own right, since this spectral property would not be sensitive to where in the fricative it is measured.

While Behrens and Blumstein did not report any statistical tests that would have indicated the extent to which peak frequency varies across the time course of a sibilant, they did report that "high frequency peaks tended to appear more often at the midpoint" of frication (p. 297), which suggests some amount of variation with time, supporting the weaker interpretation, under which a static measure is sufficient for describing only the contrast between sibilants. However, it is the stronger interpretation that seems to have persisted. For example, Behrens and Blumstein (1988) is cited as the basis of the following claims: spectral properties of sibilants "are relatively stable throughout the noise portion" (Jongman et al., 2000, p. 1255); "previous research has not found that the spectral [peak] varies greatly throughout the course of the fricative" (Munson, 2001, p. 1203); spectral "peak measures [remain] relatively constant across time" (Newman et al., 2001, p. 1184).

A recent study of the spectral dynamics of English /s/, however, provides strong evidence that only the weak interpretation of Behrens and Blumstein (1988) should be followed. Iskarous et al. (2011) analyzed adults' productions of /s/, from a speech corpus that had been recorded using x-ray microbeams to track simultaneously the movements of certain articulators like the jaw, tongue tip, and tongue blade. They found that the spectral mean frequency followed an increasing, concave trajectory across the course of the fricative, rising until reaching a global maximum around 80% of the fricative's duration, before falling off. Moreover, this temporal variation in spectral mean frequency seemed to correspond to articulatory movements such as the raising of the jaw across the first half of frication, and the release of the linguapalatal constriction near the end of frication.

The results of Iskarous et al. (2011) suggest that a static measure of spectral mean or peak frequency is insufficient to characterize the spectral properties of /s/ since these vary across the course of the fricative. From the observations of Behrens and Blumstein (1988) and Iskarous et al. (2011), one might hypothesize tha the spectral mean frequency trajectory of /ʃ/ is comparable to that of /s/, just at a lower frequency; i.e., that in a two-dimensional frequency-vs-time space, the spectral mean trajectory of /ʃ/ is a translation of that of /s/. One study presented anecdotal, graphical evidence of differences in the dynamics of the spectral properties of /s/ and /ʃ/, suggesting that "[s]uch dynamic changes may have substantial impact on the evaluation of the [/s/–/ʃ/] distinction and should be considered in future research" (Haley et al., 2010, p. 553). However, Haley and her colleagues did not test whether such apparent differences in the dynamics of the two sibilants' spectral properties were statistically significant. The central purpose of this dissertation is to investigate the differences in the spectral dynamics of adults' productions of sibilant fricatives and then use these dynamic measures to deepen the current understanding of the development of sibilant fricative categories and the sibilant fricative contrast in different populations of children.

1.2 Phonological development

The acquisition of speech sounds involves, at the very least, two associations among different time-varying patterns. First, during babbling, an infant acquires the cognitive structures that allow the infant's own articulatory patterns to be associated to the acoustic signals that they generate (Bailley et al., 1991; Jordan, 1990; Plummer, 2014). From this association develops a limited repertoire of 'vocal motor schemes' that are used to bootstrap initial phonological representations for the infant's first words (McCune and Vihman, 1987; Vihman and Keren-Portnoy, 2011). Second, a child learns to associate the acoustic signals of their own speech with those of their caretaker. Plummer (2014) provides a computational model of how this association might be acquired for vowels, through imitative interaction between infant and caretaker. Importantly, Plummer showed that through learning this association the category system that organizes the caretaker's vowel space may be transferred to the child, who in turn reorganizes an internalized acoustic vowel space to modify previously learned associations between articulator movements and their acoustic consequences. In this way, the articulatory gestures that engender relevant acoustic contrasts in the ambient language become differentiated during phonological development (cf. Browman and Goldstein, 1989).

The acquisition of voiceless sibilants is often delayed relative to other sounds in a language. For example, at age five, English-acquiring children have been found to produce

/s/ and /ʃ/ intelligibly less than 85% of the time, but at this same age they produce the voiceless stops /p/, /t/, and /k/ and the weak fricative /f/ with near perfect intelligibility (Smit et al., 1990). Furthermore, children with articulation disorders have been found to produce both /s/ and /ʃ/ as a single "undifferentiated lingual gesture" (Gibbon, 1999), and it has been suggested that typically developing children may also progress through a stage of undifferentiated production as they acquire sibilants (Li, 2012; Li et al., 2009).

The principal method for studying the acoustic differentiation of sibilant fricatives by children has been to compute a limited number of spectral properties from one or two locations in the frication noise: first, at or near the temporal midpoint of frication; second, at the boundary of the sibilant and the following sound. The first of these locations has been argued to best reflect the target acoustics, once removed from transitional effects. For example, Li (2012, p. 1306) states that the middle 40 ms of a sibilant "is the steadiest portion of the fricative noise and is least likely to be influenced by amplitude build-up at the beginning of the fricative or the transitional change into the following vowel." Similarly, Romeo et al. (2013, p. 3783) posit that measuring spectral mean frequency across the middle half of the frication noise "avoids the effects of fricative onset variation and subsequent vowel formant transitions." But the findings of Iskarous et al. (2011) strongly suggest that such transitional effects at the onset of frication and at a fricative-vowel boundary are due, respectively, to mandibular and lingual gestures—gestures that a child must learn to execute with adult-like proficiency. Thus, limiting the scope of a study to just the supposed "steady-state" interval would seem to be limiting the understanding of the development of children's ability to articulate and differentiate sibilant fricatives in an adult-like manner.

The second location from which spectral properties are computed, the offset of the frication noise, indexes the transition of the sibilant into the following sound, typically a vowel. At this location, the spectral property typically computed is one that reflects the resonances of the cavity behind the constriction, such as $F2$ frequency, which become apparent in the spectrum of the frication once the constriction is released (McGowan and

Nittrouer, 1988; Soli, 1981). Spectral properties like $F2$ frequency at the end of frication index a second aspect of phonological development other than the differentiation of sibilant gestures: the coordination of a sibilant gesture with the gesture of the following vowel.[1]

Nittrouer et al. (1989) argue that as sibilant and vowel gestures become more coordinated, they overlap to a lesser extent. As evidence they show that the vowel-context effects on $F2$ are greater in children than in adults. Recently, using high frame-rate ultrasound imaging, Zharkova et al. (2014) found that in adults' productions of sibilant-vowel syllables, lingual coarticulatory effects became apparent in the first 10 ms of the frication. At such an early point, the linguapalatal constriction would be expected to be tight enough to cancel the resonances of the back cavity. Hence, the dynamic patterns of a measure, like spectral mean or peak frequency, that indexes the resonance of the front cavity could give a much more detailed picture of the vowel-context effects on the acoustics of sibilant fricatives, and how these context effects vary in children across development.

1.3 Acoustic versus electrical hearing

The *tonotopic organization*, or *tonotopic mapping*, of the auditory system refers to the order-preserving mapping that relates the physical frequency scale (of pure tones) to the spatial organization of each structure in the auditory system, respectively. That is, the various structures in the auditory system—e.g., the basilar membrane, the auditory nerve, the cochlear nucleus, and the auditory cortex—are respectively organized in such a way that a fixed location on any of these structures—e.g., a narrow cross-section of the basilar membrane or a fiber in the auditory nerve—responds most sensitively to a narrow range of frequencies, and distinct locations in a given structure that respond to similar frequencies are located proximally to each other (Schreiner et al., 2000; von Békésy, 1960).

[1] Depending on the language, transitional information like $F2$ at fricative-vowel boundary may also index the differentiation of sibilant fricatives. For example, the Japanese sibilants /s/ and /ɕ/ differ in terms of the length of the linguapalatal constriction, which induces a difference in the size of the back cavity. This is discussed in greater detail in chapter 3.

In listeners with normal hearing, the peripheral auditory system acts as a bank of bandpass filters, parsing an incoming sound wave into its various frequency components. The intensity of each component is reflected in the vibrational pattern in the basilar membrane in the cochlea. When this membrane vibrates it innervates the auditory nerve, thus acting as a transducer of mechanical vibrational energy into neuronal electrical impulses.

In listeners who have received a cochlear implant (CI) to restore hearing after severe to profound sensorineural hearing loss, the auditory system is also tonotopically organized; however, the origin of this organization is not due to the biophysical properties of the basilar membrane, but instead to the design of the cochlear prosthetic's components. A CI includes an external sound processor and an array of electrodes that are surgically inserted into the patient's cochlea. Each electrode corresponds to a unique bandpass filter in the processor's filter bank, and each filter extracts the information from an incoming signal within a narrow frequency range, so that each electrode in the array is stimulated from a limited frequency range (Loizou, 2006).

Due to the design of contemporary cochlear prosthetics, the hearing of CI users differs from that of normal hearing listeners in two important ways that are relevant for the perception and production of sibilant fricatives. The first is that the range of audible frequencies is generally more restricted for CI users since the processor of commercially available CIs extends up to only 8 kHz (Loizou, 2006). This cutoff frequency may complicate sibilant perception in CI users since /s/ typically has its greatest concentration of energy in the 5–10 kHz range; thus, CI users may not be presented with the full frequency range relevant to discriminating sibilants.

The second is that, the frequency resolution of electrical hearing is typically poorer than that of acoustic hearing. In comparison to the frequency tuning of auditory nerves and nucleic nerve cells in a healthy auditory system, frequency tuning of nerve cells to the electrical stimulus provided by a CI has been shown to be much broader (Raggio and Schreiner, 2003; Middlebrooks et al., 2005). While the tuning of neuronal responses to electrical stimulation is affected by many listener-specific factors, such as the density of surviving neurons in particular regions of the cochlea or the distance between an electrode contact and its neural target (Kral et al., 1998), device-specific factors such as the electrode configuration have also been found to influence significantly the spread of neural activation in cochlear implant users (Bierer, 2002).

1.4 Organization of the dissertation

Because sibilant fricatives are produced by turbulence noise sources, spectral representatives of them that are based on conventional nonparametric spectral estimators, like the discrete Fourier transform, yield estimates that have a great amount of point-to-point variation; thus, some form of frequency-smoothing or averaging is necessary in order to yield a low-variance spectral representation. In chapter 2, a method of averaging is introduced that both preserves temporal resolution and intra-token variability, and its statistical properties are reviewed. Additionally, a bandpass filter bank model of the auditory system is described. It is argued that this auditory model can be construed as a sequence of smoothing windows, which when applied to the spectral estimate of a sibilant yields even further reductions in the point-to-point variation in the ordinate values, while doing so in a way that is motivated by the way that the auditory system transforms spectral information. This auditory model is then used to investigate the temporal variation in a measure, denoted peak ERB_N, of the most prominent psychoacoustic frequency across the course of sibilant fricatives.

Chapter 3 extends the acoustic analysis of Iskarous et al. (2011) by analyzing the peak ERB_N-number trajectories of English /s/ and /ʃ/ and Japanese /s/ and /ɕ/, as produced by native adult speakers of either language. These languages were chosen because they both have a binary sibilant fricative contrast, but this distinction is made in a language-specific way. In English, /s/ and /ʃ/ differ greatly in terms of front cavity size; thus, they are easily differentiable early on in the course of the frication noise. On the other hand, Japanese /s/

and /ç/ differ more so in terms of back cavity size than front cavity size, and so transitional information near the offset of frication helps differentiate these two sounds. Of interest is whether either language differentiates the two sibilants in terms of how peak ERB_N varies over time. Furthermore, English /s/ and Japanese /s/ are cross-linguistically assimilable sounds that share similar articulatory postures; thus, these sounds may be compared to determine whether the peak ERB_N trajectories are subject to language-general constraints.

In chapter 4, the development of normal-hearing English- and Japanese-acquiring children's ability to differentiate the sibilants of their ambient language in terms of the dynamic changes in peak ERB_N. Since previous research has found that the development of consonant differentiation is not uniform and is subject to tuning (cf. Nittrouer, 1995), the primary questions of interest are whether the children develop toward the same patterns of contrast as the adults in their language; and whether differentiation of sibilants in terms of dynamic versus global static properties of peak ERB_N-number trajectory occur at different ages.

Chapter 5 explores vowel-context effects on the peak ERB_N-number trajectories of English /s/ and /ʃ/, as produced by the adults and children who are reported in the previous chapters. The adults' data are analyzed in order to establish the community norms for the extent of vowel-context effects on either sibilant. Because the vowel-context effects are investigated on peak ERB_N-number trajectories, this analysis is able to determine which vowel-context effects modify dynamic properties of peak ERB_N, i.e. by changing the course of its trajectory, and which modify its global static properties, i.e. by shifting it to either a higher or lower frequency. These results are then used to explore vowel-context effects in children as an index of their development in the coordination of sibilant and vowel gestures. The primary question of interest is whether the children exhibited a developmental strengthening or weakening in vowel-context effects.

Finally, in chapter 6, the productions of sibilant fricatives by a cohort of bilaterally implanted pediatric CI users are analyzed. Previous work has found that children with CIs produce less acoustic contrast between /s/ and /ʃ/ than do children with normal hearing

(Todd et al., 2011; Uchanski and Geers, 2003); however, these studies considered only static spectral measures. The goal of this chapter is to compare the productions of the CI users with a cohort of hearing-age matched peers with normal hearing in terms of their spectral dynamic properties.

Chapter 2

Spectral estimation

2.1 Spectral representation in a statistical setting

The noise sources of sibilant fricatives arise from laminar airflow becoming turbulent as it passes through a constriction in the vocal tract. The laminar airflow exiting the lungs is channeled toward the constriction by a coronal groove formed along the tongue dorsum, behind the constriction (Stone et al., 1992). As the airstream travels through the constriction, asymmetries in the cross-sectional shape of the tongue cause random fluctuations in the airflow, and upon exiting the constriction, the flow becomes turbulent (Stevens, 2000). These random fluctuations in its noise source engender random fluctuations in the output acoustic waveform of a sibilant fricative.

Since each production of a sibilant fricative results in a waveform whose values are random, an appropriate setting in which to investigate the acoustic properties of these sounds is one where the acoustic waveform is modeled by a sequence of random variables $X_1, X_2, X_3, \ldots$. Such a sequence of random variables is referred to as a *stochastic process*, or simply *process*, and denoted either by a sequence $X_1, X_2, X_3, \ldots$ or an indexed collection $\{X_t\}$. In either case, the index t denotes the discrete, ordinal points in time at which the acoustic waveform is sampled by a digital recording device. The observed data values of a stochastic process are called a *realization* of the process and denoted either by $x_1, x_2, x_3, \ldots$ or $\{x_t\}$.

Often, speech researchers are primarily interested in spectral properties of sibilant fricatives, which is justified from both articulatory and auditory considerations. First, the distribution of peaks and troughs in the spectrum indicate the underlying configuration of the vocal tract during speech production (Fant, 1960). Second, the cochlea, a structure of the inner ear, acts as a spectral analyzer, factoring an incoming sound wave into its various frequency components (von Békésy, 1960).

Thus, speech researchers seek a representation of stochastic process $\{X_t\}$ as a sum of simple periodic functions that oscillate at different frequencies. Such a representation, though, is not guaranteed to exist for every stochastic process, but if $\{X_t\}$ is a stationary process, then its spectrum is assured to exist. A process is said to be *stationary* if it satisfies the following three conditions: first, $\mathbb{E}(X_t) = \mu$ for all t; second, $\mathrm{Var}(X_t) < \infty$ for all t; and third, $\mathrm{Cov}(X_s, X_t) = \mathrm{Cov}(X_{s+h}, X_{t+h})$ for all s, t, and h. A corollary to this definition is that if a process $\{X_t\}$ is stationary, then $\mathrm{Var}(X_t) = \sigma^2$ for all t. Thus, if a process is stationary, the expected value and variance of its variables are constant across time, and the covariance between two of its variables depends only on the amount of time that separates them.

If $\{X_t\}$ is a stationary process with an absolutely summable autocovariance function $\gamma_X(h) = \mathrm{Cov}(X_0, X_h)$, then its spectrum, or spectral density, f_X is the Fourier transform of its autocovariance function (Shumway and Stoffer, 2006, Property P4.1 and Theorem C.3):

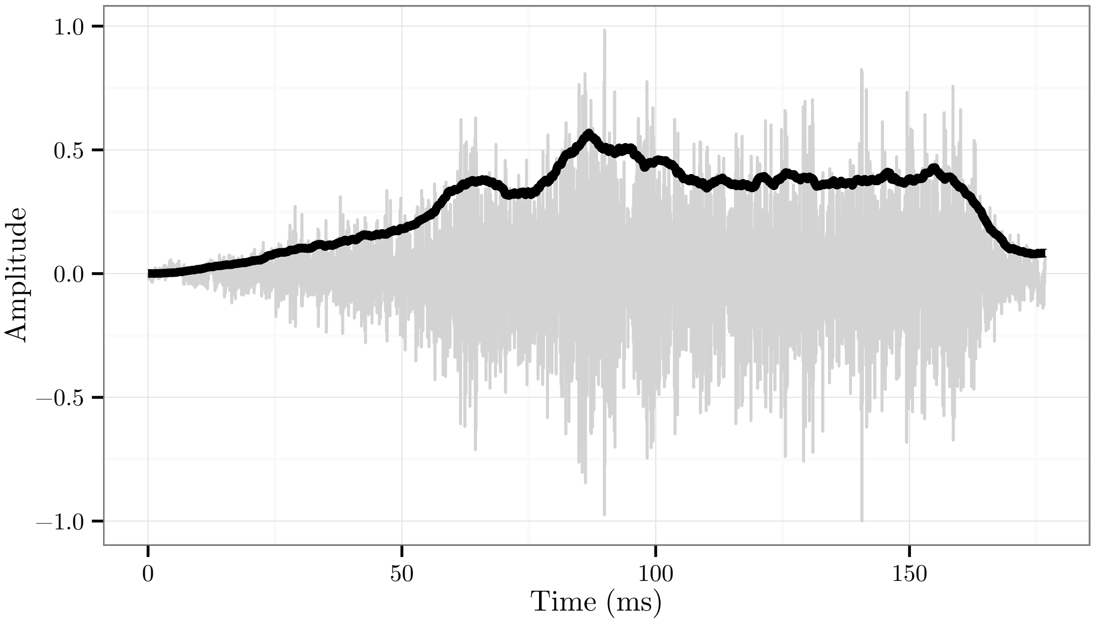
$$f_X(\omega) = \sum_{h=-\infty}^{\infty} \gamma_X(h) e^{-2\pi i \omega h}, \qquad -1/2 \leq \omega \leq 1/2. \tag{2.1}$$

The spectrum of $\{X_t\}$ denotes the distribution of variance over frequency, which corresponds, in physical terms, to the distribution of energy over frequency.

2.2 Spectral estimation

As a random quantity, the spectrum of a stochastic process cannot be accessed directly, but instead must be estimated from one or more realizations of it. A complicating factor for the spectral analysis of sibilant fricatives is that during speech production, the configuration

Figure 2.1: A realization of the waveform of /s/, extracted from an adult woman's production of the word 'soak.' The smoothed full-wave rectified amplitude envelope is shown in black.



of the articulatory system is not static, even across the duration of just a single phonetic segment. For example, the articulation of /s/ involves continuous movement of the tongue and jaw (Iskarous et al., 2011; Mooshammer et al., 2006, Fig. 1, p. 1031). In addition, the time course of neither the volume velocity of airflow nor the intraoral pressure is constant across the duration of a sibilant (Stevens, 2000). Due to these articulatory and aerodynamic changes during production, the acoustic waveform of a sibilant fricative is not stationary across its full duration, a fact which is illustrated by the realization of /s/ shown in Fig. 2.1. Specifically, the amplitude envelope of this realization $\{x_t\}$ suggests that the variance of the stochastic process $\{X_t\}$ varies considerably, especially near the onset and offset of frication.

To overcome the analytic challenges posed by the temporal dynamics inherent to speech data, most previous studies of sibilant fricatives have adopted the practice of estimating the spectrum of a sibilant waveform from a short interval, within which the waveform is assumed to be stationary. The duration of this interval varies across studies, from as short as 10 ms (e.g., Munson, 2004) to 40 ms or longer (e.g., Jongman et al., 2000; Romeo et al., 2013). In spite of their methodological differences, these studies have reported comparable spectral estimates for sibilants; thus, this dissertation assumes that, for operational purposes, a sibilant waveform remains stationary within any interval that is at most 40 ms.

Methods for estimating the spectrum of a process can be broadly classified into two groups: parametric and nonparametric. In the literature on speech, *linear predictive coding (LPC)* (Atal and Hanauer, 1971; Makhoul, 1975; Markel and Gray, 1976) is the most common parametric estimation method (e.g., Soli, 1981; Toda et al., 2002) used to analyze both vowels and consonants. An LPC spectrum is estimated by fitting to the waveform data an autoregressive model, whose coefficients are then used to compute a spectral estimate. A drawback of LPC is that it assumes that the vocal tract has no antiresonances, an assumption that does not hold for sibilant fricatives (Heinz and Stevens, 1961; Stevens, 1971). For this reason, a preponderance of studies of sibilants have opted for a nonparametric spectral estimator based on the discrete Fourier transform: the periodogram.

2.2.1 The periodogram and its discontents

If $x_1, \ldots, x_n$ are data realized from a stationary process, then the *periodogram* I_x of the data is defined as

$$I_x(\omega_j) = n^{-1} |d_x(\omega_j)|^2, \tag{2.2}$$

where d_x denotes the discrete Fourier transform[1] (DFT), and $\omega_j = j/n$ for $j = 0, \ldots, n-1$. The periodogram is, in familiar terms, the power spectrum of the data, scaled by the inverse of the number of data points observed. The presence of this scalar allows the periodogram

[1] The DFT of a finite sequence of numbers $x_1, \ldots, x_n$ is defined by $d_x(\omega_j) = \sum_{t=1}^{n} x_t e^{-2\pi i \omega_j t}$.

to be related to the sample autocovariance function[2] $\hat{\gamma}_x$ in the following way (Shumway and Stoffer, 2006, Eqn. 4.23, p. 188):

$$I_x = \sum_{|h|<n} \hat{\gamma}_x(h) e^{-2\pi i \omega_j h}, \qquad \omega_j \neq 0. \tag{2.3}$$

Thus, the periodogram's status as an estimator of the spectral density of a process is conferred by the fact that, for frequencies other than $\omega_0 = 0$, I_x and f_X are Fourier transforms of $\hat{\gamma}_x$ and γ_X, respectively. Furthermore, it can be shown that as the number of observed data becomes arbitrarily large, the asymptotic distribution of each ordinate $I_x(\omega_j)$ of the periodogram converges in distribution ($\xrightarrow{d}$) to a scaled χ^2 distribution (Shumway and Stoffer, 2006, Property P4.2, p. 193):

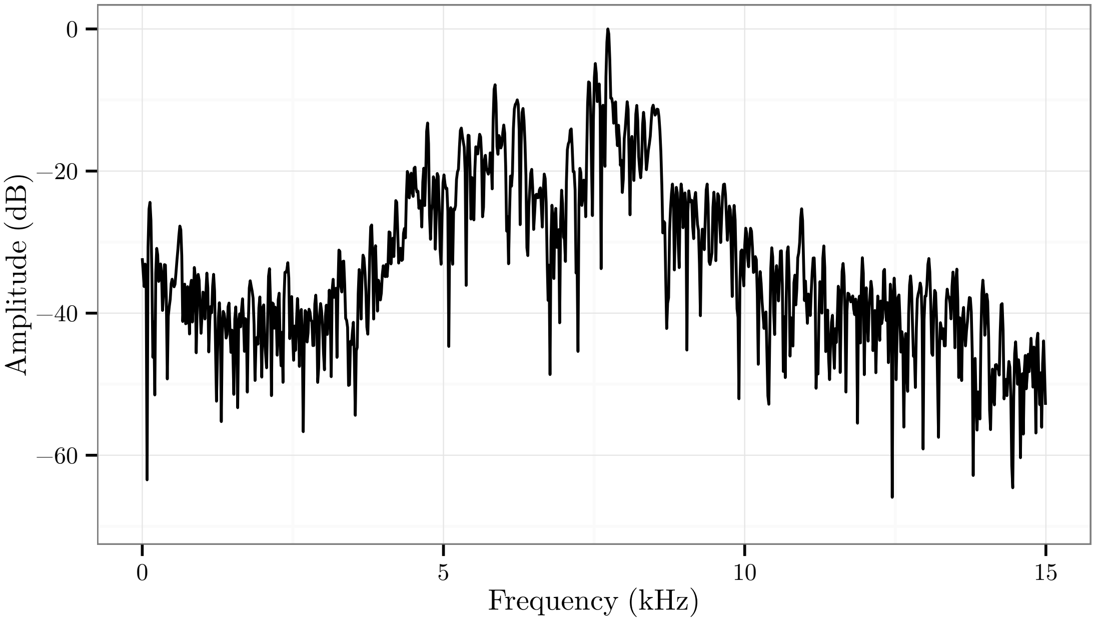
$$I_x(\omega_j) \xrightarrow{d} \frac{f_X(\omega_j)}{2} \chi_2^2. \tag{2.4}$$

As a spectral estimator, the periodogram I_x can be evaluated by considering it as a sequence of point estimators, each of which estimates the amplitude of a particular frequency component in the spectrum f_X, and then assessing the mean squared error[3] (MSE) of these point estimators. Using its asymptotic distribution to approximate the distribution of $I_x(\omega_j)$ when the data are finite in number, it follows from basic properties of the χ^2 distribution that the expected value and variance of $I_x(\omega_j)$ are $\mathbb{E}[I_x(\omega_j)] \approx f_X(\omega_j)$ and $\text{Var}\left[I_x(\omega_j)\right] \approx f_X^2(\omega_j)$, respectively. From these approximations, it follows that the bias of $I_x(\omega_j)$ is negligible, but that its variance is quite large, equal to the square of the quantity that it estimates. The extent of this variance is suggested by the extreme point-to-point variation in the periodogram estimate shown in Fig. 2.2.

[2] The sample autocovariance function is defined defined by $\hat{\gamma}_x(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x})$.

[3] If $\hat{\theta}$ is a point estimator of a random quantity θ, then the bias $\beta_{\hat{\theta}}$ of the estimator is defined as $\beta_{\hat{\theta}} = \mathbb{E}(\hat{\theta}) - \theta$, and the MSE $\mathbb{M}_{\hat{\theta}}$ is defined as $\mathbb{M}_{\hat{\theta}} = \beta_{\hat{\theta}} + \text{Var}(\hat{\theta})$.

Figure 2.2: The periodogram of the middle 20 ms of the waveform of /s/ shown in Fig. 2.1.



2.2.2 Methods for reducing the variance of the periodogram

The desire for a more accurate spectral estimator has led to the development of three techniques for reducing the variance of the periodogram: time averaging, ensemble averaging, and the multitaper spectrum. Each technique succeeds by computing multiple periodogram estimates $I_x^{(1)}, \ldots, I_x^{(K)}$ that are statistically independent from one another, and then taking their pointwise average $\bar{I}_x$. From Eqn. 2.4 and the independence of each periodogram $I_x^{(k)}$, it follows that each ordinate $\bar{I}_x(\omega_j)$ also converges to a scaled χ^2 distribution:

$$\bar{I}_x(\omega_j) = \frac{1}{K}\sum_{k=1}^{K} I_x^{(k)}(\omega_j) \xrightarrow{d} \frac{f_X(\omega_j)}{2K}\chi^2_{2K}. \qquad (2.5)$$

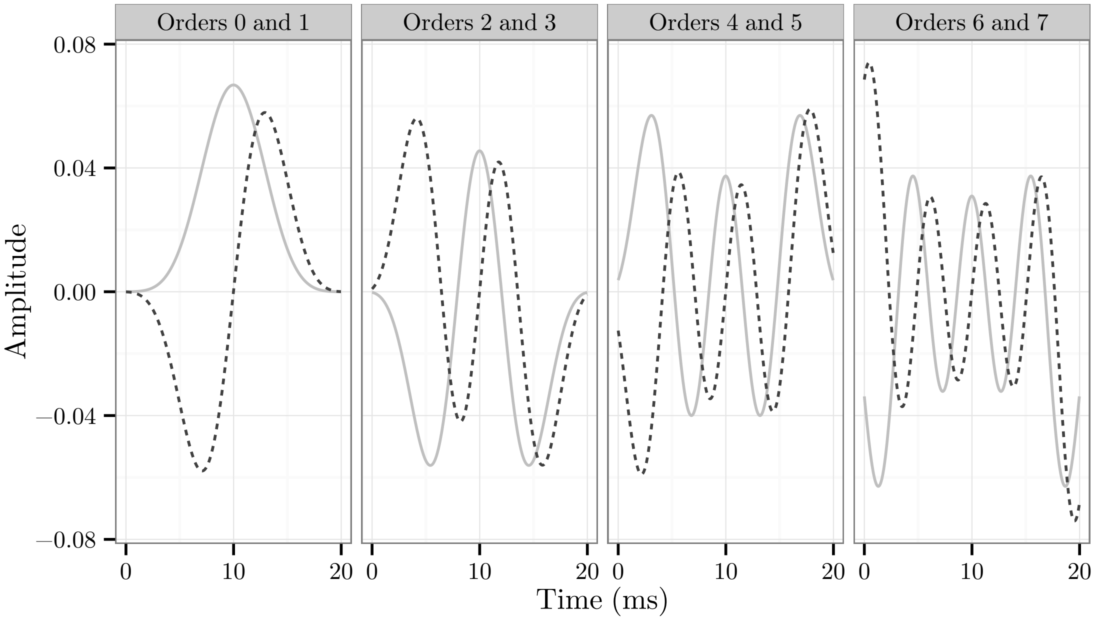
Under its asymptotic distribution, $\mathrm{Var}[\bar{I}_x(\omega_j)] \approx K^{-1} \cdot f_X^2(\omega_j) \approx K^{-1} \cdot \mathrm{Var}[I_x(\omega_j)]$; thus, the variance of $I_x(\omega_j)$ is reduced by a factor of K.

The three techniques differ from each other in how they motivate the independence of the periodogram estimates $I_x^{(1)}, \ldots, I_x^{(K)}$. Under the first technique, time averaging (Welch, 1967), within a single realization of a process, a stationary interval of duration D is divided into K subintervals of duration D/K, and from each interval, a periodogram estimate is computed. Because the periodograms are estimated from different intervals of the realization, these estimates are taken to be independent. The second technique, ensemble averaging, proceeds by estimating the spectrum from multiple realizations of the same process. For example, the spectrum of /s/ would be estimated by recording K productions of /s/, estimating the spectrum from each production, and then averaging these estimates. The multiple estimates are taken to be independent because they were computed from different realizations of the process.

While both time and ensemble averaging successfully reduce the variance of the periodogram, and while both have precedent in the literature on sibilant fricatives (e.g., Newman et al., 2001; Soli, 1981, respectively), neither method is ideal for the current dissertation, whose purpose is to investigate the spectral dynamics of sibilant fricatives and track the development of adult-like spectral dynamics in the productions of children. First, as Shadle (2006) notes, time averaging reduces temporal resolution, which directly cuts against the first goal of this dissertation. Second, when applied to speech data, ensemble averaging reduces multiple productions from a speaker to a single estimate, in effect eliding the variability produced by that speaker across multiple tokens. However, such inter-token variability has been found to index both the intelligibility of adult speakers (Newman et al., 2001) and the development of speech categories in children (Romeo et al., 2013); thus, because a speaker's inter-token variability may serve as a developmental index, ensemble averaging is poorly suited to tracking children's progression toward adult-like productions.

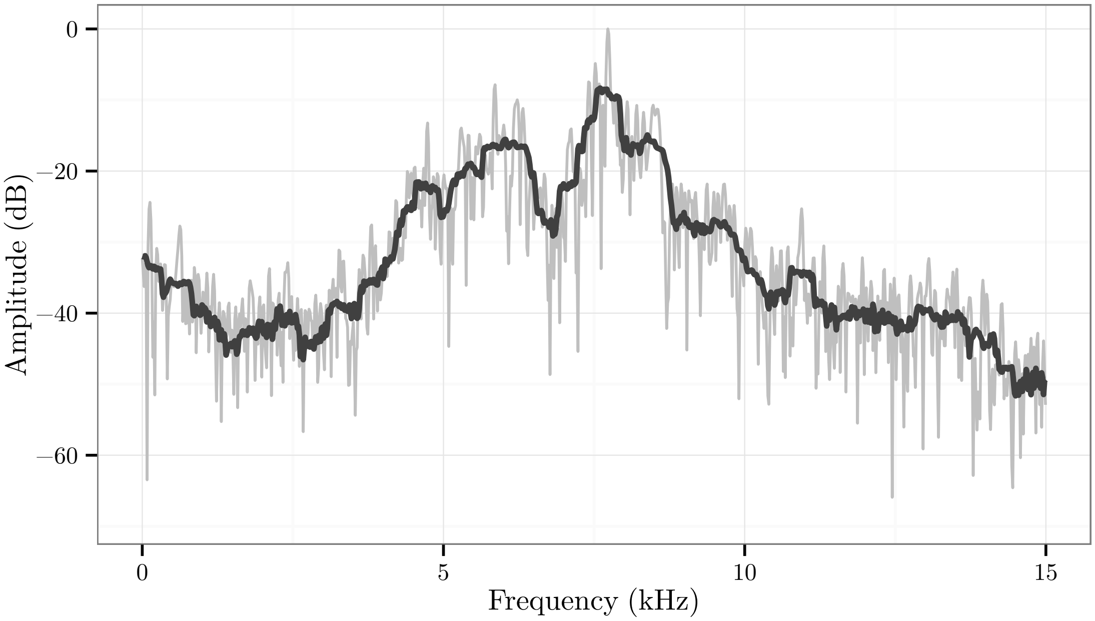
The use of the *multitaper spectrum (MTS)* technique (Thomson, 1982) is motivated by the fact that unlike time and ensemble averaging, the MTS technique is able to preserve temporal resolution and inter-token variability. The main insight behind the technique is to estimate K independent periodograms from copies of a single interval from the same

Figure 2.3: The discrete prolate spheroidal (DPS) sequences of orders $k = 0$ through $k = 7$, computed with parameters $n = 883$ and $W = 4/883$. In each panel, the solid gray line denotes the even sequence, while the dashed black line denotes the odd sequence.



realization, rather than from multiple intervals or multiple realizations. Each of the K copies of the data $x_1, \ldots, x_n$ is multiplied by a different sequence of numbers, referred to as a *taper*, drawn from the class of *discrete prolate spheroidal (DPS) sequences* (Landau and Pollak, 1961, 1962; Slepian, 1964; Slepian and Pollak, 1961). A DPS sequence is denoted by $\left\{v_t^{(k)}\right\}$, where the superscript (k) is an index denoting the *order* of the sequence. A DPS sequence has two parameters: n, the length of each sequence; and W, a bandwidth parameter, constrained such that $0 < W < 1$, that nW is an integer, and that $K \leq 2nW$. Given fixed values of n and W, the DPS sequence $\left\{v_t^{(k)}\right\}$ is the sequence of length n whose Fourier transform has the $(k+1)^{\text{th}}$ greatest concentration of energy within the frequency band $[-W, W]$; thus, $\left\{v_t^{(0)}\right\}$ has the greatest concentration of energy within this frequency band, and K consecutive DPS sequences can be ordered in terms of their spectral concentration: $\left\{v_t^{(0)}\right\} > \cdots > \left\{v_t^{(K-1)}\right\}$. Examples of DPS sequences are shown in Fig. 2.3.

Figure 2.4: The eighth-order multitaper spectrum (black) of the middle 20 ms of the waveform of /s/ shown in Fig. 2.1, plotted on top of the periodogram estimate (gray) of the same data.



These sequences were computed with parameter values $n = 883$ and $W = 4/883$, which are appropriate for $K = 8$ tapers that can be applied to a 20 ms interval of a waveform sampled at 44.1 kHz.

To compute a multitaper spectrum, K copies of the data $x_1, \ldots, x_n$ are tapered by $\left\{v_t^{(0)}\right\}, \ldots, \left\{v_t^{(K-1)}\right\}$, and then a periodogram $I_x^{(k)}$ is computed from each tapered waveform. Because the DPS sequences are orthogonal, in the sense that $\left|\left|v_t^{(a)} \cdot v_t^{(b)}\right|\right|^2 = 0$ for all orders $a \neq b$, the K periodograms are approximately independent (Percival and Walden, 1993). The number of data copies used to compute an MTS is referred to as its *order*. In this dissertation, the spectra of sibilant fricatives will be computed with eighth-order multitaper spectra, following the precedent of previous studies of sibilants that have used the MTS (cf. Blacklock, 2004; Koenig et al., 2013; Romeo et al., 2013; Todd et al., 2011). An example of an eighth-order MTS is shown in Fig. 2.4, where it is plotted on top of the periodogram from Fig. 2.2. A comparison of these two estimates reveals how a reduction in variance engenders smaller point-to-point excursions, and smoother spectral estimate.

2.3 The auditory model

Further reductions to the variance of a spectral estimate can be achieved by passing it through a filter bank model of the auditory system, where each channel acts as a smoothing window on a limited band of the spectral estimate. The output of such a model represents an "auditory spectrum" that denotes the amount of energy at the output of each filter channel in response to an input spectrum.

2.3.1 Two perspectives of the auditory system

In order to motivate the auditory model described below, the auditory system is first considered from two perspectives: the *biophysical* and the *psychoacoustic*. From the biophysical perspective, the auditory system is considered anatomically and physiologically, as the sensory system that responds to a sound wave and produces a physical response within a listener. This response lends insight into how the auditory system warps the spectrum of a sound wave and furnishes a physical basis for the psychoacoustic perspective. From this latter perspective, the auditory system is not treated as a corporeal system, but rather an unobserved link between the acoustic signal and the perceptual judgments of a listener. Such judgments have helped clarify the auditory system's selectivity at different frequencies.

The biophysical perspective

When a sound wave impinges on the outer ear, it causes fluctuations in air pressure within the auditory canal, which act as a driving force on the eardrum, causing it to vibrate. The vibration of the eardrum induces the bones of the middle ear to oscillate rotationally. The innermost of these bones, the stapes, is fused to the membrane covering the oval window of the cochlea; thus, when oscillating, the stapes acts as a piston, delivering a driving force to the cochlea.

The cochlea is a bony structure of the inner ear that is coiled in shape, resembling a nautilus. The end of the cochlea that is closest to the stapes is called the base; whereas, the innermost point of the coil is called the apex. Within the cochlea are three fluid-filled ducts: the scalae tympani, media, and vestibuli. At the base of the cochlea, the scala vestibuli terminates at the oval window, while at the apex, it merges with the scala tympani. Due to the oscillation of the stapes, the oval window vibrates, sending traveling pressure waves through the fluid that fills the scalae vestibuli and tympani. As these waves travel through the scala tympani, the pressure differences between it and the scala media induce a transverse wave along the basilar membrane, which separates these two scalae.

In response to a pure tone, this transverse wave travels from the base to the apex of the basilar membrane, gradually increasing in amplitude before peaking and then sharply decaying to a flat response (von Békésy, 1947). For pure tone stimuli, as the frequency of the tone increases, so does the distance along the basilar membrane between its apex and its point of maximum displacement during vibration (Gundersen et al., 1978; von Békésy, 1947). Such differences in the amplitude of vibration result from the physical attributes of the basilar membrane, which is narrowest at its base and broadens gradually to its widest point at the apex. Additionally, the basilar membrane is stiffest at the base and reduces in stiffness as one moves toward the apex. Due to the differences in width and stiffness along its length, each cross-section of the membrane is tuned to a particular frequency. In particular, the stiff, narrow regions near the base respond most sensitively to high frequency tones, while the pliable, wide regions near the apex are tuned to low frequencies. Hence, position x on the basilar membrane can be construed as a function of frequency ω. Such a function is called the *tonotopic map*, $x = \mathcal{T}(\omega)$.

Since the tonotopic map monotonically associates frequency to position on the basilar membrane, it can be viewed as a transformation of the physical frequency scale to a biophysical frequency scale. While this transformation of scales preserves the order of frequencies, it does not preserve the distance between them. In particular, if distance x from the apex of the basilar membrane is measured in millimeters and frequency ω in Hz, then $x \approx 16.667 \cdot \log_{10}(\omega/165.4 + 0.88)$ (Greenwood, 1961, 1990). The tonotopic map, therefore, compresses the physical frequency scale logarithmically.

Furthermore, because each cross-section of the basilar membrane is tuned to a specific frequency, its vibrational response to a complex sound represents an imperfect spectral analysis of that sound into its component frequencies. In order to resolve each frequency component independently of all others, each cross-section of the basilar membrane would be required to vibrate independently as well. However, because a section of the membrane cannot vibrate without also displacing nearby points, such representational independence does not obtain.

The psychoacoustic perspective

An early psychoacoustic demonstration of the auditory system's imperfect frequency selectivity is found in the masking experiments of Fletcher (1940). In these experiments, a pure tone signal was mixed with a bandlimited, aperiodic noise masker. In each trial, the band of noise was centered at the frequency of the tone, and the sound level of the masker was held constant, while the level of the tone was varied to determine its threshold. Across trials, the noise masker's bandwidth was varied. Of interest was the effect of the masker's bandwidth on the threshold of the pure tone. Fletcher found that as this bandwidth increased, so did the threshold of the tone, but only up to a point: once the noise masker reached a certain bandwidth, further increases had no effect.

To explain these findings, Fletcher suggested that the detection of a pure tone embedded within masking noise is mediated by a bandpass "auditory filter," whose center frequency is equal to that of the tone. The tone is detected only if, once passed through the auditory filter, the ratio of the energy of the tone to that of the masker exceeds some detection criterion. As long as the bandwidth of the masker is less than that of the auditory filter, then widening the noise band will increase its total energy under the auditory filter, raising the threshold level of the tone. However, once the masker's bandwidth equals or exceeds that of the auditory filter, further increases to its bandwidth will not affect the total energy at the output of the auditory filter.

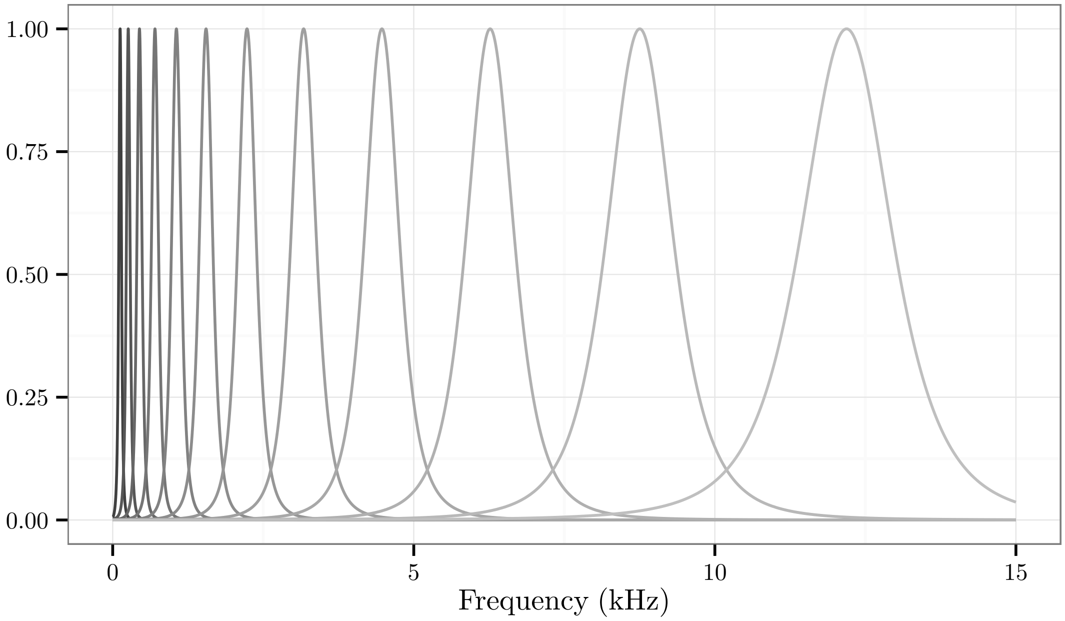
Subsequent psychoacoustic masking experiments (e.g., Patterson, 1976) have revealed that the shape of an auditory filter's frequency response resembles an asymmetric unimodal concave curve that is skewed left: the attenuation of an auditory filter is much steeper above its center frequency than below it. To approximate the filter shapes indicated by the behavioral data, psychoacousticians have suggested fourth-order gammatone filters (c.f. Glasberg and Moore, 1990; Patterson, 1976). In general, a gammatone filter can be described by its impulse response:

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi\omega t), \tag{2.6}$$

where ω is the center frequency, b is the bandwidth, n is the order, and a is the amplitude of the filter. Computing the amplitude spectrum of a filter's impulse response yields its frequency response, which summarizes the filter's attenuation of frequencies within its passband. Examples of the frequency responses of fourth-order (i.e., $n = 4$) gammatone filters are shown in Fig. 2.5.

The frequency responses shown in Fig. 2.5 make clear that as the center frequency of a gammatone filter increases, so does the width of its passband. This pattern is likewise reflected in the psychoacoustic data: auditory filters are wider at high frequencies than at low frequencies, when frequency is expressed on the linear hertz scale. In order to summarize the relationship between a filter's center frequency and its width, the notion of equivalent rectangular bandwidth has been introduced (c.f. Moore, 1997). Given a bandpass filter of arbitrary shape, its *equivalent rectangular bandwidth (ERB_N)* is equal to the total area

Figure 2.5: The frequency responses of fourth-order gammatone filters that differ in center frequency and that are spaced evenly along the ERB_N-scale.



under its passband. In other words, the ERB_N of a filter $\mathcal{F}$ determines the width of a perfectly rectangular filter such that it passes the same amount of energy as $\mathcal{F}$. Much effort has been put into measuring the ERB_N of auditory filters of various center frequencies (e.g., Moore and Glasberg, 1983, 1996; Moore et al., 1997). These studies have found that the ERB_N of an auditory filter increases linearly with its center frequency ω, such that

$$\mathrm{ERB}_N(\omega) = 0.107939\omega + 24.7, \tag{2.7}$$

when ω is expressed in hertz. Thus, the auditory system resolves higher frequencies more poorly than lower frequencies.

The frequency responses shown in Fig. 2.5 additionally indicate the auditory system's logarithmic compression of the frequency scale: the gammatone filters' center frequencies are more densely distributed in low-frequency regions. The psychoacoustic compression of the linear hertz frequency scale is derived by identifying the width of a filter with its ERB_N, and then treating Eqn. 2.7 as a differential equation that describes the rate at which the number of filters changes with frequency:

$$\frac{\mathrm{d}\,\omega}{\mathrm{d}\,\mathrm{ERB}_N} = 0.107939\omega + 24.7. \tag{2.8}$$

With the boundary condition $\mathrm{ERB}_N(0) = 0$, the solution to this differential equation is

$$\mathrm{ERB}_N\text{-number}(\omega) = \frac{1}{0.107939} \ln\left(0.00437\omega + 1\right). \tag{2.9}$$

The psychoacoustic frequency scale that results from the compression of the Hz scale in Eqn. 2.9 is referred to as the *ERB_N-scale* (Glasberg and Moore, 1990); a point on the ERB_N-scale is referred to as *ERB_N-number*.

2.3.2 A computational model of the auditory system

Taken together, the biophysical and psychoacoustic perspectives suggest that the auditory system modifies the spectrum of an acoustic signal in two important ways: First, its frequency components are not resolved independently of one another, with the width of the auditory filters increasing with their center frequency. Second, the physical frequency scale is compressed logarithmically. In this section, a model[4] of the auditory system that incorporates its differential resolvability and its frequency compression is described. The model is formulated within the theory of signals and systems (see Beerends et al., 2003, for a well-written introduction). Specifically, the auditory system is modeled as a linear, time-invariant system, assumptions that hold at moderate sound levels (Moore, 1997). As

[4] A number of other implementations for similar models exist in the literature (for a review, see Lyon et al., 2010). These other implementations have tended to emphasize on computational efficiency; whereas, the focus here is to formulate the model in such a way that makes the distributional properties of its output most apparent (see, §2.3.3).

a linear, time-invariant system, the model is formulated in the frequency-domain, taking a spectral estimate as input.

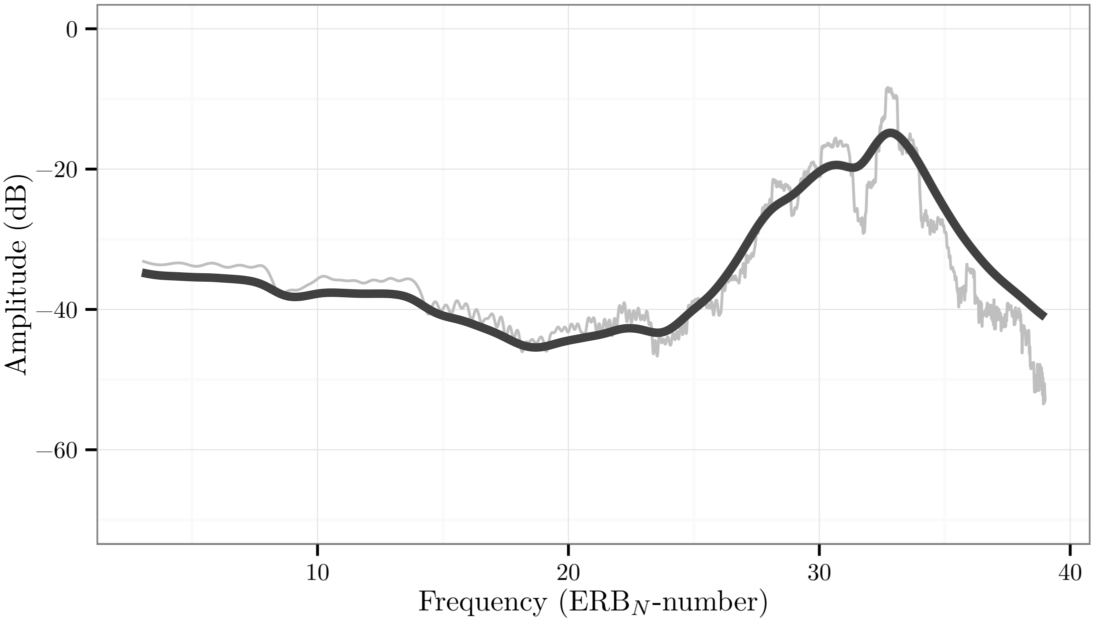
The model comprises some number C of auditory filters, each one represented by the frequency response of a gammatone filter, and is constructed in the following way: First, the center frequencies of the auditory filters are chosen so that they are equally spaced on the ERB$_N$-scale. In general, this is done by choosing the number C of filters in the model, and the interval $[\varepsilon_1, \varepsilon_C]$ on the ERB$_N$-scale within which the C center frequencies are evenly spaced. The center frequency ε_c on the ERB$_N$ scale, of the c^{th} filter is $\varepsilon_c = \varepsilon_1 + (c-1)(\varepsilon_C - \varepsilon_1)/C$, for $1 \leq c \leq C$. The center frequencies in Hz, ω_c, are then found by inverting Eqn. 2.9, $\{\omega_c = \text{ERB}_N\text{-number}^{-1}(\varepsilon_c) \mid 1 \leq c \leq C\}$. By spacing the center frequencies evenly on the ERB$_N$-scale, the auditory model incorporates the psychoacoustic compression of the physical frequency scale.

Second, for the c$^{\text{th}}$ filter, the frequency response of a gammatone filter is computed, using parameter values: $\omega = \omega_c$, $b = 1.019 \cdot \text{ERB}(\omega_c)$, $n = 4$, and $a = 1$. The values of the frequency response are then scaled so that its maximum value is one. The input spectrum is passed through the filter by multiplying it pointwise with the filter's frequency response, and the output of the filter is summed across frequency to determine the total energy of its output. Thus, the c$^{\text{th}}$ filter parses the input spectrum at ω_c with imperfect resolution. Furthermore, because the bandwidths of the filters increase with center frequency, the model incorporates the auditory system's differential frequency resolvability.

Representing the input spectral estimate $\hat{f}_x$ as an $n \times 1$ column vector, and the auditory model $\mathbf{A}$ as a $C \times n$ matrix, in which each row represents the frequency response of an auditory filter, then the response of the model to the input is the $C \times 1$ column vector, $\mathcal{E}_x = \mathbf{A}\hat{f}_x$. This output is referred to as an *excitation pattern* and denotes the total energy at the output of the auditory filter as a function of its center frequency.

Figure 2.6 shows an excitation pattern that is produced in response to the multitaper spectral estimate for /s/ shown in Fig. 2.4. This excitation pattern is the response of

Figure 2.6: The excitation pattern (black) output by the auditory model in response to the MTS estimate (gray) of the spectrum of /s/ from Fig. 2.4.



an auditory model that consisted of 361 gammatone filters whose center frequencies were spaced every 0.1 ERB$_N$-number from 3 to 39 ERB$_N$-number (i.e., from 0.1 to 15.2 kHz, approximately). Plotted underneath the excitation pattern is the MTS from Fig. 2.4, after transforming the Hz scale to the ERB$_N$-scale. A comparison of the MTS plots in Figs. 2.4 and 2.6 illustrates how the psychoacoustic compression of the frequency scale modifies the shape of the spectrum. Meanwhile, a comparison of the two plots in Fig. 2.6 illustrates how the gammatone filters imperfectly resolve the frequency components in the MTS estimate, reflecting the biophysical constraints of the basilar membrane. In particular, between 30 and 35 ERB$_N$-number, the MTS depicts two prominent peaks separated by a deep valley; in the excitation pattern, both peaks are still apparent, but the depth of the valley is greatly reduced.

2.3.3 The auditory model as a series of smoothing windows

The auditory model described in §2.3.2 can, furthermore, be viewed as a collection of smoothing windows that are applied to a spectral estimate's components that occur at the center frequencies of the auditory filters in the model. That is, the component at frequency ω_c in the spectral estimate, input to the model, is smoothed by the frequency response of the auditory filter whose center frequency is $\omega_c = \text{ERB}_N\text{-number}^{-1}(\varepsilon_c)$. Because the auditory filter is modeled as a gammatone filter, its frequency response is non-zero valued on only a limited contiguous frequency band centered at ω_c. Thus, the component of the excitation pattern, $\mathcal{E}_x(\varepsilon_c)$, is equal to a weighted sum of approximately independent estimates of the spectral components nearby ω_c (Shumway and Stoffer, 2006, §C.2). As a consequence of the overlap across adjacent gammatone filters in the model, the point-to-point variation of the input spectral estimate is reduced by the auditory model.

2.3.4 Spectral dynamics from multiple 'glimpses'

As formulated above, the auditory model outputs a static representation of the distribution of auditory excitation across psychoacoustic frequency. In order to use this model to investigate the spectral dynamics of sibilant fricatives, this dissertation subscribes to a 'glimpsing' model of speech processing (cf. Apoux and Healy, 2009; Cooke, 2003, 2006; Viemeister and Wakefield, 1991). According to such a model, the auditory periphery decomposes a sound into a spectrographic representation, whose regions denote the local time-frequency properties of the sound. Subsequent internal representations of a sound are then constructed by integrating multiple regions of this peripheral spectrographic representation. In this dissertation, the peripheral spectrographic representation of a sibilant fricative will be approximated by estimating an excitation pattern from multiple points across the interval of frication; the only subsequent representation will be an ordered sequence of ERB_N-numbers denoting the peak psychoacoustic frequency for each time interval at which an excitation pattern was computed.

2.4 Conclusion

In this chapter, the statistical setting for the spectral analysis of sibilant fricatives was reviewed. A nonparametric spectral estimator, the periodogram, was introduced and its asymptotic distributional properties were developed. Because the variance of the periodogram's ordinates is undesirably large, methods for reducing this variance were explored. The multitaper spectrum was argued to be an appropriate method for reducing the periodogram's variance without averaging across time windows or multiple realizations. Lastly, it was argued that further reductions in variance could be achieved by passing a spectral estimate through a gammatone filter bank model of the auditory system that incorporated its differential frequency selectivity and its frequency compression. In the next section, this auditory model is used to represent, in psychoacoustic terms, the time course of sibilants' spectral content.

Chapter 3

The spectral dynamics of sibilant contrast

3.1 Introduction

3.1.1 Articulatory postures of sibilant fricatives

For sibilant noise to be generated, the tongue must be positioned close to the roof of the mouth so that the oral cavity becomes constricted, but not fully occluded. Moreover, this constriction must be sufficiently narrow, so that the airstream becomes turbulent as it flows through the constriction; and the tongue surface behind and forming the constriction must be positioned in such a way that the turbulent airflow exiting the constriction is directed at the incisors downstream. While these constraints on sibilant noise generation limit the place of the constriction to a particular region of the vocal tract (i.e., no farther forward than the incisors and no farther back than the palate), this region is sufficiently spacious to permit a number of different types of constriction that produce sibilant turbulence noise. The range of this permissible variation is partly exemplified by English /s/ and /ʃ/ and Japanese /s/ and /ɕ/.

English /s/ and /ʃ/

The English sibilants differ, foremost, in terms of where along the roof of the mouth, and with which part of the tongue, the linguapalatal constriction is made: /s/ can be either apical or laminal, with either the tongue tip or blade raised to form the constriction at the upper incisors or on the alveolar ridge; whereas, /ʃ/ is typically laminal, with the constriction made by the tongue blade, posterior to the alveolar ridge (Narayanan et al., 1995). Thus, the place of articulation of /s/ is anterior to that of /ʃ/.

Additionally, /s/ and /ʃ/ differ in terms of the shape assumed by the tongue during articulation. For /s/, the surface of the tongue body is concave, with its lateral edges raised relative to its midline. For /ʃ/, however, the tongue surface is concave only along its posterior; along the middle and anterior sections, the surface of the tongue is level in the coronal plane (Stone and Lundberg, 1996). These differences in tongue shape are adapted to the palatal morphology at each consonant's place of articulation, such that, for both sibilants, the lateral edges of the tongue contact the palate across its full length (McLeod et al., 2006; Stone and Lundberg, 1996). A greater amount of lateral contact is made during the articulation of /s/, a consequence of which is that, in the coronal plane, its constriction is relatively narrower than that of /ʃ/ (Fletcher and Newman, 1991). In the midsagittal plane, however, the constriction for /s/ is both greater in height and slightly shorter, due to the cross-sectional concavity of the tongue surface (Toda and Honda, 2003).

The geometry of the front cavity, anterior to the constriction, is determined by the location of the constriction and the postures of the tongue tip and the lips. Since the constriction for /ʃ/ lies posterior to that of /s/, the front cavity is longer for /ʃ/. Additionally, the average cross-sectional area of the front cavity is wider for /ʃ/. These dimensional asymmetries result in a larger front cavity volume for /ʃ/ than for /s/ (Narayanan et al., 1995). Furthermore, these quantitative differences in volume are accompanied by qualitative differences in shape: for /ʃ/, the tongue is postured such that a sublingual cavity forms posterior to the lower incisors, whereas /s/ is more often articulated with the underside of the tongue tip making contact with the lower incisors, eliminating this sublingual cavity (Perkell et al., 2004).

Finally, /ʃ/ is articulated with a greater degree of lip rounding, which is realized as a difference in both lip shape and extent of lip protrusion. Toda et al. (2002) studied these dissociable effects of lip rounding on English sibilants, and found that, relative to the alveolar, the postalveolar is articulated with a rounder lip shape, smaller cross-sectional lip area, and more extensive lip protrusion.

Japanese /s/ and /ɕ/

The difference in articulatory posture between Japanese /s/ and /ɕ/ is traditionally described as one of the constriction's degree of palatalization, rather than its location (Akamatsu, 1997). For both Japanese sibilants, the anterior end of the constriction is located on the alveolar ridge or at the upper incisors; however, for /ɕ/ the constriction is much longer, extending posteriorly from the alveolar ridge to the anterior border of the hard palate (Toda and Honda, 2003).

Despite the similarity in the location of the anterior end of the constriction for the two Japanese sibilants, individual speakers consistently differentiate /s/ and /ɕ/ in terms of front cavity size. Analyzing magnetic resonance images (MRI) of sustained articulatory postures, Toda and Honda (2003) found that, for each participant, the area of the front-cavity in the mid-sagittal plane was greater for /ɕ/ than for /s/; however, across speakers, two-thirds of the front cavity areas for /ɕ/ fell within the observed range for /s/. Thus, the consistent speaker-internal difference in front cavity size is qualified by the significant amount of inter-speaker overlap along this articulatory dimension.

Cross-linguistic similarities between English and Japanese /s/

While the sibilant contrast is instantiated with different articulatory parameters in English and Japanese, the MRI data reported by Toda and Honda (2003) suggest that the articulatory posture for /s/ is comparable in these two languages. First, the place of articulation was observed to be either dental or alveolar, in comparable proportions across speakers. Additionally, the English and Japanese speakers exhibited similar ranges of front cavity size and of palatalization for their /s/ productions (Toda and Honda, 2003, Fig. 4).

3.1.2 Static spectral properties of sibilant fricatives

During speech production the acoustic noise sources differentially excite the various cavities of the vocal tract, whose resonances vary according to their size and shape. Consequently, as a sound wave propagates from its source, its frequency components are amplified or damped according to the resonant frequencies of the vocal tract cavities that it excites. Under the source-filter theory of speech production (Fant, 1960), the vocal tract is modeled as a system that filters the acoustic signal generated by the noise sources, which yields the output speech signal. In the frequency-domain, this action of the vocal tract on the noise sources is equivalent to the multiplication of the latter's spectrum with the former's *transfer function*, whose peaks represent the resonances of the vocal tract.

Source-filter models of vocal tract configurations like those that characterize the static articulatory postures for sibilant fricatives, where a narrow constriction divides the vocal tract and where the acoustic noise sources occur anterior to this constriction, suggest that very little acoustic energy excites the back cavity, leaving it acoustically inert (Heinz and Stevens, 1961). Because the constriction decouples the back cavity from the turbulence noise sources, it is primarily the front cavity that filters the turbulence noise sources of sibilant fricatives (Stevens, 2000). Furthermore, the size and geometry of the front cavity determine the distribution of peaks in its transfer function, with a smaller front cavity associated with a concentration of peaks at higher frequencies. Finally, the shape of the transfer function largely determines the shape of the output sibilant's spectrum since the spectrum of the turbulence noise sources is flat across a wide frequency range (Toda et al., 2010).

In the literature, two spectral properties have been proposed as indices of the frequency at which energy is concentrated: centroid and peak frequency. Centroid is computed from a spectral estimate $\hat{f}_x$ by first normalizing its amplitude values so that they sum to one. Since a spectral estimate is a non-negative function of frequency, the amplitude-normalized estimate can be treated as a discrete probability mass function. Centroid is then computed as the mean frequency of this probability mass function:

$$\text{centroid} = \sum_{\omega} \frac{\hat{f}_x(\omega)}{\sum_{\omega} \hat{f}_x(\omega)} \cdot \omega. \tag{3.1}$$

The peak frequency of $\hat{f}_x$ is simply the frequency of the maximum amplitude:

$$\text{peak} = \underset{\omega}{\operatorname{argmax}}\, \hat{f}_x(\omega). \tag{3.2}$$

While centroid and peak frequency are conceptually distinct spectral properties, they both indicate the "central frequency," in the same way that mean and mode are both measures of the center of a distribution.

Since the size and geometry of the front cavity differ between English /s/ and /ʃ/ and between Japanese /s/ and /ɕ/, these pairs of sibilants are expected to also differ in terms of where energy is concentrated in their spectrum. Specifically, energy is expected to be concentrated at higher frequencies in the spectrum of English /s/ and of Japanese /s/ relative to the spectrum of /ʃ/ or of /ɕ/. Conversely, due to the cross-linguistic similarity in how English /s/ and Japanese /s/ are articulated, these two sounds are expected to be comparable in terms of their centroid and peak frequencies (see §3.3.3).

English /s/ and /ʃ/

A number of studies of the English sibilants have found that centroid frequency is higher in /s/ than in /ʃ/. When estimated from either the beginning, middle or end of the frication noise, centroid has been found to be greater in /s/ than in /ʃ/ (Jongman et al., 2000). Similar results have been found when centroid estimates from multiple locations across the fricative are pooled together (Fox and Nissen, 2005; Maniwa et al., 2009). In each of these studies, the difference in centroid frequency was evaluated across a group of speakers. When the perspective shifts to the individual speaker, the same pattern obtains (Haley et al., 2010; Li et al., 2009). Finally, when used in classification tasks, centroid frequency has been found to yield high, sometimes even perfect, accuracy (Forrest et al., 1988; Li et al., 2009; McMurray and Jongman, 2011), indicating that these two sibilants differ reliably in terms of their centroid frequency. As is the case with centroid, peak frequency has been found to be higher in /s/ than /ʃ/, regardless or whether it is measured at the beginning, middle, or end of frication (Behrens and Blumstein, 1988; Hughes and Halle, 1956).

Japanese /s/ and /ɕ/

Acoustic analyses of the Japanese sibilants have found that /s/ and /ɕ/ exhibit significant differences in terms of either centroid or peak frequency, but that classification is typically improved by the inclusion of some other acoustic measure taken near the vowel onset. Because the linguapalatal constriction is longer during the production of /ɕ/ than of /s/, the size of the cavity behind the constriction is smaller for /ɕ/ than for /s/. Thus, when the constriction is released and the back cavity recoupled to the rest of the oral tract, the resonances of this cavity are represented in the spectrum. Li et al. (2009) found that each of their five recorded speakers produced /s/ with a significantly higher centroid frequency, but that in addition to centroid frequency, $F2$ at vowel onset need to be included in a logistic regression model in order to achieve perfect classification. Similar results were reported for a larger group of speakers in Li (2012).

3.1.3 Articulatory kinematics and spectral dynamics of English /s/

Iskarous et al. (2011) used x-ray microbeam measurements to investigate the kinematics of certain articulators—e.g., the jaw and tongue tip—and the changes to certain geometric properties of the vocal tract configuration—e.g., constriction degree and location—involved in fluent word-initial productions of English /s/. Regarding the geometric properties, it was found that when /s/ occurred pre-vocalically, the degree and location of the constriction remained approximately constant until the final 20%, or so, of the fricative's duration, at which point the constriction opened significantly in anticipation of the following vowel.

The relative constancy of these geometric properties of the constriction, across relatively long durations of the fricative, belied the kinematics of the articulators, whose movements determine the overall vocal tract geometry at any point in time. Of all the articulators, the jaw exhibited the most consistent pattern of movement: rising through the first half of /s/, and then falling through the second half. The tongue tip, on the other hand, moved in opposition to the jaw, which maintain a relatively constant degree of constriction.

In addition to these articulatory measurements, Iskarous et al. (2011) also recorded the acoustic signal of each /s/ production in order to examine how its centroid and peak frequency vary across its duration, as a consequence of the kinematics of the articulators. These two measures were said to vary in similar ways; hence, only centroid was discussed in detail. Centroid was found to follow an increasing concave downward trajectory, similar to that of the jaw, which led the authors to suggest that the increase in centroid frequency across the first half of /s/ was partially due to the rise of the jaw, the height of which affects the noise source generated when turbulent airflow impinges on the lower incisors.

3.1.4 Purpose and hypotheses

Because Iskarous et al. (2011) studied only English /s/, it is unknown whether the temporal variation in centroid frequency that they observed is specific to this sibilant or whether it is a general property of the acoustics of sibilant fricatives that holds across sibilants within a given language or across similarly articulated sibilants cross-linguistically. The analysis in this chapter extend the acoustic results of Iskarous et al. (2011) in three ways.

First, instead of investigating temporal variation in centroid frequency, the analyses here investigate the dynamics of a psychoacoustic measure of peak frequency, peak ERB$_N$-number, which denotes the psychoacoustic frequency of the auditory filter that is most activated by an incoming acoustic signal. Peak frequency is preferred over centroid because it has been found to correlate more strongly than centroid[1] with speakers' perceptual

[1] A number of studies of the English sibilant contrast have found a significant correlation between centroid frequency and some aspect of perception. For example, a speaker's degree of contrast between /s/ and /ʃ/,

prototypes of English /s/ and /ʃ/ (Newman, 2003).

Second, the peak ERB_N-number trajectories of English /s/ and Japanese /s/ will be compared to those of English /ʃ/ and Japanese /ɕ/, respectively. The purpose of these within-language analyses is to determine whether contrastive sibilants in a given language differ not just in terms of static spectral properties at a given point (cf. §3.1.2), but also in terms of the pattern of temporal variation in one such property. Specfically at interest here is whether the peak ERB_N-number trajectories of English /s/ and /ʃ/, or of Japanese /s/ and /ɕ/, differ in terms of their shape, such that one is not simply the translation of the other along the peak ERB_N-number scale. Since, within a given language, the voiceless sibilants are articulated with different postures, it is hypothesized that differences in how the articulators must move to form and release these postures will lead to differences in the shapes of the two sibilants' peak ERB_N-number trajectories.

Third, since any observed differences across sibilants within a language may be due to kinematic requirements on the articulators as they move to form, maintain, and release the linguapalatal constriction, two sibilants that have comparable articulatory postures, English and Japanese /s/, will be compared cross-linguistically. This comparison is intended as an exploration of potential sources of trajectory shape difference, and so no specific predictions are made regarding whether the peak ERB_N-number trajectory of English and Japanese /s/ will differ in shape. If cross-linguistic differences are not observed, it would suggest that the spectral dynamics of sibilant fricatives reflect only the articulatory kinematics necessary to achieve a stable posture. On the other hand, the presence of such differences would suggest that the spectral dynamics reflect language-specificity either in the gesture for the sibilant itself or in the coordination of the sibilant gesture with that for the neighboring phonetic

in terms of centroid frequency, has been found to correlate with their auditory and articulatory acuity (Brunner et al., 2011; Ghosh et al., 2010; Perkell et al., 2004) and to affect other listener's perception and processing of their productions of /s/ and /ʃ/ (Hazan and Baker, 2011; Newman et al., 2001); however, none of these studies compared how well peak frequency correlated with the perceptual measure. To the author's knowledge, only Newman (2003) has made such a comparison, finding peak frequency to have a stronger association with perception.

context—in this case, the following vowel.

3.2 Method

Adults' productions of sibilant fricatives were drawn from the English and Japanese portions of the Paidologos corpus, which resulted from a large-scale cross-sectional and cross-linguistic study of the acquisition of obstruent consonants by young children (Edwards and Beckman, 2008a,b). For each language, the corpus also includes productions from native adult speakers that were elicited using the same language materials and procedure as with the children.

3.2.1 Participants

Twenty adult, native speakers of each language were recruited to complete a picture-prompted word-repetition task. Each group of twenty speakers was balanced across gender. The English-speaking participants were recruited from the Columbus, OH metropolitan area, while the Japanese-speaking participants were recruited from Tokyo, Japan. Using a test of otoacoustic emissions at 2, 3, 4, and 5 kHz, all participants were determined to have normal hearing at the time of testing. Furthermore, none of the adult participants reported any history of speech, language, or hearing disorders.

3.2.2 Materials

The target words for the repetition task included sibilant-initial real words, in which the sibilant occurred before a vowel. Tables 3.1 and 3.2 show the target words for English and Japanese, respectively. As a part of a larger study of the acquisition of consonants by children, these words were chosen because they are familiar to young children and because their referent is easily picturable.

Table 3.1: The /s/- and /ʃ/-initial target words used in the English word-repetition task.

Vowel context	**/s/-initial words**	**/ʃ/-initial words**
/i/	sister	sheep
	seal	shield
	seashore	ship
/e/	safe	shape
	same	shell
	seven	shepherd
/ɑ/	sauce	shark
	soccer	shop
	sun	shovel
/o/	soak	shore
	sodas	shoulder
	soldier	show
/u/	soup	chute
	suitcase	shoe
	super	sugar

Since the vowel inventory of English is larger than that of Japanese, the English monophthongs were grouped into classes that correspond roughly to the five Japanese monophthongs /i, e, ɑ, o, u/. These English vowel classes elided certain features, like the tense-lax distinction, that are not likely to induce coarticulatory effects on sibilants. Specifically, the English /i/ category comprised /i, ɪ/; /e/, /e, ɛ/; /ɑ/, /ʌ, ɑ, ɔ/; and /u/, /u, ʊ/

For English, three words for each sibilant in each vowel context were chosen, which yielded a total of 15 target words for /s/ and /ʃ/, respectively. In Japanese, /si/ is not a phonotactically legal sequence, and /ɕe/ occurs only in words that are unlikely to be familiar to young children (e.g., /ɕeri/ 'sherry'). Consequently, no target words were included for these two sibilant-vowel sequences, which left 12 target words for either sibilant in Japanese. In addition to these target words, the word list for each language included stop- and affricate-initial words.

For each language, an adult, female, native speaker, who had received phonetic training, produced multiple repetitions of each word in a child-directed register. These productions

Table 3.2: The /s/- and /ɕ/-initial target words used in the Japanese word-repetition task.

Vowel context	/s/-initial words		/ɕ/-initial words	
	Gloss	Transcription	Gloss	Transcription
/i/			‘bullet train’	/ɕiɴkɑɴseɴ/
			‘seesaw’	/ɕiːsoː/
			‘zebra’	/ɕimɑumɑ/
/e/	‘back’	/senɑkɑ/		
	‘cicada’	/semi/		
	‘teacher’	/seɴseː/		
/ɑ/	‘cherry blossom’	/sɑkurɑ/	‘rice paddle’	/ɕɑmodʒi/
	‘fish’	/sɑkɑnɑ/	‘shampoo’	/ɕɑwɑː/
	‘monkey’	/sɑru/	‘shower’	/ɕɑmpuː/
/o/	‘sausage’	/soːseːdʒi/	‘bread’	/ɕokupɑɴ/
	‘sky’	/sokːusu/	‘fire engine’	/ɕoːboːɕɑ/
	‘socks’	/sorɑ/	‘soy sauce’	/ɕoːju/
/u/	‘sand’	/sunɑ/	‘dumpling’	/ɕuːmɑi/
	‘sparrow’	/sudzume/	‘creme puff’	/ɕuːkuriːmu/
	‘watermelon’	/suikɑ/	‘shoes’	/ɕuːdzu/

were recorded digitally at 22.5 kHz, and from these recordings, three repetitions of each word were chosen to combine with other words to create six lists of auditory stimuli (i.e., two ordered-lists for each of the three sets of audio recordings of the words). The order within each list was pseudo-randomized, so that the words for each target sibilant-vowel pair were distributed evenly across the list. Finally, the auditory stimuli were paired with digital images of the referent of the target word, and these audiovisual pairs were used as prompts in the repetition task.

3.2.3 Elicitation and recording procedure

The English speakers completed the repetition task inside a sound-attenuated room on the campus of The Ohio State University. The Japanese speakers were tested in a quiet room in Tokyo, Japan. Prior to the task, the participants were instructed that they would be completing an experiment that was designed primarily for young children, and that this task would require them to repeat real words of their native language after being prompted by paired images and audio recordings of those words.

The adults completed the task at their own pace, using a custom software program that allowed them to initiate each trial and to track their progress through the task. On a given trial, the software program first displayed the associated image on a computer screen, and then, after a 300 ms delay, played the audio recording over speakers.

The adults' repetitions of the test words were spoken into an AKG C5900M condenser microphone with a cardioid response and recorded using a Marantz PMD660 flash card recorder with 44.1 kHz sampling frequency. The full duration of the task was recorded for subsequent annotation and acoustic analysis.

3.2.4 Annotation of fricative events

A team of trained phoneticians marked the frication onsets and the fricative-vowel boundaries of the target sibilants using a custom Praat script that displayed the recording's waveform and spectrogram simultaneously and that allowed it to be played at will. Frication onset was marked at the earliest point at which an increase in the waveform's amplitude coincided with the presence of high-frequency energy in the spectrogram. For the fricative-vowel boundary, the onset of periodicity in the vocalic portion was first determined. The fricative-vowel boundary was then marked at the zero-crossing of the waveform's first upswing following the onset of periodicity.

In the English data, two productions of /se/ were not annotated, and were omitted from subsequent analysis, because participants e9gt03fw and e9gt10fw produced 'fame' instead of the target 'same'. This seemed to be due to an ambiguous initial fricative in the audio prompt for 'same' that was used in wordlist enrw111. This left 298 productions of /s/ and 300 of /ʃ/ for acoustic analysis.

In the Japanese data, six productions were not annotated because the participant's repetition overlapped the audio prompt that was still playing in the background. These six omitted productions included one token each of /se/, /sɑ/, /so/, and /ɕo/ and two tokens of /ɕi/. This left 237 productions of both /s/ and /ɕ/ for acoustic analysis.

3.2.5 Peak ERB$_N$-number trajectories

For each annotated sibilant production, the times of frication onset and fricative-vowel boundary were used to define 17 20-ms analysis intervals that were spaced evenly across its duration, such that the first interval was left-aligned with the frication onset and the last, right-aligned with the fricative-vowel boundary. The amount of overlap or separation between consecutive intervals, thus, depended on the duration of the sibilant, which ranged from 75.088 ms to 382.197 ms in English, and from 62.663 to 264.955 in Japanese. For the English data, interval spacing ranged from 16.76 ms overlap to 1.31 ms separation ($M = 10.04$ ms overlap, $SE = 2.61$ ms), and for the Japanese data, from 17.33 ms to 4.69 ms overlap ($M = 12.42$ ms overlap, $SE = 1.86$ ms).

The waveform of each analysis interval was read from the source wave file into an R programming environment, and its spectrum was estimated by computing a multitaper spectrum (MTS) (Thomson, 1982), using parameter values $K = 8$ and $nW = 4$. The MTS estimate was then passed through a bank of 361 fourth-order gammatone filters, which modeled the differential frequency selectivity of the auditory system. The center frequencies $\{\omega_c\}$ in Hz of the gammatone filters were chosen such that their projections onto the ERB$_N$-number scale were evenly spaced, ranging from 3 to 39, with 0.1 ERB$_N$-number spacing between adjacent filters. For a given filter with center frequency ω_c in Hz, its bandwidth was set to $1.019 \times \text{ERB}_N(\omega_c)$; its phase, to zero; and its amplitude, such that its frequency response had a maximum of one. From the excitation pattern output by this gammatone filter bank, the center frequency on the ERB$_N$-number scale of the channel with the greatest excitation was determined. This frequency is referred to as *peak ERB$_N$-number*.

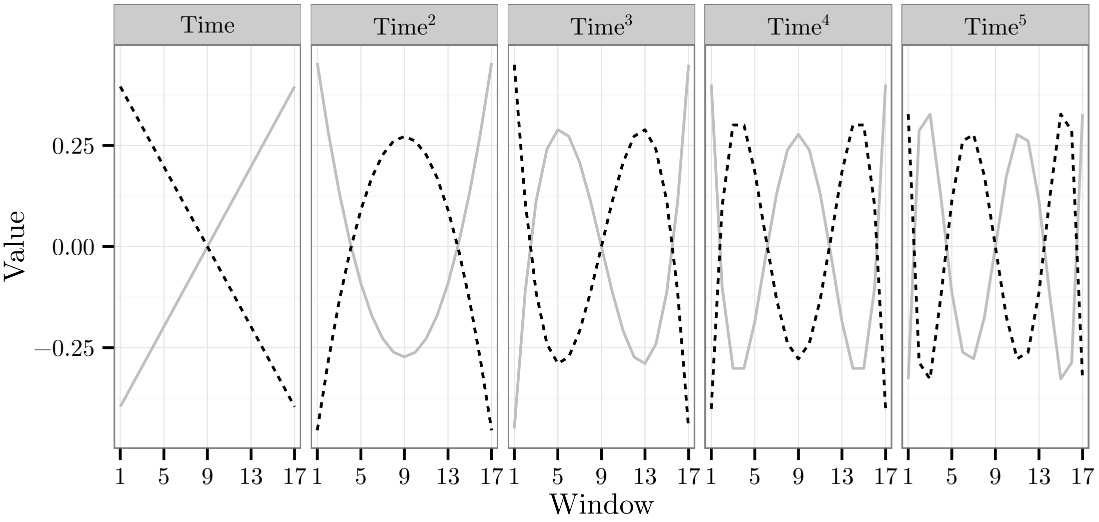
In this way, peak ERB$_N$-number was computed for each analysis interval of a sibilant production; the resulting sequence of 17 peak ERB$_N$-number values is referred to as a *peak*

ERB_N-number trajectory. The spectral dynamics of a sibilant production were represented by its peak ERB_N-number trajectory.

3.2.6 Growth-curve analysis of peak ERB_N-number trajectories

The peak ERB_N-number trajectories were analyzed with fifth-order orthogonal polynomial growth curve models (Mirman et al., 2008; Mirman, 2014), which are a special class of linear mixed-effects model where the effect of time on the observed variable is modeled as a linear combination of polynomial functions of time. During model fitting, coefficients for these functions are estimated. The sign and magnitude of these coefficients detect shape characteristics of the observed peak ERB_N-number trajectories and determine the shape of the peak ERB_N-number trajectory predicted by the fitted model. Examples of each polynomial function, scaled with a positive and negative coefficient, are shown in Fig. 3.1. From these plots it is clear that the linear power of time indicates an overall increase (positive coefficient) or decrease (negative coefficient) in peak ERB_N-number across the duration of the fricative; the quadratic power, whether the trajectory is convex (positive coefficient) or concave (negative coefficient); and the higher powers of time, the presence of minor inflections and local maxima or minima. Because the extremes of the quartic power point in the same direction, it can, along with the quadratic power, affect the convexity or concavity of the predicted trajectory. Conversely, because the tails of the cubic and quintic power point in opposite directions, they can induce asymmetries in the tails of the predicted trajectory. In a fitted model, the coefficient estimates were determined to be significant or insignificant by estimating a 95% Wald confidence interval; if the interval did not cover zero, then the coefficient was considered considered significant.

Figure 3.1: The orthogonal polynomial functions used in the growth curve models. In each panel, the solid grey line denotes the function whose coefficient is $\beta = 1$, and the dashed black line denotes the same function with coefficient $\beta = -1$.



3.3 Results

3.3.1 Contrast between English /s/ and /ʃ/

To determine how the contrast between English /s/ and /ʃ/ is realized in the peak ERB$_N$-number trajectories produced by speakers of either sex, the trajectories estimated from the English-speaking adults' productions were pooled across vowel context, and a growth curve model was fitted to them. The model's fixed-effects structure was built up following a forward, stepwise selection protocol, where the base model included only the intercept term. Effects for consonant, talker sex, polynomial time, and interactions between a time-factor and the other factors were added to the model only if a likelihood ratio test found that the model fit was significantly improved. To account for the repeated measures design of the word-repetition task, in which each participant produced multiple tokens of either sibilant, the consonant-contrast models included random effects of intercept and of each power of time, for participant and for consonant within participant. The random-effects structure

Table 3.3: Fixed-effects structure of the consonant-contrast model fitted to the adult English speakers' sibilant productions. Italicized fixed effects denote the terms used in the base vowel-context models for English /s/ and /ʃ/ (see §5.3.1 and §5.3.3, respectively).

	Likelihood ratio test		
Fixed effect	**Deg. freedom**	**χ^2 statistic**	**p-value (<)**
Consonant	1	69.98	0.001
Sex	1	18.56	0.001
Consonant × Sex	1	12.87	0.001
Time	5	103.53	0.001
Time × *Sex*	1	9.68	0.01
Time^2	7	1333.30	0.001
Time^2 × Consonant	1	14.86	0.001
Time^3	9	394.01	0.001
Time^3 × Consonant	1	8.77	0.01
Time^3 × *Sex*	1	13.55	0.001
Time^4	11	247.93	0.001
Time^4 × Consonant	1	18.84	0.001
Time^4 × Consonant × Sex	2	10.27	0.01
Time^5	13	111.89	0.001
Time^5 × *Sex*	1	3.86	0.05
Time^5 × Consonant × Sex	2	9.42	0.01

was augmented during the model-building protocol, such that a power of time was added to it only if that time-factor was also included in the model's fixed effects. The fitted model is referred to as a *consonant-contrast model*.

Table 3.3 shows the fixed effects of the consonant-contrast model fitted to the peak ERB$_N$-number trajectories of the English-speaking adults' productions of /s/ and /ʃ/. The results of the likelihood ratio tests, by which these effects were added to the model, are also reported there. The results reported on the first row correspond to the likelihood ratio test between the base model and the model whose only fixed effects were the intercept and consonant. For all other rows, the results reported on row r correspond to the likelihood ratio test that compared the model that comprised the fixed effects listed up through row $r-1$, to the model that included the effects up through row r.

The fixed-effects coefficients that of the fitted model are shown in Table 3.4. By setting to

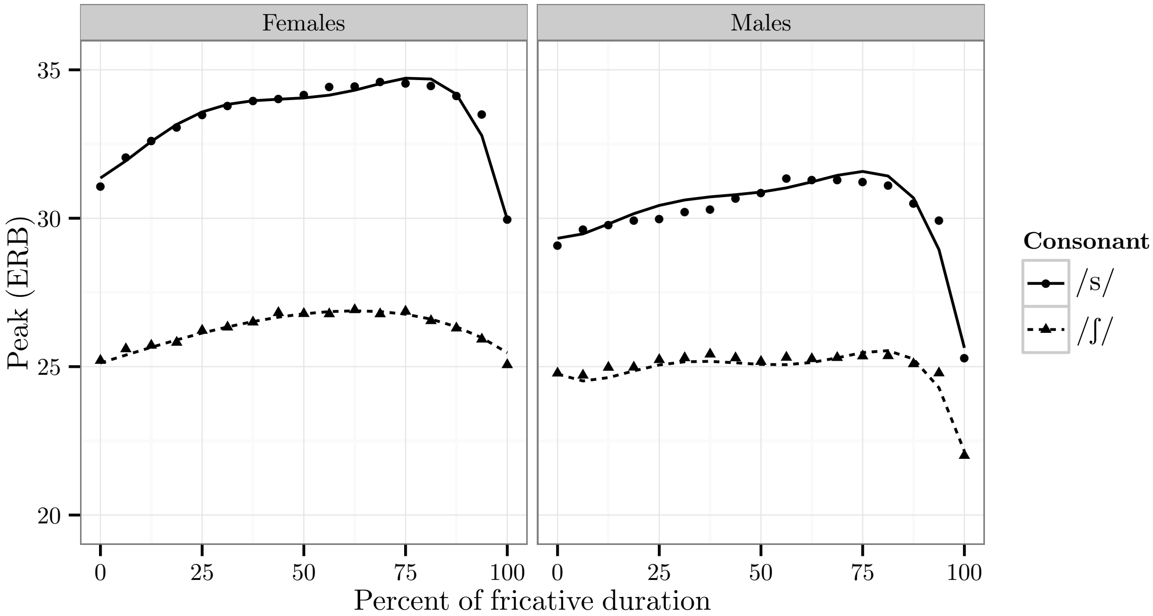
Table 3.4: Fixed-effects coefficients of the consonant-contrast model fitted to the adult English speakers' sibilant productions. The reference levels of the consonant and sex factors were /s/ and females, respectively. Shaded rows indicate fixed-effects coefficients whose estimate was not significantly different from zero, as determined by a 95% Wald confidence interval.

Fixed effect term	**Coeff. ($\hat{\beta}$)**	**Std. error**
Intercept	33.400	0.349
Consonant[/ʃ/]	−7.150	0.319
Sex[Males]	−3.156	0.490
Consonant[/ʃ/] × Sex[Males]	1.763	0.436
Time	0.964	0.327
Time × Sex[Males]	−1.624	0.445
Time^2	−4.346	0.480
Time^2 × Consonant[/ʃ/]	2.318	0.484
Time^3	−1.687	0.303
Time^3 × Consonant[/ʃ/]	1.240	0.335
Time^3 × Sex[Males]	−1.126	0.293
Time^4	−1.912	0.295
Time^4 × Consonant[/ʃ/]	1.831	0.320
Time^4 × Consonant[/s/] × Sex[Males]	−0.060	0.378
Time^4 × Consonant[/ʃ/] × Sex[Males]	−1.128	0.378
Time^5	−0.976	0.264
Time^5 × Sex[Males]	0.001	0.321
Time^5 × Consonant[/ʃ/] × Sex[Females]	0.948	0.300
Time^5 × Consonant[/ʃ/] × Sex[Males]	−0.041	0.299

zero the values of all the random effects, predictions for the peak ERB$_N$-number trajectories of /s/ and /ʃ/ can be made on just the fixed effects of the fitted consonant-contrast model. These predictions represent the peak ERB$_N$-number trajectory for each sibilant and each sex "at the population level" (Bates et al., 2014, p. 36), and are shown in Fig. 3.2.

The fitted consonant-contrast model included significant effects of all powers of polynomial time. The positive effect of linear time ($\hat{\beta} = 0.964$, $SE = 0.327$, $CI = [0.323, 1.606]$) indicated that the trajectory tended to rise across the duration of either sibilant. The negative effect of quadratic time ($\hat{\beta} = -4.346$, $SE = 0.480$, $CI = [-5.286, -3.405]$) indicated that the peak ERB$_N$-number trajectories followed a concave downward curve. The

Figure 3.2: Predicted peak ERB$_N$-number trajectories for the English-speaking adults' consonant-contrast model. Data means are shown as points.



negative effect of quartic time ($\hat{\beta} = -1.912$, $SE = 0.295$, $CI = [-2.490, -1.333]$) further contributed to the concave curvature and added inflections within the middle half of the trajectory. The effects of cubic ($\hat{\beta} = -1.687$, $SE = 0.303$, $CI = [-2.281, -1.093]$) and quintic time ($\hat{\beta} = -0.976$, $SE = 0.264$, $CI = [-1.493, -0.458]$) shifted the global maximum of each trajectory to a point later than its midpoint and contributed to the asymmetry in the slope of the left and right tails of the trajectory.

Effects of consonant were estimated with /s/ as the reference level. A significant negative effect of consonant on the intercept term ($\hat{\beta} = -7.150$, $SE = 0.319$, $CI = [-7.774, -6.525]$) indicated that the peak ERB$_N$-number was lower for /ʃ/ than for /s/ on average. A significant positive effect of consonant on quadratic time ($\hat{\beta} = 2.318$, $SE = 0.484$, $CI = [1.369, 3.266]$) indicated that the peak ERB$_N$-number trajectory of /ʃ/ was less concave than that of /s/. Finally, the effects of consonant on cubic ($\hat{\beta} = 1.240$, $SE = 0.335$, $CI = [0.583, 1.896]$) and quartic ($\hat{\beta} = 1.831$, $SE = 0.320$, $CI = [1.203, 2.459]$) time were significant and positive. Because these effects of consonant on the higher powers of time are positive, they reduce the magnitude of the (negative) cubic and quartic time effects, indicating that, relative to /s/, the peak ERB_N-number trajectory of /ʃ/ has an earlier global maximum and shallower inflections, respectively. The reduced concavity and suppressed inflections, due to effects of consonant on polynomial time, suggest that peak ERB_N-number varies less across the time course of /ʃ/ than of /s/.

Effects of sex were estimated relative to the females' productions. On the intercept term, there was a significant negative effect of sex ($\hat{\beta} = -3.156$, $SE = 0.490$, $CI = [-4.116, -2.195]$) and a significant positive interaction between consonant and sex ($\hat{\beta} = 1.763$, $SE = 0.436$, $CI = [0.909, 2.617]$). Together these terms suggest that at frication midpoint, the male speakers produced /s/ and /ʃ/ with lower peak ERB_N-number and with less acoustic differentiation.

3.3.2 Contrast between Japanese /s/ and /ɕ/

The fitted consonant-contrast model for the Japanese sibilants comprised the fixed effects listed in Table 3.5, and the fixed-effects coefficients of the fitted model are listed in Table 3.6. Predictions on these fixed effects are shown in Fig. 3.3. For both the males' and the females' productions of /s/ and /ɕ/, the peak ERB_N-number trajectory increased throughout the beginning of the fricative before reaching a global maximum near 25% of its duration, at which point it decreased until the temporal midpoint, after which it rose to a local maximum at 75% of the fricative duration and then fell steeply across the final quarter of the consonant's time course.

These commonalities in the shape of the peak ERB_N-number trajectories across consonant and talker sex are determined by the significant effects of time in the fitted model. Specifically, the Japanese consonant-contrast model included significant negative effects of linear ($\hat{\beta} = -7.209$, $SE = 0.725$, $CI = [-8.629, -5.789]$) and quadratic time ($\hat{\beta} = -9.542$, $SE = 0.864$, $CI = [-11.236, -7.849]$), which indicate, respectively, that the trajectories

Table 3.5: Fixed-effects structure of the consonant-contrast model fitted to the adult Japanese speakers' sibilant productions. Shaded rows denote the terms used as fixed effects in the base vowel-context models for Japanese /s/ and /ɕ/.

	Likelihood ratio test		
Fixed effect	**Deg. freedom**	**χ^2 statistic**	***p*-value (<)**
Consonant	1	20.88	0.001
Sex	1	7.69	0.01
Time	5	942.09	0.001
Time × Consonant	1	10.04	0.01
Time^2	7	2217.40	0.001
Time^2 × Consonant	1	32.281	0.001
Time^2 × Sex	1	4.77	0.05
Time^2 × Consonant × Sex	1	6.13	0.05
Time^3	9	647.08	0.001
Time^3 × Consonant	1	14.34	0.001
Time^3 × Sex	1	5.89	0.05
Time^4	11	590.43	0.001
Time^4 × Consonant	1	7.48	0.01
Time^4 × Sex	1	12.55	0.001
Time^5	13	231.71	0.001
Time^5 × Consonant	1	4.66	0.05
Time^5 × Sex	1	6.16	0.05

decreased across their midpoints and that their overall shapes approximated a concave downward curve. Significant effects of higher powers of polynomial time introduced the two inflection points in the middle half of the trajectory, so that a local minimum occurred near the temporal midpoint (cubic: $\hat{\beta} = -4.961$, $SE = 0.533$, $CI = [-6.007, -3.916]$; quartic: $\hat{\beta} = -3.438$, $SE = 0.533$, $CI = [-4.483, -2.393]$; quintic: $\hat{\beta} = -2.550$, $SE = 0.417$, $CI = [-3.367, -1.732]$).

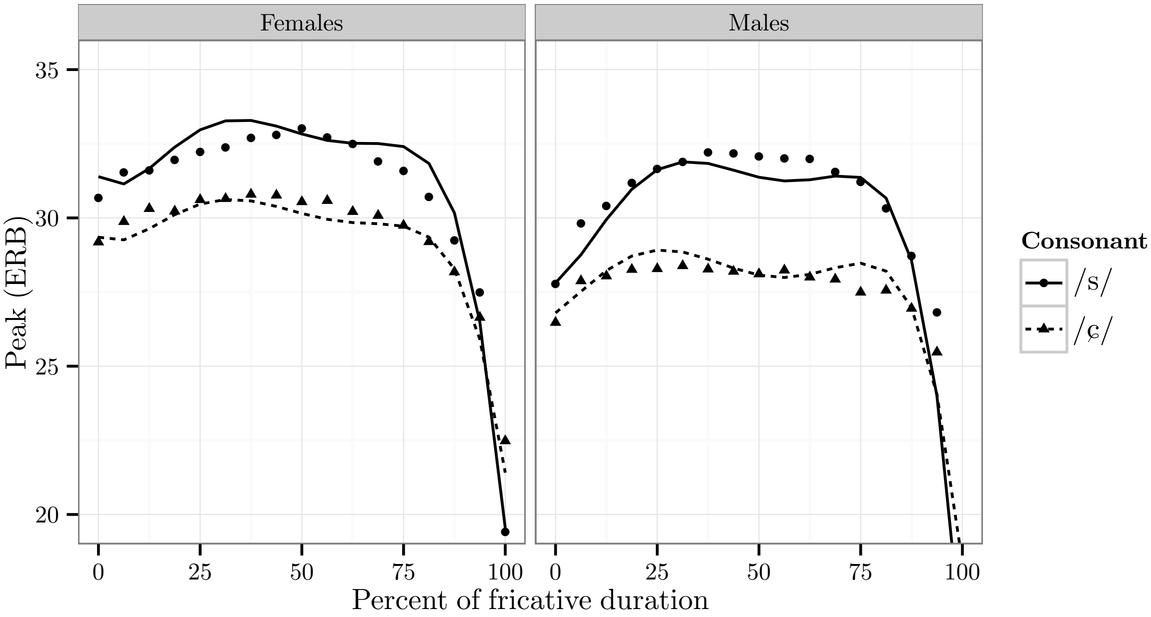
Effects of consonant were estimated with /s/ as the reference level. A significant negative effect of consonant on the intercept ($\hat{\beta} = -2.077$, $SE = 0.351$, $CI = [-2.764, -1.389]$) indicated that on average peak ERB$_N$-number was lower for /ɕ/. The effect of consonant on each power of time was significant and positive (linear: $\hat{\beta} = 1.950$, $SE = 0.555$, $CI = [0.863, 3.037]$; quadratic: $\hat{\beta} = 3.369$, $SE = 0.635$, $CI = [2.124, 4.613]$; cubic: $\hat{\beta} = 1.991$,

Table 3.6: Fixed-effects coefficients of the consonant-contrast model fitted to the adult Japanese speakers' sibilant productions. The reference levels of the consonant and sex factors were /s/ and females, respectively. Shaded rows indicate fixed-effects coefficients whose estimate was not significantly different from zero, as determined by a 95% Wald confidence interval.

Fixed effect term	**Coeff. ($\hat{\beta}$)**	**Std. error**
Intercept	31.184	0.456
Consonant[/ɕ/]	−2.077	0.351
Sex[Males]	−1.782	0.592
Time	−7.209	0.725
Time × Consonant[/ɕ/]	1.950	0.555
Time^2	−9.542	0.864
Time^2 × Consonant[/ɕ/]	3.369	0.635
Time^2 × Sex[Males]	−2.950	1.097
Time^2 × Consonant[/ɕ/] × Sex[Males]	2.273	0.862
Time^3	−4.961	0.533
Time^3 × Consonant[/ɕ/]	1.991	0.447
Time^3 × Sex[Males]	−0.524	0.482
Time^4	−3.438	0.533
Time^4 × Consonant[/ɕ/]	1.137	0.396
Time^4 × Sex[Males]	−1.734	0.575
Time^5	−2.550	0.417
Time^5 × Consonant[/ɕ/]	0.871	0.395

$SE = 0.447$, $CI = [1.115, 2.866]$; quartic: $\hat{\beta} = 1.137$, $SE = 0.396$, $CI = [0.360, 1.913]$; quintic: $\hat{\beta} = 0.871$, $SE = 0.395$, $CI = [0.098, 1.645]$). These effects reduced the magnitude of the effects of polynomial time on /ɕ/. Specifically, the effect on linear time indicated that peak ERB$_N$-number decreased less steeply overall for /ɕ/ than for /s/. The effect on quadratic time indicated that the peak ERB$_N$-number trajectory was less concave for /ɕ/, which may have been primarily due to differences across the final quarter of the time course, during which peak ERB$_N$-number dropped noticeably more for /s/. Finally, the effects on the higher powers of time reduced the magnitude of the main effects of cubic, quartic, and quintic time, the most noticeable consequence of which seems to be a reduction in the asymmetry between the left and right tails of the trajectory of /ɕ/, relative to that of /s/. That is, the slope of the final, relative to the initial, quarter of the trajectory is

Figure 3.3: Predicted peak ERB$_N$-number trajectories for the Japanese-speaking adults' consonant-contrast model.



greater for /s/ than for /ɕ/.

Effects of sex were estimated relative to the females' productions. A significant negative effect on the intercept ($\hat{\beta} = -1.782$, $SE = 0.592$, $CI = [-2.942, -0.622]$) indicated that peak ERB$_N$-number was lower on average in the males' productions. A significant negative effect of sex on quadratic time ($\hat{\beta} = -2.950$, $SE = 1.097$, $CI = [-5.101, -0.800]$) indicated that peak ERB$_N$-number followed a more concave trajectory for the men than for the women. Finally, a significant negative effect of quartic time ($\hat{\beta} = -1.734$, $SE = 0.575$, $CI = [-2.862, -0.607]$) indicated that the inflections were deeper in the males' productions, giving their trajectories a more pronounced bimodal shape. The effect on quartic time also indicated that the trajectories were more concave in the males' productions.

3.3.3 Cross-linguistic comparison of English and Japanese /s/

The cross-linguistic similarities in the articulatory postures that are assumed during sustained production of English and Japanese /s/ suggest that these two sounds have comparable peak ERB_N-number values across the middle half of their time course. To compare the "steady-state" spectral properties of /s/ across the two languages, the values that were estimated from windows 5 through 13 were pooled, and to them was fitted a linear mixed-effects model that included a fixed effect of language and random effects of intercept for participant. The fitted model revealed that peak ERB_N-number of /s/ is slightly lower in Japanese, but that this difference is not significant ($\hat{\beta} = -0.326$, $SE = 0.591$, $CI = [-1.483, 0.832]$).

Despite having comparable peak ERB_N-number values across their middle half, the trajectories for English and Japanese /s/ in Figs. 3.2 and 3.3, respectively, suggest significant cross-linguistic differences in peak ERB_N-number trajectory-shape for /s/. To determine whether these apparent differences are significant, the English and Japanese speakers' productions of /s/ were pooled and analyzed with a growth curve model. The fixed effects structure of the model was built up in the same manner as the consonant-contrast models, using a forward, stepwise selection protocol, except that effects of language, rather than of consonant, were considered. The fitted model is referred to as a *cross-linguistic comparison model*. The fixed-effects structure of the fitted cross-linguistic comparison model is shown in Table 3.7, and the fixed-effects coefficients of the fitted model are listed in Table 3.8. Predictions made on these fixed effects are shown in Fig. 3.4, which resemble the predictions for /s/ made in the two consonant-contrast models.

The fitted model included significant, negative effects of all powers of polynomial time other than linear. The negative effect of quadratic time ($\hat{\beta} = -4.558$, $SE = 0.801$, $CI = [-6.127, -2.989]$) indicated that peak ERB_N-number followed a concave trajectory in either language. The effects of cubic ($\hat{\beta} = -1.746$, $SE = 0.466$, $CI = [-2.659, -0.833]$), quartic ($\hat{\beta} = -1.977$, $SE = 0.515$, $CI = [-2.986, -0.968]$), and quintic time ($\hat{\beta} = -1.128$, $SE = 0.423$, $CI = [-1.956, -0.299]$) indicated that this concave trajectory was modified by minor

Table 3.7: Fixed-effects structure of the cross-linguistic comparison model fitted to the adult English and Japanese speakers' productions of /s/.

	Likelihood ratio test		
Fixed effect	**Deg. freedom**	χ^2 **statistic**	p**-value (<)**
Language	1	5.83	0.05
Sex	1	17.27	0.001
Language × Sex	1	5.37	0.05
Time	3	747.8	0.001
Time × Language	1	44.065	0.001
Time^2	4	2395.30	0.001
Time^2 × Language	1	35.09	0.001
Time^2 × Language × Sex	2	10.61	0.01
Time^3	5	705.87	0.001
Time^3 × Language	1	19.47	0.001
Time^3 × Sex	1	4.746	0.05
Time^4	6	543.98	0.001
Time^4 × Language	1	13.66	0.001
Time^4 × Sex	1	4.929	0.05
Time^4 × Language × Sex	1	8.95	0.01
Time^5	7	210.81	0.001
Time^5 × Language	1	10.182	0.01
Time^5 × Language × Sex	2	6.16	0.05

inflections.

Effects of language were estimated with English as the reference language. There was a significant effect of language on the intercept ($\hat{\beta} = -2.401$, $SE = 0.639$, $CI = [-3.654, -1.148]$), indicating that average peak ERB_N-number across the full duration of /s/ was lower in Japanese than in English. Since the previous analysis of just the middle half of /s/ found an insignificant difference between the two languages, this effect on the intercept may be due the extreme differences across the final quarter of /s/, where in Japanese the drop in peak ERB_N-number is much steeper than it is in English. The fitted model also included significant, negative effects of language on linear time ($\hat{\beta} = -7.656$, $SE = 0.878$, $CI = [-9.377, -5.936]$), indicating that peak ERB_N-number dropped more overall in the Japanese speakers' productions of /s/; on quadratic time ($\hat{\beta} = -4.680$,

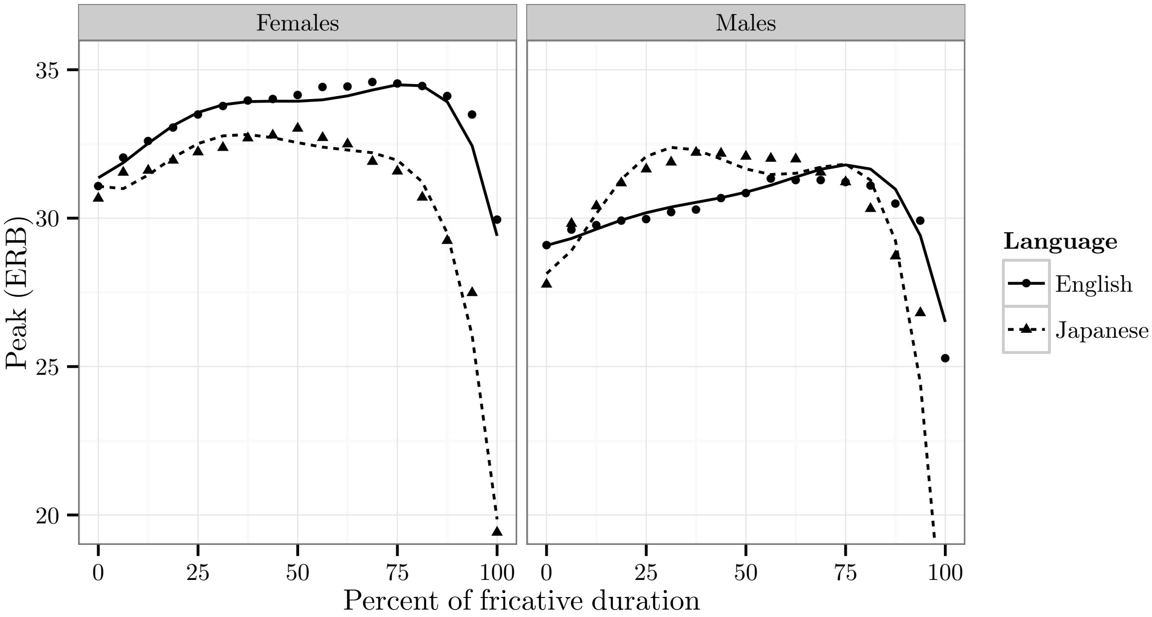
Table 3.8: Fixed-effects coefficients of the cross-linguistic comparison model fitted to the adult English and Japanese speakers' productions of /s/. The reference levels of the language and sex factors were English and females, respectively. Shaded rows indicate fixed-effects coefficients whose estimate was not significantly different from zero, as determined by a 95% Wald confidence interval.

Fixed effect term	**Coeff. ($\hat{\beta}$)**	**Std. error**
Intercept	33.246	0.453
Language[Japanese]	−2.401	0.639
Sex[Males]	−2.947	0.618
Language[Japanese] × Sex[Males]	1.842	0.867
Time	0.451	0.617
Time × Language[Japanese]	−7.656	0.878
Time^2	−4.558	0.801
Time^2 × Language[Japanese]	−4.680	1.122
Time^2 × Language[English] × Sex[Males]	0.626	1.082
Time^2 × Language[Japanese] × Sex[Males]	−3.558	1.099
Time^3	−1.746	0.466
Time^3 × Language[Japanese]	−2.978	0.579
Time^3 × Sex[Males]	−1.008	0.466
Time^4	−1.977	0.515
Time^4 × Language[Japanese]	−0.976	0.731
Time^4 × Sex[Males]	0.186	0.691
Time^4 × Language[Japanese] × Sex[Males]	−2.882	0.967
Time^5	−1.128	0.423
Time^5 × Language[Japanese]	−0.799	0.600
Time^5 × Language[English] × Sex[Males]	0.428	0.521
Time^5 × Language[Japanese] × Sex[Males]	−1.243	0.558

$SE = 1.122$, $CI = [-6.879, -2.481]$), indicating that the concavity of the peak ERB$_N$-number trajectory was more pronounced in Japanese; and on cubic time ($\hat{\beta} = -2.978$, $SE = 0.579$, $CI = [-4.112, -1.845]$), indicating greater asymmetry between the tails of the trajectory in Japanese.

Effects of sex were estimated relative to the females' productions. There was a significant, negative effect of gender on the intercept ($\hat{\beta} = -2.947$, $SE = 0.618$, $CI = [-4.159, -1.735]$) and a significant, positive interaction between language and sex ($\hat{\beta} = 1.842$, $SE = 0.867$, $CI = [0.142, 3.542]$), which together indicate that men tended to pro-

Figure 3.4: Predicted peak ERB$_N$-number trajectories from the cross-linguistic comparison model for /s/.



duce /s/ with lower peak ERB$_N$-number, but that the difference between the English and Japanese peak ERB$_N$ values was smaller in the men's, as opposed to the women's, productions.

3.4 Discussion

The significant effects of consonant on powers of polynomial time in the consonant contrast models for English and Japanese confirm the hypothesis that the peak ERB$_N$-number trajectories of a language's sibilant fricatives differ in shape. Furthermore, the significant effects of language on powers of time in the cross-linguistic comparison model for /s/ indicate that the shape of a sibilant's peak ERB$_N$-number trajectory is language-specific, not purely a matter of general kinematic constraints.

3.4.1 Potential artifacts of time normalization

The productions of the English sibilants ranged in duration from 75.088 to 382.197 ms. Likewise, the Japanese sibilant productions were not uniform in duration, ranging from 62.663 to 264.955 ms. However, in the foregoing analysis, the peak ERB_N-number trajectory of each production was represented simply as a 17-point sequence. Thus, this representation normalized the duration of each production, imposing uniformity where there was variability. Furthermore, due to this normalization, as the duration of a production decreased, the overlap between adjacent analysis intervals increased. The data overlapped by adjacent intervals were used to compute consecutive points in the peak ERB_N-number trajectory, which raises the possibility that differences in the spectral dynamics of different consonants appeared as artifacts of the time normalization.

In English, /ʃ/ exhibited a shallower peak ERB_N-number trajectory than /s/, while in Japanese the peak ERB_N-number trajectory of /ɕ/ was flatter and shallower than that of /s/. If the productions of English /ʃ/ or of Japanese /ɕ/ were significantly shorter than those of English /s/ or Japanese /s/, respectively, then the relative stability of the more posterior sibilants may simply have been due to multiple points in their peak ERB_N-number trajectories having been computed from the same data. To assess this possibility, a linear mixed-effects model of consonant duration, with random effects for participant and for consonant-within-participant, was built for each language. For either language, the fixed effect of consonant was estimated with /s/ as the reference level. For English, a significant positive effect of consonant ($\hat{\beta} = 10.835$, $SE = 3.147$, $CI = [4.666, 17.003]$ ms) indicated that the productions of /ʃ/ ($\hat{\mu} = 196.072$, $SE = 43.466$ ms) were longer than the /s/ productions ($\hat{\mu} = 185.155$, $SE = 44.658$ ms). Similarly, a significant positive effect of consonant was found for the Japanese data ($\hat{\beta} = 9.312$, $SE = 2.497$, $CI = [4.419, 14.205]$ ms), indicating that /ɕ/ ($\hat{\mu} = 146.044$, $SE = 27.355$ ms) had a greater duration than /s/ ($\hat{\mu} = 136.620$, $SE = 31.377$ ms). Thus, in either language, the effects of consonant on polynomial time, particularly those on linear and quadratic time, do not seem to be artifacts of time-normalization since the effects are in the opposite direction from what would be expected given the observed durational differences between /s/ and /ʃ/ in English and between /s/ and /ɕ/ in Japanese.

Similarly, a linear mixed-effects model of /s/ duration cross-linguistically reveals that /s/ is shorter ($\hat{\beta} = -48.342$, $SE = 9.201$, $CI = [-66.376, -30.308]$) in Japanese than in English. In the cross-linguistic comparison model, the effects of language on polynomial time served to increase the magnitude of their main effects. If these effects were due to cross-linguistic differences in sibilant duration, they would be expected to be in the opposite direction, to decrease the main effects of polynomial time, since Japanese /s/ is shorter than English /s/; hence, the cross-linguistic differences for /s/ do not seem to have been introduced by time normalization.

3.4.2 Articulatory kinematics

Since the observed variation in peak ERB_N-number across the time course of sibilant fricatives does not seem to be an artifact of time normalization, it is likely that the temporal changes in this psychoacoustic property reflect the kinematics of the articulatory system. In particular, since peak ERB_N-number is affiliated to the cavity that is most excited by a noise source, there are three aspects of the articulatory configuration, to which changes would, in theory, engender variation in peak ERB_N-number: the location of the constriction, the degree of constriction, and the height of the jaw.

First, the location of the constriction directly affects the size of the front cavity. As the constriction moves in the anterior direction, the front cavity decreases in length and volume; a move in the posterior direction would effect increases in length and volume. Assuming the constriction is narrow enough to decouple the back cavity, the excitation pattern primarily reflects the resonances of the front cavity; thus, increases in peak ERB_N-number would suggest a decrease in front cavity size, caused by an anterior movement of the constriction, and vice versa. From the concave downward trajectories predicted by the consonant-contrast models, the constriction would be expected to move first anteriorly and then posteriorly across the time course of each fricative; however, Iskarous et al. (2011) reported that during the production of English /s/, the location of the constriction is approximately constant across the first 80% of the fricative, and, across the final 20%, it moves anteriorly approximately 1 mm. Therefore, the observed peak ERB$_N$-number trajectories for English /s/ cannot be explained in terms of movement in the location of the constriction. Comparable articulatory data would need to be collected to determine whether movement in constriction location is plausible for the other three sibilants investigated above.

Second, the degree of constriction affects the coupling of the back cavity, which is excited by turbulence noise produced at the glottis. When coupled to the anterior portion of the oral tract, the back cavity's resonant frequencies appear as peaks in the excitation pattern, near the frequency of the second formant, $F2$, of the following vowel. In native adults' productions of English /s/ and /ʃ/ or Japanese /s/ and /ɕ/, the $F2$ frequency near the fricative-vowel boundary has been observed to be less than 2.5 kHz (22.96 ERB$_N$) (Li et al., 2009; McGowan and Nittrouer, 1988; Soli, 1981). Across the final 20% of English /s/, Iskarous et al. (2011) observed an increase of 0.5–1 mm in the distance between the tongue and the palate, suggesting that the constriction is released in anticipation of the following vowel. The trajectories predicted from the consonant-contrast models similarly showed a sharp decrease in peak ERB$_N$-number across the final 20–25% of each sibilant. For English, the drop was not extreme enough to indicate that the affiliation of peak ERB$_N$-number switched from the front to the back cavity in any of the conditions except for perhaps /ʃ/ when produced by a male. Even if peak ERB$_N$-number remained affiliated to the front cavity across the full duration of the English sibilants, the release of the constriction may nevertheless have decreased peak ERB$_N$-number by increasing the effective length of the front cavity at its posterior end. For Japanese, each trajectory dropped below 22.5 ERB$_N$ by the final analysis interval, suggesting that the back cavity was coupled by this point.

The release of the constriction may further contribute to the drop in peak ERB_N-number across the final portion of each fricative by reducing the flow velocity of the turbulent jet; thus, although there does not seem to be much direct evidence for back cavity coupling in the English sibilants, the observed drop in peak ERB_N-number across the end of the fricative may reflect a reduction in the degree of constriction.

Finally, the height of the jaw affects the cross-sectional area of the front cavity: a higher jaw entails a smaller cross-sectional area and a higher resonant frequency. Furthermore, jaw height determines the extent to which the lower incisors are positioned within the stream of the turbulent airflow exiting the constriction. When the turbulent jet collides with an obstacle positioned orthogonal to the airflow, high frequency noise is generated. Computational and mechanical models of sibilant fricatives have found that the presence of such a noise source, resulting from the jet's collision with an obstacle, like the lower incisors, is necessary to produce frication whose spectra match those of natural sibilants (Nakamura et al., 2011; Narayanan and Alwan, 2000; Nozaki, 2010; Shadle, 1985; Toda et al., 2010). Each trajectory predicted from the consonant-contrast models suggests that, across the time course of a sibilant, peak ERB_N-number follows a concave downward curve, modified with shallow inflections. If such changes are consequences of jaw height, then the jaw would be expected to rise and then fall during sibilant production. Such movement was observed by Iskarous et al. (2011); however, as they did not test powers of polynomial time beyond quadratic, they did not report any inflections in the trajectory of the jaw.

These relationships between peak ERB_N-number and the kinematics of the articulators suggest that, across the first 75–80% of the English and Japanese sibilants, peak ERB_N-number indicates the resonance of the front cavity, and variation within this interval is due primarily to movement of the jaw. Across the final 20–25%, peak ERB_N-number indicates the resonance of the front cavity in English, but either the front or back cavity in Japanese; variation within this interval may reflect either lowering of the jaw or the release of the constriction.

In the consonant-contrast model for English, there were significant, positive effects of consonant on quadratic, cubic, and quartic time, which led to predicted peak ERB$_N$-number trajectories that were flatter for /ʃ/ than for /s/. This raises the question of whether /ʃ/ is produced with a more restricted range of jaw motion or produced in such a way that makes peak ERB$_N$-number more resistant to variation in jaw height. Using a computational model of the vocal tract, Toda and Maeda (2006) simulated its transfer function in response to turbulence noise sources, while varying the size of the front cavity and the constriction. They used the resonant frequencies of these simulated transfer functions to develop a map of how changes in front cavity size relate to changes in resonant frequency. When the subjects whose vocal tracts were imaged in Toda and Honda (2003) were placed on this map, according to their articulatory configuration for sustained /ʃ/, they all fell in regions where increases or decreases in front cavity size would not much affect the resonant frequencies of the transfer function (see Toda and Maeda, 2006, Fig. 8, p. 579). Their simulation did not place where the subjects' configurations for /s/ fell on this map; however, their results do suggest that resonant frequency becomes more sensitive to perturbations in front cavity size as it decreases.

3.4.3 Language specificity of gestural coproduction

Under the task-dynamic model of speech production (Fowler and Saltzman, 1993; Saltzman and Munhall, 1989), individual sounds are planned and produced as gestures involving certain articulators. Sequences of sounds are composed through gestural scores that determine how the individual gestures are coproduced. For example, the production of a sibilant-vowel sequence involves blending the gesture for the sibilant with the gesture for the vowel such that the articulators move smoothly as they execute the two. More concretely, the observation, by Iskarous et al. (2011), that the constriction is released at 80% of the duration of /s/ rather than being maintained across its full duration can be understood as a consequence of gestural coproduction.

One interpretation of the cross-linguistic differences between the peak ERB_N-number trajectories of English and Japanese /s/ is that they are due to language-specific differences in gestural coproduction: In Japanese, the gestures for the sibilant and vowel are composed in such a way that by the end of frication the linguapalatal constriction has been released enough for the back cavity to become the dominant resonating cavity; whereas, in English, this is not the case. Temporal variation in peak ERB_N-number, then, is epiphenomenal, falling out from the coordination of gestures.

Even if such a view is adopted, the importance of a sibilant's specific spectral dynamics is maintained from the perspective of language acquisition. If the coproduction of gestures is governed by language-specific principles, then they must be learned during acquisition. Furthermore, since an infant may have, at best, only partial access to the articulators of the adult caretakers—i.e., through the infant's vision of the adult's lip and jaw movement—the spectral dynamics and other time-varying properties of speech must be leveraged to learn fluent, language-specific gestural coordination.

Chapter 4

The development of contrastive spectral dynamics in sibilant fricatives

4.1 Introduction

4.1.1 Ages of acquisition of intelligible sibilants, cross-linguistically

The acquisition of the ability to produce intelligible sibilant fricatives is protracted in both English and Japanese, extending, in some cases, into early or middle childhood. In a large-scale cross-sectional survey of English-acquiring children aged three to nine years, Smit et al. (1990) found that the girls produced an intelligible /s/ at least 75% of the time from age 4;6 years;months onward, but the boys did not reach this level of performance until age 6;0. By age 7;0, the differences between the sexes in rate of /s/ intelligibility had disappeared, and by age 9;0 both produced /s/ intelligibly 90% of the time. For /ʃ/, the girls and boys surpassed the 75% intelligibility criterion by ages 4;0 and 5;0, respectively. By age 7;0, both sexes produced an intelligible /ʃ/ more than 90% of the time; however, 100% intelligibility still was not reached by age 9;0.

In a study of Japanese-acquiring three-year-old children, split into younger and older age groups, Yasuda (1970) found that in the older cohort, only 27.6% of the boys and 37.1% of the girls consistently produced /s/ correctly; whereas 68.9% of the boys and 71.4% of the girls produced /ɕ/ accurately. These proportions were even lower in the younger cohort of children. Additionally, a cross-sectional survey (Umebayashi and Takagi, 1965, cited in Yasuda, 1970) of three- through five-year-old children found that by age 4;6, 75% of children consistently produced /ɕ/ accurately; whereas, the same level of accuracy for /s/ was not achieved until the end of the fifth year.

4.1.2 Language-specific differentiation and tuning of sibilant fricatives

Nittrouer (1995, p. 521), citing Browman and Goldstein (1989), argues that the development of adult-like productions of sibilant fricatives, or any speech sound for that matter, involves "the differentiation and tuning of individual gestures." She continues,

> For example, a young child may produce one or two roughly specified fricative constrictions. Eventually, this limited set must be differentiated into all the precisely specified fricative constrictions of the child's native language. A child's earliest gestures must also be tuned to meet the specific requirements of the native language.

A number of studies have examined the spectral properties of young English- and Japanese-acquiring children's productions of target sibilant fricatives, in order to better understand the time course of their differentiation and tuning during language acquisition. From these studies, reviewed below, it is clear that such differentiation proceeds in a language-specific manner.

As was discussed in §3.1.2, native adults' productions of English /s/ and /ʃ/ can be classified, with high accuracy, using just a single measure, e.g. centroid or peak frequency, of where energy is concentrated in the spectrum of the frication noise. On the other hand, accurate classification of the Japanese sibilants required an additional spectral feature that was estimated near the onset of the following vowel, like the frequency of $F2$ (Li et al., 2009). These results suggest that the spectral features that signal the contrast between sibilants in English or Japanese are language-specific.

Given this specificity in the spectral features in which adult speakers contrast the sibilants of the two languages, it is perhaps unsurprising that studies of the spectral properties of English- and Japanese-acquiring children's productions of sibilant fricatives suggest that they are differentiated only in terms of those features that are contrastive in their ambient language. In a cross-linguistic study, Li (2012) analyzed the development of sibilant contrast in two- through five-year-old children who were natively acquiring either English or Japanese. From each production, the centroid frequency at the midpoint of frication and the $F2$ frequency at vowel onset were computed. For each sibilant, in each language, the children's mean centroid and vowel-onset $F2$ values were regressed against their age. For English /s/ and /ʃ/, it was found that when fitted to the mean centroid values, the regression lines for the two sibilants diverged as age increased, suggesting a developmental increase in the degree of contrast; however, when fitted to the mean $F2$ values, the regression lines did not diverge from each other. The regression lines for Japanese /s/ and /ɕ/ diverged from each other regardless of whether they were fitted to the centroid or $F2$ estimates. Thus, in both languages, the sibilant contrast developed only in terms of the spectral features that distinguish the sibilants in adults' productions.

Furthermore, Li's (2012) results also suggest that the differentiation of sibilants in terms of centroid frequency is not the same in English and Japanese, even though this feature helps distinguish the two sibilants in both languages. The youngest children's productions were undifferentiated in terms of centroid frequency; however, the region of centroid space occupied by these undifferentiated sibilant categories differed cross-linguistically. In particular, the Japanese-acquiring two-year-olds' productions were more similar to the adults' productions of /ɕ/ than of /s/. The opposite was found in English, the two-year-olds' centroid values being closer to those of the adults' /s/ than their /ʃ/. Additionally, the regression lines diverged from each other differently in each language. In Japanese, the regression line for /ɕ/ remained approximately flat, while the line for /s/ increased significantly, suggesting that the contrast between these sibilants develops, in part, due to differences in how children produce /s/ as they get older. In English, the regression line for /s/ increased only slightly, while the one for /ʃ/ decreased significantly, suggesting that the acquisition of the sibilant contrast in this language is due more to changes in how speakers produce /ʃ/ as they age.

In her regression analysis of the English-acquiring children, Li (2012) found that the mean centroid of /s/ and /ʃ/ began to differ significantly from each other around 35 months of age. Although this difference suggests that the sibilant contrast becomes acoustically realized at a young age, other studies indicate that preschool-aged children's productions of sibilants are not yet adult-like in their spectral properties. For example, it is possible for young children to produce /s/ and /ʃ/ in such a way that a statistical analysis evinces an acoustic contrast that is nonetheless imperceptible to adult listeners[1] (Li et al., 2009). Furthermore, Weismer et al. (1980) reported that normally-articulating preschool age children produced /s/ with a spectral shape different from that observed in adults' productions.

Moreover, differences between children's and adults' productions of the English sibilants can be found into middle childhood and adolescence. For example, the difference between centroid frequency of /s/ and /ʃ/ has been found to be greater in adults' productions than in the productions of either seven-year-olds (Nittrouer, 1995; Nittrouer et al., 1989) or eight- and nine-year-olds (Fox and Nissen, 2005). Romeo et al. (2013) investigated the structure of sibilant categories, beyond just their mean centroid, in a group of adults and adolescents between the ages of 9 and 14. Each speaker's /s/ and /ʃ/ categories were represented as a sample of centroid values, and three measures of the sibilant categories and their contrast were computed: cross-category distance, defined as the distance between the mean of each category; mean category dispersion, defined as the mean standard deviation of the two sibilant categories; and discriminability.[2] The adolescent speakers were grouped into three cohorts, each spanning two years. It was found that each cohort tended to produce

[1] This phenomenon is known as *covert contrast* in the literature and has been found for other contrasts in English. Covert contrast has also been observed in young Japanese-acquiring children's productions of /s/ and /ɕ/ (Tsurutani, 2004).

[2] The *discriminability*, $d_{(a)}$, of two categories X and Y, whose mean and variance are μ_X and σ_X^2, and μ_Y and σ_Y^2, respectively, is defined as $d_{(a)} = \dfrac{|\mu_X - \mu_Y|}{\sqrt{(\sigma_X^2 + \sigma_Y^2)/2}}$ (Simpson and Fitter, 1973).

the English sibilants with greater cross-category distance than the adults did; however, this increased separation in means was accompanied by an increase in mean category dispersion. The net effect of the adolescents producing /s/ and /ʃ/ with greater separation but also greater dispersion, was that the discriminability of the two categories tended to increase across the age groups, with the adults' categories being the most discriminable. These results suggest a continual tuning, with respect to centroid frequency, of the English sibilants /s/ and /ʃ/ and the contrast in which they participate.

Finally, returning to the results of Li (2012), there is evidence that the differentiation and tuning of sibilants in a given language may proceed at different rates for different spectral features. In her regression analyses of the Japanese-acquiring children's productions, the age at which the regression lines for /s/ and /ɕ/ became significantly different from one another depended on whether the models were fitted to $F2$ or centroid data. In terms of vowel-onset $F2$, /s/ and /ɕ/ were differentiated by approximately 40 months; whereas, for centroid frequency, differentiation was delayed until close to 50 months.

4.1.3 Purpose and hypotheses

Building on the findings of chapter 3, which suggest that native adult speakers of English and Japanese contrast the sibilant fricatives of their language in terms of dynamic spectral properties, the analyses in the current chapter extend the results reviewed in the previous section by examining how the differentiation of sibilant fricatives, in terms of their peak ERB$_N$-number trajectories, develops cross-linguistically in children between the ages of two and five. While the literature reviewed in §4.1.2 characterized the development of sibilant contrast in terms of spectral properties other than peak ERB$_N$-number or peak frequency, it still offers insight into how contrastive dynamic spectral properties might develop in each language.

Li (2012) observed that the spectral properties of the two-year-olds' undifferentiated sibilant category was more /s/-like in English, but more /ɕ/-like in Japanese. The analyses in this chapter examined whether a similar pattern obtains for the shape properties of the peak ERB$_N$-number trajectories of the English and Japanese sibilants. That is, since Li (2012) analyzed centroid and $F2$ from a single analysis interval, it is possible that, even if the two-year-olds produce undifferentiated sibilants whose static spectral properties are similar to the adults' /s/ in English, and more like /ɕ/ in Japanese, the articulatory gestures made by two-year-olds are nothing like those of adults. This gestural difference would be better revealed through the temporal variation in spectral properties. The following hypotheses are made regarding how the differentiation of sibilants develops with respect to dynamic aspects of peak ERB$_N$-number trajectory: First, in English, the peak ERB$_N$-number trajectories estimated from the youngest children's productions will not differ in terms of their curvature, and their concavity will be more similar to that of the adults' productions of /s/ than of /ʃ/. Second, in Japanese, the youngest children's peak ERB$_N$-number trajectories will differ neither in terms of their concavity, nor in terms of the asymmetry in their tails; furthermore, with respect to these shape properties, the children's peak ERB$_N$-number trajectories will more closely resemble the adults' /ɕ/ productions than their /s/ productions.

It is also hypothesized that there will be a developmental increase in the extent to which the sibilants of either language are differentiated in terms of dynamic spectral features. For the English-speaking adults' productions, it was argued that temporal variation in peak ERB$_N$-number reflected jaw movement and the release of the linguapalatal constriction, but that peak ERB$_N$-number still likely reflected the resonance of the front cavity. Furthermore, the centroid values computed from English-acquiring children's productions of /s/ and /ʃ/ exhibit statistically significant differences from a young age. These facts suggest that English /s/ and /ʃ/ are differentiated primarily in terms of the size of the front cavity. The gestures for /s/ and /ʃ/, then, might be first differentiated in terms of where the linguapalatal constriction is made in the oral cavity, and later fine-tuned in terms of increasingly skilled control of the movement of the articulators during the formation, maintenance, and release of this constriction. Thus, it is hypothesized that children's productions of English /s/ and

/ʃ/ are first differentiated in terms of average peak ERB_N-number value across the duration of the fricative, and then tuned in terms of less and less erratic variation in the dynamic properties of each sibilant's peak ERB_N-number trajectory.

In the Japanese-speaking adults' productions of /s/ and /ɕ/, the affiliation of peak ERB_N-number appeared to switch from the front cavity to the back cavity across the final 20–25% of the sibilants' duration. Additionally, Japanese-acquiring children distinguish /s/ and /ɕ/ in terms of vowel-onset $F2$ frequency at an earlier age than they make this distinction in terms of the centroid frequency of the frication. Together these observations suggest that the gestures for Japanese /s/ and /ɕ/ might first be differentiated in terms of the length of the linguapalatal constriction, rather than in terms of the front cavity size: the length of the constriction determines the size of the back cavity, whose resonances may appear in the spectrum or excitation pattern only once the constriction is released. Thus, it is hypothesized that the children's productions of Japanese /s/ and /ɕ/ are first differentiated in terms of effects on powers of polynomial time, reflecting the switch of peak ERB_N-number affiliation from the front to the back cavity, and later in terms of average peak ERB_N-number across the duration of the sibilants.

4.2 Method

4.2.1 Participants

English-acquiring children

Eighty-one English-acquiring children were recruited from daycare centers and preschools in the Columbus, OH metropolitan area. The distribution of these children across age and sex is shown in Table 4.1. The participants were screened for speech, language, and hearing disorders with parental report and a small battery of clinical assessments. These assessments included a hearing test either of 25 dB HL pure tones at 0.5, 1, 2, and 4 kHz or of otoacoustic emissions at 2, 3, 4, and 5 kHz; the *Goldman-Fristoe Test of Articulation*

Table 4.1: The distribution, across age and sex, of English-acquiring children, who participated in the repetition task.

	2-year-olds	**3-year-olds**	**4-year-olds**	**5-year-olds**	**Total**
Females	9	10	10	11	40
Males	11	10	11	9	41
Total	20	20	21	20	81

(GFTA-2) (Goldman and Fristoe, 2000), a test of articulatory accuracy for consonants; the *Expressive Vocabulary Test (EVT)* (Williams, 1997); and the *Receptive One-Word Picture Vocabulary Test (ROWPVT-2)* (Brownell, 2000). The articulation and vocabulary tests are all norm-referenced for each age group, allowing each child's raw score to be standardized relative to a large sample of age-matched peers. The norm-referenced scores for each test have a mean of 100 and a standard deviation of 15.

None of the parents reported any history of speech, language, or hearing disorders for the children. Each child passed a hearing test in at least one ear. Each child's standardized scores on the articulation and vocabulary tests were no less than 1.5 standard deviations below the norm-referenced means (GFTA-2: $M = 107.2$, $SD = 10.4$, range: $[83, 130]$; EVT: $M = 109.0$, $SD = 13.8$, range: $[79, 145]$; ROWPVT-2: $M = 108.8$, $SD = 13.2$, range: $[83, 145]$).

Japanese-acquiring children

Seventy-eight Japanese-acquiring children were recruited in Tokyo, Japan to participate in the study. The children were screened for speech and language disorders with parental report. At the time of test, none of the parents reported any history of disorders. They were also screened for hearing disorders with the same test of otoacoustic emissions that was administered to the English-acquiring children. The distribution of the participants across age and sex is shown in Table 4.2.

Table 4.2: The distribution, across age and sex, of Japanese-acquiring children, who participated in the repetition task.

	2-year-olds	**3-year-olds**	**4-year-olds**	**5-year-olds**	**Total**
Females	9	10	9	9	37
Males	10	10	11	10	41
Total	19	20	20	19	78

4.2.2 Materials

Productions of the sibilant-initial real words listed in Tables 3.1 and 3.2 were elicited from the children using the same stimuli that were used for the adults.

4.2.3 Elicitation and recording procedure

The children completed the task in the presence of an adult experimenter, who controlled a software program, described in §3.2.3, that presented the audiovisual stimulus of each trial. The program allowed the experimenter to replay the audio stimulus of a trial if, for example, the child did not respond or if the response was either inaudible or masked by background noise. The experimenter sometimes interacted with the child to encourage their involvement in the task (e.g., "Great job!" or "What did the computer say?"). On rare occasions, the experimenter provided a live-voice example of the target word (e.g., "Can you say 'seashore?'").

4.2.4 Segmentation, transcription, and annotation of target productions

Because the elicitation procedure permitted multiple repetitions of a target to be recorded, and because children may substitute an incorrect phone for a target sibilant (e.g., produce [t] for target /s/), the recordings of the repetition task underwent two stages of annotation before the onset and offset of frication were marked. First, each recording was segmented by marking the boundaries of the intervals that followed either a presentation of the audio stimulus or a voice prompt from the experimenter. The child's response within each of these intervals was classified either as a repetition of the target word, as a production of a word other than the target, or as nonresponsive (e.g., no response or uncooperative groaning).

After segmentation, the repetitions of the target word were transcribed by a trained phonetician. The symbol set used in transcription included the full International Phonetic Alphabet (IPA). The atomic IPA symbols could be joined in series to denote a sequence of consonants; e.g., [θs] denoted frication where the place of articulation changed from interdental to alveolar. Additionally, two IPA symbols could be connected with an infix colon to denote a production that was intermediate between two canonical sounds. In such cases, the symbol to the left of the colon denoted the more dominant of the two sounds; e.g., [s:θ] denoted frication that was intermediate between [s] and [θ], but more like [s] than [θ], and vice versa for [θ:s]. Productions that were transcribed as a weak (e.g., /f, θ/) fricative were excluded from analysis for two reasons. First, the spectra of these sounds are typically more diffuse, lacking prominent spectral peaks in the high-frequency range (cf. Fox and Nissen, 2005; Hughes and Halle, 1956), which compromises the interpretability of a measure like peak ERB_N-number for weak fricatives. Second, in §3.4.2, it was argued that the variation in peak ERB_N-number across the first half of adults' productions of sibilant fricatives is partly due to changes in the turbulence noise produced from the turbulent air flow striking the lower incisors; however, the production of weak fricatives does not involve such a noise source.

The onset and offset of frication were annotated on the transcribed productions, using the criteria described in §3.2.4 to determine the location of these events.

Some of the 2430 target trials recorded from the English-acquiring children had to be excluded from analysis for various reasons: One trial was never elicited due to an experimenter error. In 52 trials, the child never repeated the target word because they either produced a different word (26 trials) or were nonresponsive (26 trials). One hundred eleven trials were not transcribed because there was too much background noise. One hundred seventy-two trials were not aligned for various reasons; e.g., the production was a stop substitution or a

Table 4.3: The distribution of transcribed, aligned sibilant productions, across age group, sex and consonant.

	2-year-olds		**3-year-olds**		**4-year-olds**		**5-year-olds**		**Total**	
	/s/	/ʃ/	/s/	/ʃ/	/s/	/ʃ/	/s/	/ʃ/	/s/	/ʃ/
Females	70	86	84	112	116	146	133	148	403	492
Males	89	90	92	104	119	147	110	122	410	463
Total	159	176	176	216	235	293	243	270	813	955

distortion, or it overlapped with the audio prompt. This left 2094 trials whose productions were transcribed and aligned. Of these, 268 were excluded because the production was transcribed either as a weak or voiced fricative, or as a sibilant fricative interrupted by a stop; five were excluded because the interval of sibilant frication was too long, greater than 500 ms; and three were excluded because the sibilant interval was too short, less than 60 ms. This left 1818 annotated sibilant productions, whose peak ERB$_N$-number trajectories could be analyzed. These productions were tabulated across consonant and vowel context for each participant. Some participants had to be excluded because they did not produce enough analyzable sibilants for their random effects to be estimated in the vowel-context models. Specifically, one two-year-old female (e2bt14fw: three trials), one two-year-old male (e2bt11mw: 14 trials), one three-year-old female (e3bt16fw: 12 trials), and two three-year-old males (e3bt19mw: two trials; e3bt20mw: 19 trials) were excluded. This left 1768 productions, from 76 participants, with which to build the consonant-contrast and vowel-context models for the English-acquiring children. The distribution of these productions across age group, gender, and consonant is shown in Table 4.3.

Of the 1872 target trials recorded for the Japanese-acquiring children, 204 were excluded because, for example, the production could not be either transcribed or aligned due to background noise, or the production was not transcribed as a sibilant. Another 76 trials were excluded because the sibilant produced by the child was either too short (less than 60 ms) or too long (more than 500 ms); consequently, two two-year-old participants (j2at05fw: seven trials; j2bt18mw: 12 trials), one of each gender, were excluded. Because they did not

Table 4.4: The distribution of transcribed, aligned sibilant productions, across age group, sex and consonant, for the Japanese-acquiring children.

	2-year-olds		**3-year-olds**		**4-year-olds**		**5-year-olds**		**Total**	
	/s/	/ɕ/	/s/	/ɕ/	/s/	/ɕ/	/s/	/ɕ/	/s/	/ɕ/
Females	50	54	100	103	103	98	104	106	357	361
Males	72	76	95	100	123	129	115	116	405	421
Total	122	130	195	203	226	227	219	222	762	782

produce enough sibilant productions for their random effects to be estimated in the vowel-context models, two two-year-old females (j2at06fw: 12 trials; j2bt23fw: eight trials), two two-year-old males (j2at11mw: 14 trials; j2at12mw: three trials), one three-year-old female (j3bt24fw: seven trials), and one three-year-old male (j3at05mw: four trials) were excluded. These exclusions left 1544 productions, from 70 participants to build the consonant-contrast and vowel-context models for the Japanese-acquiring children.

4.2.5 Peak ERB$_N$-number trajectories

A peak ERB$_N$-number trajectory was computed from each transcribed and annotated sibilant production, with the same method that was used for the adults (see §3.2.5). Again, the amount of overlap between consecutive analysis intervals depended on the duration of the production. For the English data, the spacing between intervals ranged from 17.49 ms overlap to 9.08 ms separation ($M = 8.27$ ms overlap, $SE = 4.15$ ms), and, for the Japanese data, from 17.29 ms overlap to 6.54 ms separation ($M = 11.59$ ms overlap, $SE = 3.06$ ms).

4.2.6 Community-norm consonant-contrast models

Because the differentiation of sibilant fricatives in a given language has been found to progress primarily in terms of spectral features that are contrastive in adults' productions (cf. Li, 2012), the development of sibilant contrast in English and Japanese was evaluated with respect to only those terms that were included in the adults' consonant-contrast models. For example, the fixed-effects structure of the model fitted to the English-acquiring children's productions included all and only those terms that were included in the English-speaking adults' consonant-contrast model, listed in Table 3.3. Thus, the development of the English sibilant contrast was investigated only in terms of effects of consonant on intercept and quadratic, cubic, and quartic time. On the other hand, the development of the Japanese sibilant contrast was evaluated in terms of effects of consonant on the intercept and each power of polynomial time. In this way, the adults' consonant-contrast models were used as landmarks that delineate the community norms toward which the children progress.

4.3 Results

4.3.1 The development of sibilant contrast in English-acquiring children

The English consonant-contrast model (cf. Table 3.3) was fitted to all the children's productions, pooled across age group. The fixed-effects coefficients for this fitted model are shown in the second column ('All') in Table 4.5. The fitted model included a significant, positive effect of linear time ($\hat{\beta} = 2.279$, $SE = 0.688$, $CI = [0.930, 3.627]$), indicating that peak ERB$_N$-number tended to increase across the children's productions of /s/ and /ʃ/. Significant, negative effects of quadratic ($\hat{\beta} = -10.147$, $SE = 0.545$, $CI = [-11.214, -9.080]$) and quartic time ($\hat{\beta} = -3.629$, $SE = 0.345$, $CI = [-4.305, -2.953]$) were also found, indicating that peak ERB$_N$-number followed a concave downward trajectory across the time course of either sibilant. Effects of consonant were evaluated relative to /s/, and were found to be significant on the intercept ($\hat{\beta} = -3.257$, $SE = 0.417$, $CI = [-4.074, -2.439]$), indicating that on average peak ERB$_N$-number is lower in productions of /ʃ/ than of /s/; and on quadratic ($\hat{\beta} = 2.752$, $SE = 0.526$, $CI = [1.722, 3.782]$) and quartic time ($\hat{\beta} = 1.163$, $SE = 0.414$, $CI = [0.352, 1.974]$), indicating that the peak ERB$_N$-number trajectory of /ʃ/ is less concave than that of /s/.

Table 4.5: Fixed-effects coefficients (and their standard errors, in parentheses) for the consonant-contrast models fitted to the English-acquiring children's sibilant productions. Each column denotes a single fitted model. Shaded cells denote coefficients whose estimate was not significantly different from zero in the fitted model, as determined by a 95% Wald confidence interval.

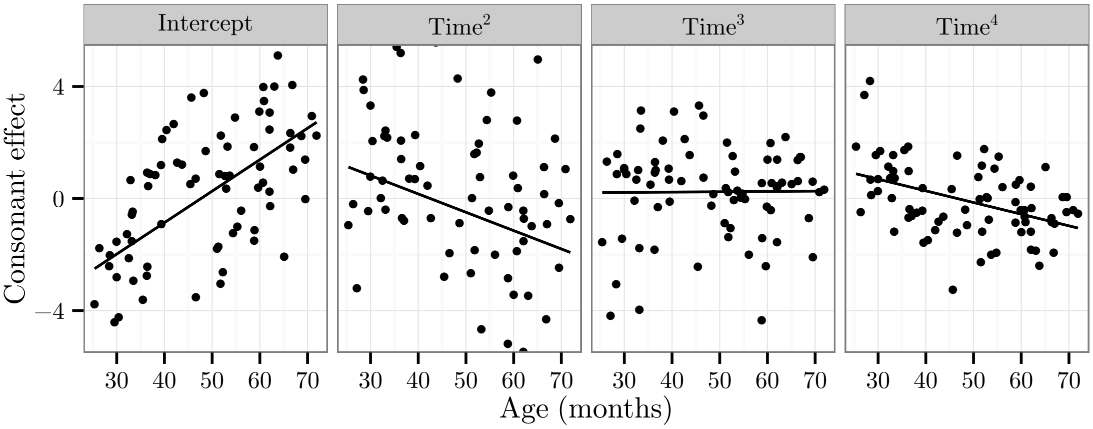
	Age group to which consonant-contrast model was fitted				
Fixed effect	**All**	**2 years**	**3 years**	**4 years**	**5 years**
Intercept	31.294 (0.386)	29.675 (0.960)	31.755 (0.647)	31.204 (0.608)	32.328 (0.640)
Cons[/ʃ/]	−3.257 (0.417)	−0.483 (0.714)	−3.664 (0.563)	−3.285 (0.634)	−5.148 (0.556)
Sex[M]	−0.485 (0.518)	0.361 (1.195)	−0.335 (0.897)	−0.058 (0.733)	−1.846 (0.949)
Cons[/ʃ/] × Sex[M]	0.852 (0.578)	1.107 (0.955)	0.917 (0.705)	0.339 (0.857)	0.735 (0.819)
Time	2.279 (0.688)	2.916 (2.014)	1.476 (1.374)	2.041 (1.234)	2.654 (1.126)
Time × Sex[M]	0.014 (0.969)	0.836 (2.652)	0.561 (1.997)	0.432 (1.695)	−1.837 (1.663)
Time^2	−10.147 (0.545)	−9.811 (1.320)	−9.713 (0.944)	−11.706 (1.347)	−9.179 (0.530)
Time^2 × Cons[/ʃ/]	2.752 (0.526)	1.424 (1.075)	1.824 (1.018)	4.288 (1.162)	2.962 (0.630)
Time^3	−0.480 (0.405)	−0.518 (1.056)	−0.119 (0.874)	−1.957 (0.623)	0.660 (0.780)
Time^3 × Cons[/ʃ/]	0.583 (0.321)	1.261 (0.830)	−0.396 (0.670)	1.472 (0.565)	0.217 (0.422)
Time^3 × Sex[M]	−0.499 (0.520)	−0.230 (1.280)	−0.100 (1.148)	0.028 (0.757)	−2.075 (1.103)
Time^4	−3.629 (0.345)	−2.020 (1.074)	−3.589 (0.593)	−4.182 (0.736)	−3.805 (0.462)
Time^4 × Cons[/ʃ/]	1.163 (0.414)	0.503 (1.252)	1.436 (0.772)	1.209 (0.928)	0.884 (0.570)
Time^4 × Cons[/s/] × Sex[M]	0.450 (0.468)	0.478 (1.412)	0.140 (0.799)	−0.192 (0.960)	0.524 (0.658)
Time^4 × Cons[/ʃ/] × Sex[M]	−0.132 (0.447)	−1.076 (1.372)	−0.289 (0.728)	0.003 (0.915)	0.573 (0.631)
Time^5	0.040 (0.296)	1.683 (0.828)	−0.348 (0.599)	−1.362 (0.559)	0.355 (0.511)
Time^5 × Sex[M]	−0.241 (0.412)	−0.396 (1.103)	0.306 (0.827)	0.428 (0.762)	−0.863 (0.757)
Time^5 × Cons[/ʃ/] × Sex[F]	−0.346 (0.361)	−1.998 (1.022)	−0.421 (0.756)	0.910 (0.699)	0.048 (0.534)
Time^5 × Cons[/ʃ/] × Sex[M]	0.169 (0.364)	−1.586 (0.947)	0.071 (0.751)	1.091 (0.689)	0.437 (0.588)

These effects of consonant suggest that as a group the children acoustically differentiate /s/ from /ʃ/ in similar ways as adults do, but they do not say anything about how, or even whether, this differentiation develops with age. This development was investigated through the random effects of the consonant-contrast model fitted to all the children's productions and through the fixed effects of separate consonant-contrast models fitted to each cross-sectional age group.

First, within the model fitted to all the children's productions, random effects to the intercept and to each power of polynomial time were estimated for each consonant within each subject. These random effects reflect the participants' deviations from the prediction made on the fixed effects, and, as such, can be used to determine the individuals difference in the effect of consonant on the intercept or on any power of time. For a given participant and factor (i.e., the intercept or a power of time), the *individual consonant effect* is defined to be the random effect for /ʃ/ subtracted from that of /s/.

Each child's individual consonant effect was computed for the intercept and for quadratic, cubic, and quartic time, since these were the factors that were subject to a significant effect of consonant in the adults' community-norm consonant-contrast model. These individual effects are plotted against age in Fig. 4.1. In the fitted community-norm model, a negative fixed effect of consonant reduced the magnitude of the intercept; thus, for individual consonant effects on the intercept, more extreme positive values indicate stronger effects. Conversely, positive fixed effects of consonant reduced the magnitudes of the three powers of time affected by consonant; thus, for individual consonant effects on a power of time, more extreme negative values indicate stronger effects. To determine whether the individual consonant effects got stronger with age, Kendall's rank correlation coefficient, τ, was computed and a one-sided test of significant was carried out. The individual effects on the intercept increased significantly with age ($\tau = 0.430$, $z = 5.490$, $p < 0.001$), while those on quadratic ($\tau = -0.186$, $z = -2.377$, $p < 0.01$) and quartic time ($\tau = -0.286$, $z = -3.651$, $p < 0.001$) significantly decreased with age. The individual effects on cubic time did not sig-

Figure 4.1: The English-acquiring children's individual consonant effects plotted against age, shown with a linear regression model fitted to the data.

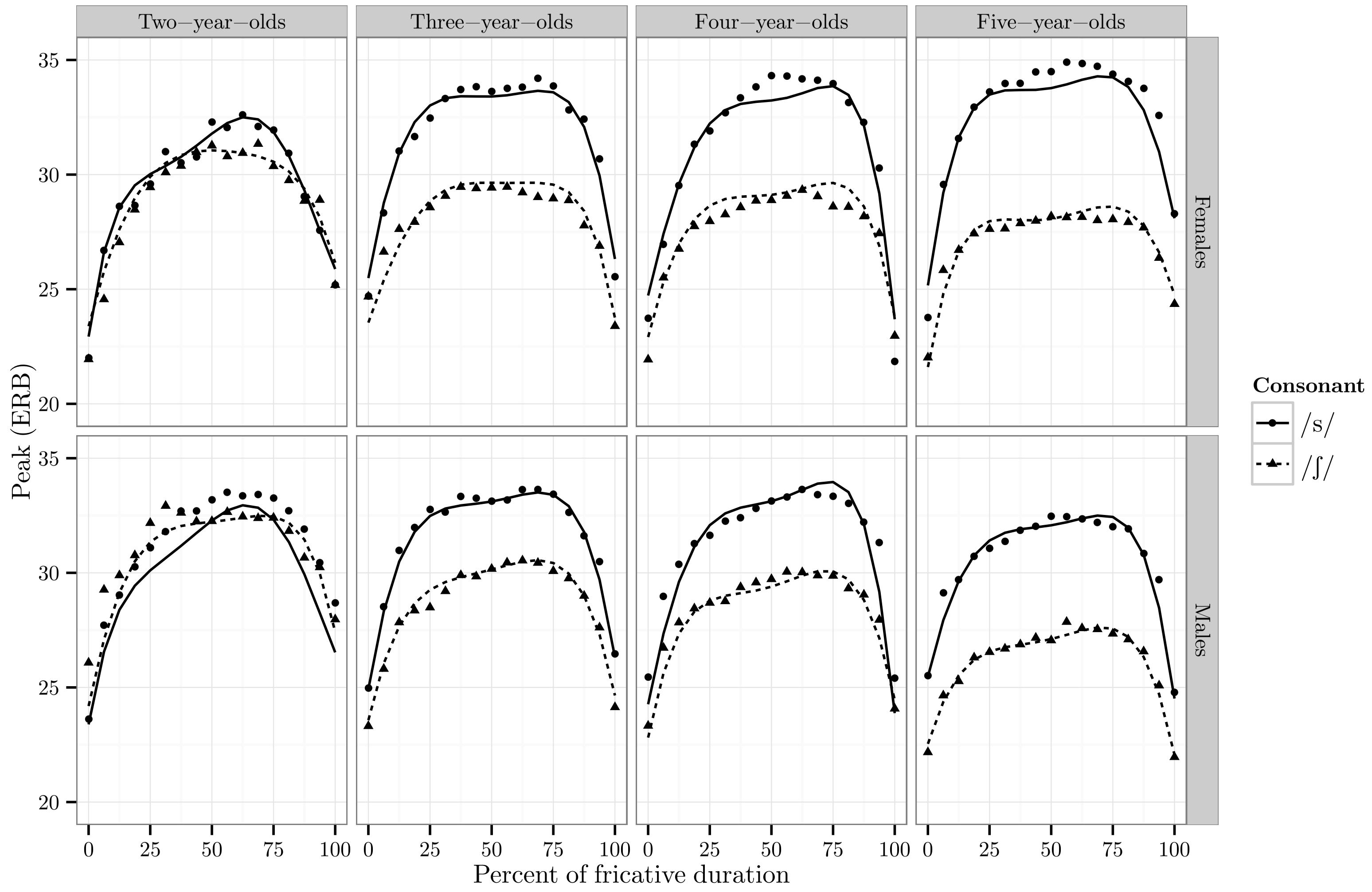


nificantly decrease with age ($\tau = -0.031$, $z = -0.395$, $p > 0.34$). The individual consonant effects, therefore, suggest that the effects on the intercept and on quadratic and quartic time become stronger with age.

While the individual consonant differences suggest that the children develop toward the community-norm, they give no clear indication of the age at which the effects of consonant become significant. To investigate this development, a separate consonant-contrast model was fitted to the data from each age group of English-acquiring children. In each model, effects of consonant were evaluated relative to /s/. The predicted trajectories of each model, made on its fixed effects, are shown in Fig. 4.2.

The fixed-effects coefficients of the consonant-contrast model fitted to the two-year-olds' productions are shown in the third column ('2 years') of Table 4.5. This fitted model included significant effects of quadratic ($\hat{\beta} = -9.8112$, $SE = 1.320$, $CI = [-12.398, -7.224]$) and quintic time ($\hat{\beta} = 1.683$, $SE = 0.828$, $CI = [0.059, 3.306]$), indicating that, for both consonants, the peak ERB$_N$-number trajectory followed a concave downward curve that was modified with minor inflections. No effects of consonant were significant, suggesting that, the two-year-olds did not distinguish /s/ from /ʃ/ in terms of peak ERB$_N$-number.

Figure 4.2: Fixed-effects predictions of the consonant-contrast models fitted to the productions of the English-acquiring children.



The fixed-effects coefficients of the fitted consonant-contrast model for the three-year-olds are shown in the fourth column (‘3 years’) of Table 4.5. The fitted model included significant effects of quadratic ($\hat{\beta} = -9.713$, $SE = 0.944$, $CI = [-11.562, -7.863]$) and quartic time ($\hat{\beta} = -3.589$, $SE = 0.593$, $CI = [-4.752, -2.426]$). As with the two-year-olds, the negative effect of quadratic time indicated that the peak ERB$_N$-number trajectory of each consonant followed a concave downward curve; however, the effect of quartic, rather than quintic, time indicates that the trajectory remained flatter through its middle portion of the three-year-olds’ productions. The effect of consonant ($\hat{\beta} = -3.664$, $SE = 0.563$, $CI = [-4.768, -2.560]$) on the intercept was significant and negative, suggesting that, at age three, the peak ERB$_N$-number trajectory of children’s productions of /ʃ/ is lower overall than that of /s/, but not different in shape.

The fifth row (‘4 years’) of Table 4.5 lists the fixed-effects coefficients of the consonant-contrast model when fitted to the four-year-olds’ productions. This model included significant, negative effects of quadratic ($\hat{\beta} = -11.706$, $SE = 1.347$, $CI = [-14.346, -9.066]$), cubic ($\hat{\beta} = -1.957$, $SE = 0.623$, $CI = [-3.178, -0.736]$), quartic ($\hat{\beta} = -4.182$, $SE = 0.736$, $CI = [-5.625, -2.738]$), and quintic time ($\hat{\beta} = -1.362$, $SE = 0.559$, $CI = [-2.459, -0.266]$). The fitted model also included a significant, negative effect of consonant on the intercept term ($\hat{\beta} = -3.285$, $SE = 0.634$, $CI = [-4.527, -2.043]$) and significant, positive effects on quadratic ($\hat{\beta} = 4.288$, $SE = 1.162$, $CI = [2.011, 6.564]$) and cubic time ($\hat{\beta} = 1.472$, $SE = 0.565$, $CI = [0.365, 2.578]$). The effects of consonant on non-zero powers of polynomial time suggest that the peak ERB$_N$-number trajectory of /ʃ/ differs from that of /s/, not just in terms of average value, but also in terms of shape, the positive effects indicating that the trajectory of /ʃ/ is less concave.

The fixed-effects coefficients of the fitted consonant-contrast model for the five-year-olds are listed in the sixth row (‘5 years’) of Table 4.5. This fitted model included a significant, positive effect of linear time ($\hat{\beta} = 2.654$, $SE = 1.126$, $CI = [0.447, 4.860]$) and significant, negative effects of quadratic ($\hat{\beta} = -9.179$, $SE = 0.530$, $CI = [-10.219, -8.140]$) and quartic time ($\hat{\beta} = -3.805$, $SE = 0.462$, $CI = [-4.711, -2.899]$), indicating that the peak ERB$_N$-number trajectory of either consonant followed an increasing, concave downward curve. There were also significant effects of consonant on the intercept term ($\hat{\beta} = -5.148$, $SE = 0.556$, $CI = [-6.238, -4.059]$) and on quadratic time ($\hat{\beta} = 2.962$, $SE = 0.630$, $CI = [1.728, 4.196]$), suggesting that the peak ERB$_N$-number trajectory of /ʃ/ was lower overall and less concave than that of /s/.

4.3.2 The development of sibilant contrast in Japanese-acquiring children

Using the fixed effects listed in Table 3.5, a consonant-contrast model was fitted to all the children's productions, pooled across age group. The fixed-effects coefficients of this fitted model are shown in the second column ('All') in Table 4.6. In the fitted model, there were significant, negative effects of quadratic ($\hat{\beta} = -9.802$, $SE = 0.836$, $CI = [-11.440, -8.163]$), quartic ($\hat{\beta} = -3.359$, $SE = 0.398$, $CI = [-4.139, -2.580]$), and quintic time ($\hat{\beta} = -0.742$, $SE = 0.241$, $CI = [-1.214, -0.269]$), indicating that peak ERB$_N$-number followed an inflected, concave downward curve across the time course of either /s/ or /ɕ/. The fitted model also included a significant, negative effect of consonant on the intercept ($\hat{\beta} = -0.670$, $SE = 0.231$, $CI = [-1.123, -0.217]$), indicating that average peak ERB$_N$-number across the duration of the fricative was lower for /ɕ/ than for /s/.

None of the effects of consonant on the powers of polynomial time were significant; however, these effects were positive (linear: $\hat{\beta} = 0.217$; quadratic: $\hat{\beta} = 1.405$; cubic: $\hat{\beta} = 0.172$; quartic: $\hat{\beta} = 0.505$; quintic: $\hat{\beta} = 0.457$), as was the case for the adults (linear: $\hat{\beta} = 1.950$; quadratic: $\hat{\beta} = 3.369$; cubic: $\hat{\beta} = 1.991$; quartic: $\hat{\beta} = 1.137$; quintic: $\hat{\beta} = 0.871$). That the effect of consonant was, in each case, positive, but smaller in magnitude for the children than for the adults, leaves open the possibility that the children are on a developmental path toward the adults' community-norm model of sibilant contrast, but that they are just not far enough along for effects of consonant to have become regular enough to be detected statistically.

Table 4.6: Fixed-effects coefficients (and their standard errors, in parentheses) for the consonant-contrast models fitted to the Japanese-acquiring children's sibilant productions. Each column denotes a single fitted model. Shaded cells denote coefficients whose estimate was not significantly different from zero in the fitted model of that column, as determined by a 95% Wald confidence interval.

	Age group to which consonant-contrast model was fitted				
Fixed effect	**All**	**2 years**	**3 years**	**4 years**	**5 years**
Intercept	29.149 (0.427)	26.856 (1.227)	29.743 (0.717)	29.247 (0.668)	29.768 (0.772)
Cons[/ɕ/]	−0.670 (0.231)	0.144 (0.551)	−0.205 (0.416)	−1.515 (0.375)	−0.764 (0.395)
Sex[M]	0.222 (0.557)	1.337 (1.509)	−1.400 (0.964)	2.086 0.844	−0.639 (1.013)
Time	−0.568 (0.627)	−0.405 (1.550)	−4.312 (1.165)	−0.332 (0.888)	2.555 (1.314)
Time × Cons[/ɕ/]	0.217 (0.685)	0.520 (1.288)	3.641 (1.482)	−0.131 (0.834)	−2.718 (1.278)
Time^2	−9.802 (0.836)	−5.079 (1.643)	−8.467 (1.363)	−13.293 (1.676)	−10.606 (1.614)
Time^2 × Cons[/ɕ/]	1.405 (0.881)	0.353 (1.622)	1.079 (1.564)	2.538 (1.705)	0.526 (1.212)
Time^2 × Sex[M]	−1.610 (1.119)	−3.717 (2.149)	−4.294 (1.656)	2.280 (2.214)	−1.776 (2.136)
Time^2 × Cons[/ɕ/] × Sex[M]	0.536 (1.163)	−0.982 (2.065)	1.744 (1.581)	−0.803 (2.205)	3.242 (1.444)
Time^3	−0.630 (0.469)	−0.731 (0.999)	−0.593 (0.869)	0.535 (1.028)	−0.671 (0.912)
Time^3 × Cons[/ɕ/]	0.172 (0.378)	−0.373 (0.913)	0.018 (0.665)	0.080 (0.805)	0.707 (0.677)
Time^3 × Sex[M]	0.640 (0.548)	1.647 (1.141)	0.109 (1.103)	−0.861 (0.980)	−0.044 (1.114)
Time^4	−3.359 (0.398)	−2.695 (1.036)	−2.758 (0.840)	−4.245 (0.707)	−2.978 (0.705)
Time^4 × Cons[/ɕ/]	0.505 (0.354)	−0.977 (0.872)	1.873 (0.759)	0.993 (0.575)	−0.431 (0.658)
Time^4 × Sex[M]	0.963 (0.482)	1.914 (1.235)	−0.480 (1.039)	1.399 (0.836)	0.319 (0.853)
Time^5	−0.742 (0.241)	−0.784 (0.599)	−0.766 (0.472)	−0.790 (0.525)	−0.643 (0.448)
Time^5 × Cons[/ɕ/]	0.457 0.308	0.262 (0.789)	0.288 (0.635)	0.496 (0.558)	0.686 (0.582)

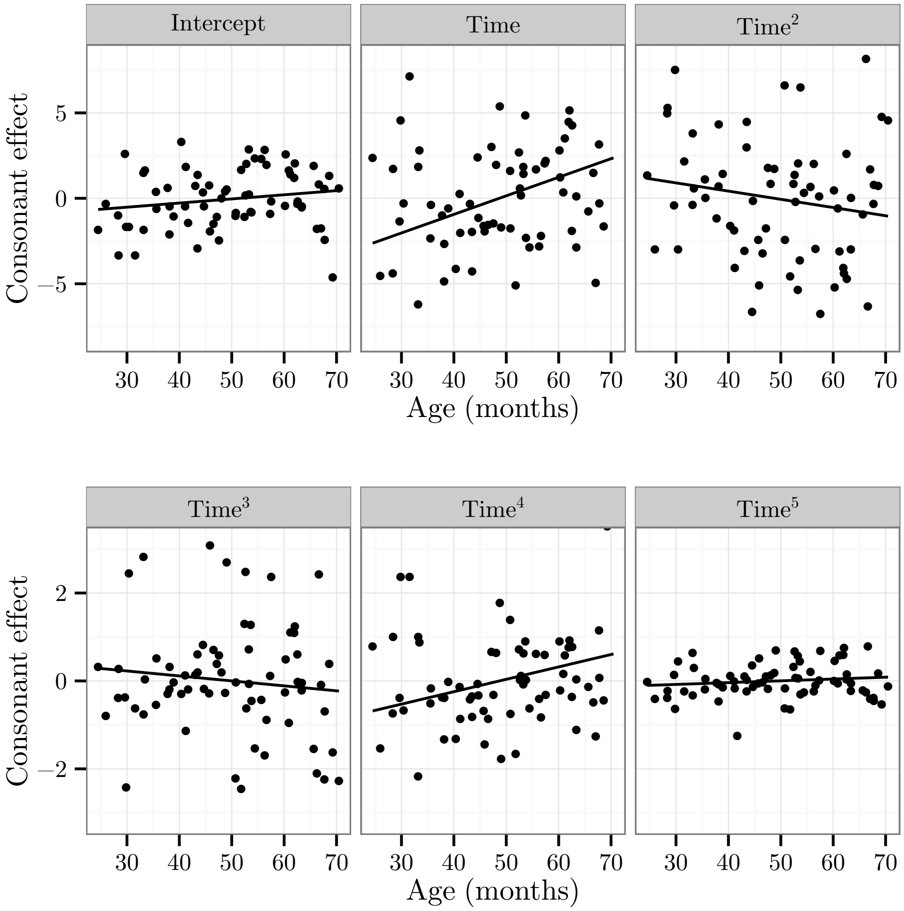
To examine whether the effects of consonant strengthen developmentally in the children, the random effects of the fitted model were used to determine the individual consonant effects for the intercept and for each power of time, by subtracting the random effects for /ɕ/ from the ones for /s/. These individual effects are shown in Fig. 4.3. For the individual consonant effects on the intercept, a developmental strengthening would be indicated by a positive association with age; whereas for the effects on the powers of polynomial time, this would be signaled by a negative association. Kendall's rank correlation τ-tests revealed that the individual consonant effects on the intercept ($\tau = 0.137$, $z = 1.668$, $p < 0.05$), linear time ($\tau = 0.190$, $z = 2.310$, $p < 0.01$), and quartic time ($\tau = 0.165$, $z = 1.999$, $p < 0.05$) were significantly, positively associated with age. The association with age was also positive, but not significant, for quintic time ($\tau = 0.069$, $z = 0.839$, $p > 0.20$). Quadratic ($\tau = -0.098$, $z = -1.191$, $p > 0.011$) and cubic time ($\tau = -0.042$, $z = -0.508$, $p > 0.30$) were both negatively, though not significantly, associated with age.

To determine the age at which the effects of consonant on the intercept and the powers of time became significant, separate consonant-contrast models were fitted to the data from each age group. The fixed-effects coefficients of the models fitted to the individual age groups are listed in the four rightmost columns of Table 4.6. The predicted trajectories, made on the fixed effects of each fitted model, are shown in Fig. 4.4.

When fitted to the two-year-olds' data, the consonant-contrast model included significant effects of quadratic ($\hat{\beta} = -5.079$, $SE = 1.643$, $CI = [-8.299, -1.859]$) and quartic time ($\hat{\beta} = -2.695$, $SE = 1.036$, $CI = [-4.726, -0.664]$), indicating that peak ERB$_N$-number traced a concave trajectory across the duration of both consonants. There were no significant effects of consonant, suggesting that /s/ and /ɕ/ are not differentiated acoustically by the two-year-olds.

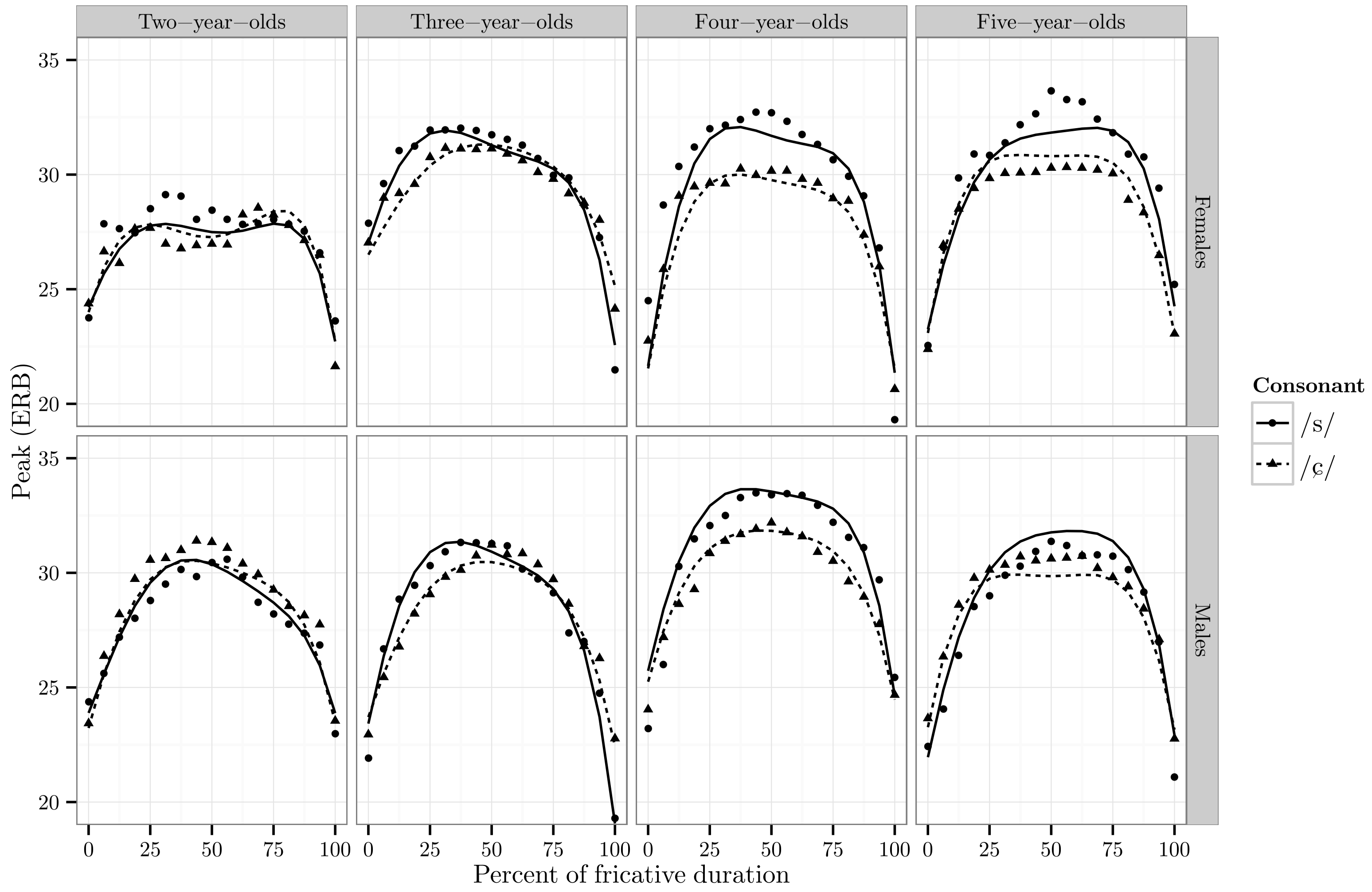
The consonant-contrast model fitted to the three-year-olds' productions included significant, negative effects of linear ($\hat{\beta} = -4.312$, $SE = 1.165$, $CI = [-6.595, -2.029]$), quadratic ($\hat{\beta} = -8.467$, $SE = 1.363$, $CI = [-11.137, -5.796]$), and quartic time ($\hat{\beta} = -2.758$, $SE =$

Figure 4.3: The Japanese-acquiring children's individual consonant effects, plotted against age.



0.840, $CI = [-4.405, -1.111]$), indicating that peak ERB$_N$-number followed a concave curve, and tended to decrease across the full duration of either sibilant. There were significant, positive effects of consonant on linear ($\hat{\beta} = 3.641$, $SE = 1.482$, $CI = [0.7356, 6.546]$) and quartic time ($\hat{\beta} = 1.873$, $SE = 0.759$, $CI = [0.385, 3.361]$), suggesting that the concavity and global, downward slope of the trajectory was reduced for /ɕ/, compared to /s/.

Figure 4.4: Fixed-effects predictions of the consonant-contrast models fitted to the productions of the Japanese-acquiring children.



For the four-year-olds, the fitted consonant-contrast model included significant effects of quadratic ($\hat{\beta} = -13.293$, $SE = 1.676$, $CI = [-16.578, -10.007]$) and quartic time ($\hat{\beta} = -4.245$, $SE = 0.707$, $CI = [-5.630, -2.860]$), again indicating a concave peak ERB$_N$-number trajectory for both sibilants. There was a significant, negative effect of consonant on the intercept ($\hat{\beta} = -1.515$, $SE = 0.375$, $CI = [-2.250, -0.781]$), indicating that peak ERB$_N$-number was lower on average across the time course of /ɕ/, compared to /s/. No effects of consonant on powers of time were significant.

When fitted to the five-year-olds' data, the consonant-contrast model included significant, negative effects of quadratic ($\hat{\beta} = -10.606$, $SE = 1.614$, $CI = [-13.770, -7.441]$) and quartic time ($\hat{\beta} = -2.978$, $SE = 0.705$, $CI = [-4.361, -1.595]$), indicating a concave curvature to the peak ERB$_N$-number trajectory of either sibilant. There was a significant effect of consonant on linear time ($\hat{\beta} = -2.718$, $SE = 1.278$, $CI = [-5.223, -0.213]$); however, this effect of consonant was negative, in the opposite direction as would be expected from the adults' model.

4.4 Discussion

4.4.1 English-acquiring children

The results for the English-acquiring children generally supported the hypotheses on how the differentiation of sibilant fricatives would develop in terms of their spectral dynamic properties. The first hypothesis proposed that the two-year-old children's peak ERB$_N$-number trajectories would be undifferentiated in terms of their shape properties, in particular their concavity, and that they would more closely resemble the adults' /s/ than their /ʃ/. That the peak ERB$_N$-number trajectories for /s/ and /ʃ/ are not differentiated by the two-year-olds is suggested by the absence of any significant effects of consonant in the consonant-contrast model fitted to their data. From this fitted model, the estimates for the coefficients on quadratic time for /s/ and /ʃ/ are $\hat{\beta} = -9.811$ and $\hat{\beta} = -8.388$, respectively. From the adults' consonant-contrast model, the analogous estimates are $\hat{\beta} = -4.345$ and $\hat{\beta} = -2.028$. Thus, in terms of concavity, the two-year-olds' peak ERB$_N$-number trajectories are more similar to those of the adults' /s/ productions than their /ʃ/ productions.

Furthermore, comparing the estimates of the quadratic coefficient across age groups suggests that in the children, the concavity of /s/ mostly stays the same (two-year-olds: $\hat{\beta} = -9.811$; three-year-olds: $\hat{\beta} = -9.712$; four-year-olds: $\hat{\beta} = -11.706$; five-year-olds: $\hat{\beta} = -9.179$), but the concavity of /ʃ/ decreases (two-year-olds: $\hat{\beta} = -8.3876$; three-year-olds: $\hat{\beta} = -7.889$; four-year-olds: $\hat{\beta} = -7.418$; five-year-olds: $\hat{\beta} = -6.217$). This pattern is reminiscent of what has been observed for the development of sibilant contrast, in terms of centroid frequency, wherein the mean centroid values for /s/ and /ʃ/ diverge primarily due to the centroid of /ʃ/ decreasing (cf. Fox and Nissen, 2005; Li, 2012; Nissen and Fox, 2005; Nittrouer, 1995).

The second hypothesis supposed that the effects of consonant on polynomial time would increase with age. The correlations between age and the individual consonant effects on quadratic, cubic, and quartic time partially support this hypothesis. In particular, the individual effects on quadratic and quartic time were significantly, negatively correlated with age, which indicated a developmental strengthening. The individual effects on cubic time were not significantly correlated with age. Effects on cubic time indicate an asymmetry in the tails of the trajectory. In the adults' productions, the slope of the left tail was much shallower than that of the right tail, indicating a gradual increase in peak ERB$_N$-number at the beginning of the sibilant, and a more sudden drop in peak ERB$_N$-number at the end. The children's productions, on the other hand, show rapid increases in peak ERB$_N$-number across the first quarter of either sibilant. Given the resemblance of the left and right tails of the children's peak ERB$_N$-number trajectories, and that the right tails indicate the release of the constriction, it is likely that the steep increase in peak ERB$_N$-number at the beginning of the sibilant is, at least partly, due to it being initiated while the linguapalatal constriction is still open enough to permit back-cavity coupling. Since the target words in the repetition task were not embedded within a carrier phrase, the participants' productions would not be subject to any perseverative coarticulatory effects; thus, it is possible that the adults initiated each sibilant-initial word with an already tightly-formed constriction. The articulatory results from Iskarous et al. (2011), who found the degree of constriction to be comparable at the onset and midpoint of frication for sibilants in word-initial position, support this possibility.

The third hypothesis proposed that the effect of consonant on the intercept would reach statistical significance at an earlier age than its effect on the powers of polynomial time. The results of the consonant-contrast models separately fitted to each cross-sectional age group support this hypothesis. The effect on intercept was significant for the three-, four-, and five-year-olds, but effects on polynomial time were significant only for the four- and five-year-olds. Comparing the fitted models across age groups, a picture of the development of sibilant contrast emerges: In the youngest children, /s/ and /s/ are indistinguishable in terms of peak ERB$_N$-number. Later, these two sounds are differentiated, first in terms of the average peak ERB$_N$-number across the time course of the fricative, then in terms of dynamic aspects of the peak ERB$_N$-number trajectory, like its concavity.

4.4.2 Japanese-acquiring children

The first hypothesis proposed that the two-year-olds' productions would be undifferentiated in terms of their shape properties, and that they would be more similar to the adults' productions of /ɕ/ in terms of their concavity and the asymmetry between their tails. This hypothesis was partially supported. The consonant-contrast model fitted to the two-year-olds' productions suggested that, at this age, /s/ and /ɕ/ are undifferentiated acoustically. From the coefficient estimates for this fitted model, the mean quadratic estimates for /s/ and /ɕ/ were $\hat{\beta} = -6.937$ and $\hat{\beta} = -7.075$, respectively; whereas, from the adults' consonant-contrast model, these estimates were $\hat{\beta} = -11.0175$ and $\hat{\beta} = -6.512$, respectively. Thus, in terms of concavity, the two-year-olds' peak ERB$_N$-number trajectories are more similar to the adults' /ɕ/. The adults' cubic and quintic coefficients were negative for both /s/ and /ɕ/, though smaller in magnitude for the latter. Since the effects of cubic and quintic time, which affect the asymmetry of the trajectory's tails, were never significant in the children's models, they were, for the two-year-olds, more similar to the adults' /ɕ/. However, this similarity is specious and should not be interpreted as support for the hypothesis.

The second hypothesis proposed that the effects of consonant on polynomial time would strengthen with age. A developmental strengthening of these effects would have been evinced by a negative association between age and the individual consonant effects on any power of time; however, only the individual effects on quadratic and cubic time exhibited a negative association with age, and neither was significant. Thus, this hypothesis was not supported.

The third hypothesis suggested that when tested cross-sectionally, the effects of consonant on polynomial time, which indicate a change of the affiliation of peak ERB$_N$-number from the front cavity to the back cavity, would reach statistical significance at an earlier age than the effects on intercept. This hypothesis was partially supported. Effects of consonant were first seen in the three-year-olds, and these effects were on linear and quadratic time. The earliest (and only) age at which an effect of consonant on the intercept was significant was four years old. Thus, significant effects on polynomial time were found at an earlier age than those on the intercept; however, the effects on time did not remain significant in the older age groups.

For the five-year-olds, the only effect of consonant was in the opposite direction as would be expected; i.e. a negative effect on linear time. To make sure that there was not a preponderance of phonemically incorrect productions (e.g., [ɕ]-for-/s/ substitutions or vice versa) in the five-year-olds data, the phonemic accuracy was computed for each age group. The five-year-olds were most accurate, with 88.89% of their productions judged to be phonemically correct; hence, the odd results for the five-year-olds does not seem to be due to their having produced a great number errors.

The results for the Japanese-acquiring five-year-olds suggest a developmental regression in terms of the amount of acoustic contrast between /s/ and /ɕ/. While the overall trend of the contrast between sibilants is surely an increasing one, when considered from infancy to adulthood, it is possible that local regressions in development do occur in children. For example, in the data on English consonant acquisition reported by Smit et al. (1990, Fig. 10, p. 790), both boys and girls can be seen to regress in the accuracy of /s/ from 3;6 to 4;0. It is possible that a similar regression happens later in Japanese and because it occurred in the oldest age cohort, the increasing trend of sibilant contrast was compromised. This possibility would be better investigated with longitudinal data, rather than cross-sectional data as was presented here.

Chapter 5

Vowel-context effects on spectral dynamics

5.1 Introduction

The preceding chapters have argued that English-speaking adults contrast /s/ and /ʃ/ in terms of dynamic spectral properties, and that the development of English-acquiring children's differentiation of the gestures for /s/ and /ʃ/ is observable in the peak ERB_N-number trajectories of their productions of these consonants. In addition to the differentiation of gestures, the acquisition of adult-like speech production skills has been argued to involve development in the coordination of the gestures for the individual segments that compose a syllable or word (see Goffman et al., 2008; Smith and Goffman, 1998).

A common method for investigating the coordination of gestures has been through coarticulatory effects on a segment imposed by its phonetic context. Such coarticulatory effects arise from the need for the articulatory gestures of adjacent sounds to be blended together in order for speech to be fluent, not a variegated chain of sounds separated by perceivable breaks or silences. As such, any sound or gesture may be subject to the influence of either its preceding or following sound. Effects of the former have generally been interpreted as consequences of the mechanical or inertial properties of the articulators, while those of the latter, referred to as *anticipatory effects*, as indices of language-specific patterns of coordination and inter gestural timing (cf. Katz and Bharadwaj, 2001). Anticipatory effects of a following vowel on the English sibilants /s/ and /ʃ/ have been investigated through both articulatory and acoustic methodologies. In this dissertation, the phrase *coarticulatory effect* is reserved for an anticipatory effect that is observed in articulatory data, while *vowel-context effect* denotes an anticipatory effect in the acoustics.

5.1.1 Coarticulatory effects on sibilants

Lingual coarticulatory effects

Zharkova et al. (2011, 2012) investigated spatial coarticulatory effects of a following vowel on /s/ and /ʃ/, in both adults and children, between the ages of six and ten. An ultrasound transducer was used to image the midsagittal contour of the tongue during production of the syllables /si/, /sɑ/, /su/, /ʃi/, /ʃɑ/, and /ʃu/. From each production, only the frame that was recorded at the midpoint of frication was analyzed. For each speaker and each syllable type, the distance between each pair of tongue contours was computed. These distances, referred to as *within-set (WS) distances*, represented the amount of token-to-token variation that a speaker exhibits across different productions of the same target syllable. Similarly, for each speaker and each pair of syllable types with the same sibilant, the distance between any two tongue contours from different syllable types was computed. These distances, referred to as *across-set (AS) distances*, represented the amount of variation in midsagittal tongue shape that was conditional on the following vocalic context.

For a given pair of syllable types, a coarticulatory effect was established if the AS distances were found to be significantly greater than the WS distances of both syllable types, individually. Using this method, both children and adults were found to exhibit coarticulatory effects between /ʃɑ/–/ʃi/ and /ʃɑ/–/ʃu/. The extent of coarticulation across the age groups was examined for these two pairs separately by comparing the AS distances. Children were found to have significantly greater AS distances than the adults for either pair of syllables, suggesting that during the articulation of /ʃ/, the shape of the tongue is adapted to the following vowel to a greater extent by children than by adults.

For the /s/-initial syllables, the adults exhibited coarticulatory effects in each pair of syllable types, but the children exhibited no coarticulatory effects. It was suggested that the absence of coarticulatory effects on /s/ in the children's productions was due to their inability to differentially control the dorsum of the tongue from the tip. For example, in the adults' productions of /si/ and /sa/, the coarticulatory effect was apparent in the posterior part of the tongue, which was more advanced in the context of /i/ than of /ɑ/, but no such difference was observed in the children's productions (Zharkova et al., 2012, Figs. 2 and 3, pp. 198–199). If the children were unable to differentially control the tongue tip and the tongue dorsum, then the tongue dorsum would be less likely to adapt to the following vowel because it would be actively recruited for the posture necessary to maintain the linguapalatal constriction during production of the fricative.

Expanding on their previous work, Zharkova et al. (2014) investigated the spatial and temporal coarticulatory effects of a following vowel in adults' and preadolescents' (ages 10 to 12) productions of /si/, /sɑ/, /ʃi/, and /ʃɑ/. From each production, nine tongue contours spaced equally in time across the duration of the sibilant were analyzed. Spatial coarticulatory effects were investigated for each time point independently of the others. The adults exhibited spatial coarticulatory effects at each time point for both sibilants, while the children exhibited such effects from the second and fourth time points for /s/ and /ʃ/, respectively. Unlike in their previous studies where younger children were studied, Zharkova et al. (2014) found no group-related differences between the preadolescents and the adults in the extent of spatial coarticulatory effects.

Temporal coarticulatory effects were investigated by first determining for each participant and syllable type, the earliest time point at which AS distances became significantly greater than WS distances. For the adults, the mean time point of the onset of coarticulatory effects was 1.8 and 1.7 for /s/ and /ʃ/, respectively. The preadolescents, on average, showed coarticulatory effects on /s/ by time point 2.0, and on /ʃ/ by 3.5. The difference in the temporal coarticulatory effects on /ʃ/ was significant across the two age groups, suggesting that the onset of coarticulation occurs relatively later in the production for preadolescent children than for adults. Given that children typically have a slower speech rate than adults, it is possible that this temporal coarticulatory effect arose simply from the time-normalization. However, the authors found that the adults actually produced slightly longer sibilants than the children: the adults' mean segmental durations were 177 ms and 181 ms for /s/and /ʃ/, respectively, while the children's were 169 ms and 177 ms for /s/ and /ʃ/, respectively.

Temporal coarticulatory effects on the tongue have also been investigated by Katz and Bharadwaj (2001) who recorded productions of the syllables /si/, /su/, /ʃi/, and /ʃu/ spoken by small cohorts of five-year-olds, seven-year-olds, and adults. The movements of each participant's tongue tip and tongue body were recorded using electromagnetic midsagittal articulography, in which small metallic pellets were fixed onto the midline of the tongue at the two measurement points. Coarticulatory effects were investigated on the horizontal position of the tongue tip and tongue body since these structures were expected to move anteriorly for the articulation of /i/, but posteriorly for /u/. Furthermore, this coarticulatory effect was investigated in terms of absolute time, rather than relative time. The position of the tongue was compared across consonants and groups at the fricative-vowel boundary, and 30 ms and 100 ms prior to this point. For /s/, it was found that the coarticulatory effect was earlier in the children's productions than in the adults'; however, for /ʃ/, no difference in the onset of a coarticulatory effect was observed between age groups.

Mandibular coarticulatory effects

In their study of the articulatory kinematics of /s/, Iskarous et al. (2011) found that the time course of the jaw height varied according to the height of the following vowel. In both high- and low-vowel contexts, the jaw rose comparable amounts across the first half of the fricative. Over the second half, though, the jaw fell significantly more when /s/ preceded a low vowel, with the difference in jaw position across the two contexts becoming significant around 75% of the fricative's duration.

Labial coarticulatory effects

Toda et al. (2002) investigated anticipatory effects of a following vowel on the protrusion of and area enclosed by the lips during the production of alveolar and postalveolar sibilants,[1] measured at the temporal midpoint of the consonant. For the alveolars, the lips were more protruded in the context of /u/ than in the context of either /i/ or /ɑ/, and the lips enclosed a smaller area in the context of /u/ than of /ɑ/. During the production of the postalveolars, the lips were more protruded in the context of /u/ than of /ɑ/, but no other coarticulatory effects were found. Thus, the alveolar sibilants were found to be more sensitive than the postalveolars to anticipatory lip rounding. This asymmetry was suggested to be due to lip rounding being saturated in postalveolar sibilants, as they are articulated with lip rounding even in unrounded vowel contexts.

5.1.2 Acoustic vowel-context effects

Soli (1981) analyzed the spectral consequences of sibilant-vowel coarticulation across the final 60 ms of frication. He found that the spectral prominence that is associated with the following vowel's $F2$ frequency grows in amplitude with closer proximity to the fricative-vowel boundary. The amplitude of this prominence was found to be greater in the context of a high vowel, /i/ or /u/, and greater in the spectra of /ʃ/ than of /s/. Indeed, in the syllable /ʃɑ/, the $F2$ prominence was the highest spectral peak at the end of frication.

Using similar methodology as Soli (1981), McGowan and Nittrouer (1988) found that the spectra of children's productions of /s/ and /ʃ/ also show low frequency peaks affiliated with $F2$ of the following vowel. As was the case with the adults, these peaks tended to increase in amplitude the closer to the fricative-vowel boundary. Two differences between the children's and the adults' $F2$ peaks emerged. First, the amplitude of the $F2$ peak, relative to the high frequency (i.e. front cavity) peak was greater in children's productions. A possible

[1] In their study the voiceless sibilant fricatives /s/ and /ʃ/ were not analyzed separately, but grouped with other alveolars /s, z, ts, dz/ and the postalveolars /ʃ, ʒ, tʃ, dʒ/.

explanation for this is that the size of the glottal opening relative to the constriction is smaller for children than for adults, from which it would follow that the level of the frication at the glottis would be relatively greater in children than adults. Second, the $F2$ peaks occurred at higher frequencies for the children than the adults. This observation follows from children's vocal tracts being smaller than adults' causing the resonant frequencies of the former to be relatively higher.

Nittrouer et al. (1989) examined the contextual effects of the vowels /i/ and /u/ on the sibilants /s/ and /ʃ/ produced by adults and three-, four-, five-, and seven-year-old children. Since these two vowels differ in terms of both lip rounding and tongue placement, the authors estimated centroid frequency at 100 ms prior to vowel onset, in order to index effects of labial coarticulation, and $F2$ frequency at 30 ms prior to vowel onset, in order to index effects of lingual coarticulation. An effect of labial rounding was found on /s/, but not /ʃ/: for /s/, centroid was lower in the context of /u/ (cf. Iskarous et al., 2011; Toda et al., 2002). Their analysis of $F2$ reached similar conclusions as McGowan and Nittrouer (1988). Additionally, Nittrouer et al. (1989) investigated the extent of vowel-context effects across age groups. Relative to adults, the children were found to exhibit stronger vowel-context effects on $F2$ frequency, but comparable effects on centroid frequency (Nittrouer et al., 1996, cf.).

5.1.3 Contextual effects as a window into phonological organization

Building on a previous perceptual study (Nittrouer and Studdert-Kennedy, 1987), which found children to be more sensitive to transitional information than the frication when identifying fricative-vowel syllables, Nittrouer et al. (1989) interpreted the stronger effect of vowel context on $F2$ frequency in children as evidence that they organize their speech production more syllabically than segmentally. From this it would follow that coarticulatory effects should decrease as children matured. Furthermore, the absence of an effect on centroid across /i/ and /u/ contexts could simply mean that labial coarticulation is already adult-like by age three; crucially, the extent of the vowel-context effect on centroid was not greater in adults.

Zharkova et al. (2014), however, did find that both spatial and temporal coarticulatory effects were stronger in adults than in preadolescents. In light of their finding, they suggested an alternative interpretation of coarticulatory effects in children and adults: "that age-related differences ... may depend not on whether the units are syllables or segments but on the nature of the task to be performed by the articulators in each case" (p. 384). Thus, even if children have a greater propensity than adults to program their speech in terms of syllables rather than segments, some contextual effects would be expected to be stronger in adults since these effects are also sensitive to the maturation of articulator control which is still ongoing in children.

5.1.4 Purpose and hypotheses

Due to differences in methodology, the contrasting results of Nittrouer et al. (1989) and Zharkova et al. (2014) are not easily reconciled. For one, the children studied by Nittrouer and her colleagues were younger (between three and seven years old) than those reported by Zharkova and her colleagues, who were between 10 and 12 years of age. Bearing in mind the results of Romeo et al. (2013), who found that the cross-category distance between /s/ and /ʃ/ exhibited a developmental overshoot between the ages of nine and twelve before decreasing toward adult-like levels in 13 and 14-year-olds, it is possible that Zharkova and her colleagues observed preadolescents during a similar stage of developmental overshoot, arrived at through too great of reductions in the coproduction of the fricative and vowel gestures.

Alternatively, it is possible that the scope of Nittrouer's and her colleagues' analysis was too narrow to appreciate the complexity of vowel-context effects on the spectral properties of sibilant fricatives. The preceding chapters have argued that the spectral properties of English /s/ and /ʃ/ differ in terms of both static and dynamic aspects, which raises the possibility that the phonetic context may also affect the dynamic aspects of a spectral property's trajectory through the course of a sibilant. However, the analysis in Nittrouer et al. (1989) only allowed for contextual effects on either centroid or $F2$ to be evaluated in static terms.

The analyses in this chapter evaluate the presence and extent of vowel-context effects on static and dynamic aspects of peak ERB_N-number trajectories in sibilants produced by adults and by two- through five-year-old children. Analyzing children in this age range makes comparison with Nittrouer et al. (1989) more straightforward, avoiding the potential for differences due to the participants' developmental stages of motor control. It is hypothesized that vowel context will affect both global and time-varying aspects of the peak ERB_N-number trajectories of either fricative. Additionally, the development of the strength of these effects in children will be tested. If, following Nittrouer, contextual effects indicate the size of the units used to plan speech gestures, and if children plan their productions using larger units (i.e. syllables rather than segments), then the vowel-context effects would be expected to weaken developmentally. On the other hand, if contextual effects indicate more the complexity of gestural coordination, as Zharkova and her colleagues suggest, then vowel-context effects would not necessarily be expected to weaken with age; some may, in fact, strengthen, reflecting greater coordination between the sibilant and vowel gestures.

5.2 Method

The productions analyzed here are the same productions from the same adults and children, who were reported in chapters 3 and 4. The language materials, stimuli, and procedure used to elicit these productions are described fully in §§3.2.1–3.2.3 and §§4.2.1–4.2.3. The procedure for annotating the onset and offset of frication and the criteria for inclusion in the analysis are described in §3.2.4 and §4.2.4. The method with which peak ERB_N-number trajectories were computed is described in §3.2.5.

5.2.1 Vowel-context models

Because English /s/ and /ʃ/ have been found to exhibit different anticipatory coarticulatory and vowel-context effects, a separate vowel-context model was fitted for each sibilant. For the adults' productions, the fixed-effects structure of the vowel-context models was built up using a stepwise forward selection protocol, as with the consonant-contrast models. The base model in this procedure was determined from the adults' fitted consonant-contrast model by removing from it all fixed effects that involved the consonant factor. This base model represents the time and sex-demographic factors that affect the peak ERB_N-number trajectories of a language's sibilants, generally. To this base model, fixed effects of vowel were added if a likelihood ratio test found that the model fit was significantly improved. First, a main effect of vowel was considered. Next, a binary interaction between vowel and polynomial time and a ternary interaction between vowel, sex, and time were considered at increasing powers of time. Since each sibilant-vowel sequence was produced multiple times by each participant, each vowel-context model included random effects of intercept and of each power of time, for participant and for vowel within participant. For the adults, the vowel factor had five levels, corresponding to the five vowel categories: /i/, /e/, /ɑ/, /o/, /u/.

Because the children were less accurate than the adults at producing the target consonant as a sibilant, there were too few children who produced a sibilant target for each vowel context to analyze context effects in the same way as was done for the adults. Instead of treating each vowel context separately, the vowels were releveled according to their rounding, frontness, and height features. For rounding, the vowels were releveled as *rounded* (/o, u/) and *unrounded* (/i, e, ɑ/), with effects determined relative to unrounded vowels. For frontness, the vowels were releveled as *front* (/i, e/) and *back* (/ɑ, o, u/), with the latter as the reference level. For height, the vowels were releveled as *high* (/i, u/) and *non-high* (/e, ɑ, o/), with effects determined relative to non-high vowels. These feature classes were used to interpret the effects in the adults' fitted vowel-context models, and an effect of a

Table 5.1: Fixed effects structure of the vowel-context model for English-speaking adults' peak ERB_N-number trajectories for /s/.

	Likelihood ratio test		
Fixed effect term	*df*	χ^2 statistic	*p*-value (<)
Vowel	4	36.19	0.001
Time^3:Vowel	4	15.14	0.01
Time^5:Vowel	4	16.46	0.01

particular feature was investigated in the children only if it was found to be significant for the adults.

5.3 Results

5.3.1 Community-norm vowel-context effects on English /s/

The base vowel-context model for English /s/ included the fixed effects that appear on the italicized rows of Table 3.3. In addition to these fixed effects, the fitted vowel-context model further included those effects shown in Table 5.1. The fixed-effects coefficients of the vowel-context model fitted to the adults' productions of English /s/ are shown in Table refVowelContext.EnglishAdultsCoefficients. Figure 5.1 shows the peak ERB_N-number trajectories for /s/ in each vowel context, as predicted from the fixed effects of the vowel-context model.

In the fitted vowel-context model for English /s/, the vowel factor had five levels corresponding the five English vowel classes (/i/, /e/, /ɑ/, /o/, /u/), and effects of vowel were estimated with /ɑ/ as the reference level. Each level of the vowel factor had a significant effect on the intercept (/i/: $\hat{\beta} = 0.435$, $SE = 0.222$, $CI = [0.001, 0.870]$; /e/: $\hat{\beta} = 0.608$, $SE = 0.222$, $CI = [0.173, 1.044]$; /o/: $\hat{\beta} = -0.520$, $SE = 0.222$, $CI = [-0.955, -0.085]$; /u/: $\hat{\beta} = -0.494$, $SE = 0.222$, $CI = [-0.928, -0.059]$). Here, the effects for /o/ and /u/ were negative, which suggests that, across its time course, the peak ERB_N of /s/ was on average lower in rounded vowel contexts; conversely, the effects for /i/ and /e/ were positive,

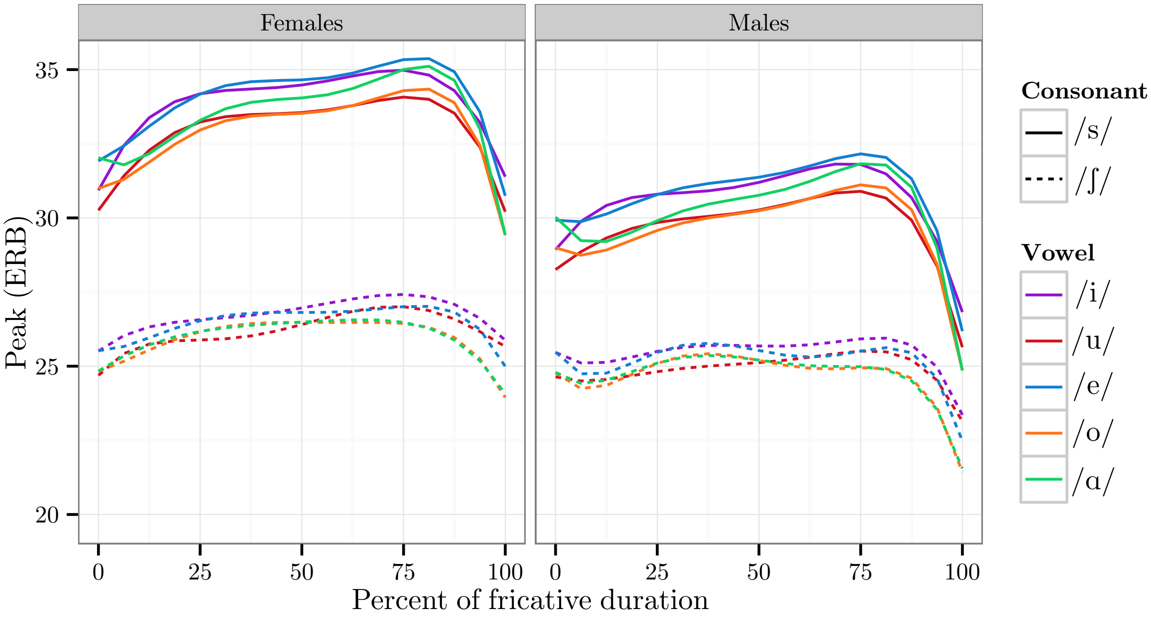
Table 5.2: Fixed-effects coefficients of the vowel-context model fitted to the English-speaking adults' productions of /s/. The reference levels for the vowel and sex factors were /ɑ/ and females, respectively. Shaded rows indicate fixed-effects coefficients whose estimate was not significantly different from zero, as determined by a 95% Wald confidence interval.

Fixed effect term	**Coeff. ($\hat{\boldsymbol{\beta}}$)**	**Std. error**
Intercept	33.412	0.376
Vowel[/i/]	0.435	0.222
Vowel[/u/]	-0.494	0.222
Vowel[/e/]	0.608	0.222
Vowel[/o/]	-0.520	0.222
Sex[Males]	-3.282	0.494
Time	1.250	0.439
Time × Sex[Males]	-1.606	0.620
Time^2	-4.247	0.563
Time^3	-2.762	0.626
Time^3 × Vowel[/i/]	2.218	0.636
Time^3 × Vowel[/u/]	1.971	0.636
Time^3 × Vowel[/e/]	1.171	0.639
Time^3 × Vowel[/o/]	0.829	0.636
Time^3 × Sex[Males]	-1.456	0.678
Time^4	-1.887	0.277
Time^5	-1.686	0.420
Time^3 × Vowel[/i/]	1.620	0.435
Time^3 × Vowel[/u/]	1.190	0.435
Time^3 × Vowel[/e/]	0.566	0.439
Time^3 × Vowel[/o/]	0.425	0.435
Time^3 × Sex[Males]	0.022	0.449

which suggests that on average the peak ERB$_N$ of /s/ was higher before front vowels.

The high vowels, /i/ and /u/, had significant positive effects on both cubic (/i/: $\hat{\beta} = 2.218$, $SE = 0.636$, $CI = [0.972, 3.464]$; /u/: $\hat{\beta} = 1.971$, $SE = 0.636$, $CI = [0.724, 3.217]$) and quintic time (/i/: $\hat{\beta} = 1.620$, $SE = 0.435$, $CI = [0.767, 2.473]$; /u/: $\hat{\beta} = 1.190$, $SE = 0.435$, $CI = [0.336, 2.043]$). Because these effects were positive, they reduced the magnitude of the negative effects of cubic and quintic time, making the tails of the peak ERB$_N$-number trajectory more symmetric in high-vowel contexts. Thus, while the drop in peak ERB$_N$ across the final quarter of /s/ is comparable in all vowel contexts, the rise in

Figure 5.1: Predicted peak ERB$_N$-number trajectories for the English-speaking adults' vowel-context models.



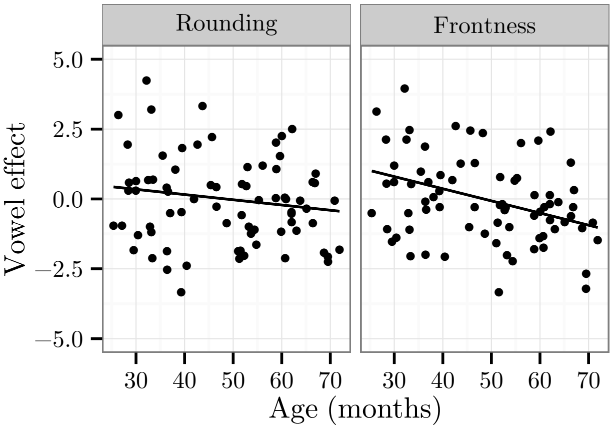
peak ERB$_N$ is steeper in high-vowel contexts than otherwise.

5.3.2 The development of vowel-context effects on English /s/

Rounding

The English-speaking adults' productions of /s/ showed effects of anticipatory rounding on the intercept, with peak ERB$_N$ being lower on average across the duration of the fricative when /s/ occurred before a rounded as opposed to an unrounded vowel. To investigate such rounding effects in the English-acquiring children, a rounding-context model was built for all the children's productions. The structure of this model included the fixed effects listed on the italicized rows in Table 3.3 plus an effect of vowel rounding on the intercept. The fitted model included a significant effect of vowel rounding on the intercept ($\hat{\beta} = -1.113$, $SE = 0.243$, $CI = [-1.589, -0.637]$). The individual rounding effects, shown in the left

Figure 5.2: The English-acquiring children's individual rounding and frontness effects on the intercept, for /s/, plotted against age.



panel of Fig. 5.2, were computed by subtracting a participant's random effect for rounded vowels from their random effect for unrounded vowels. A positive correlation with age would indicate an increase in the strength of rounding-context effects developmentally; however, the children's individual rounding differences exhibited a weak negative association with age ($\tau = -0.092$, $z = -1.175$, $p > 0.12$).

When a rounding-context model was fitted to each age group, the rounding effect was always negative, and its magnitude decreased monotonically as age increased. Furthermore, the effect was significant for every age group except the five-year-olds (two-year-olds: $\hat{\beta} = -1.694$, $SE = 0.539$, $CI = [-2.749, -0.638]$; three-year-olds: $\hat{\beta} = -1.146$, $SE = 0.527$, $CI = [-2.178, -0.114]$; four-year-olds: $\hat{\beta} = -1.105$, $SE = 0.558$, $CI = [-2.198, -0.012]$; five-year-olds: $\hat{\beta} = -0.656$, $SE = 0.344$, $CI = [-1.331, 0.018]$). In the vowel-context model fitted to the adults' productions of /s/, the estimated coefficients for the effects of /o/ and /u/ on the intercept were $\hat{\beta} = -0.520$ and $\hat{\beta} = -0.494$, respectively.

Frontness

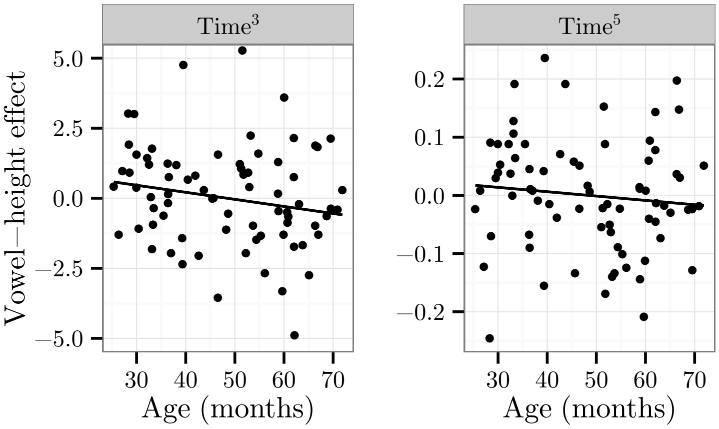
In the English-speaking adults' productions of /s/, peak ERB$_N$ was on average greater, across the time course of the fricative, in the context of front, as opposed to back, vowels. A frontness-context model, which included the fixed effects listed on the italicized rows in Table 3.3 as well as an effect of vowel frontness on the intercept, was fitted to all the children's /s/ productions. The fitted model included a significant, positive effect of frontness ($\hat{\beta} = 0.762$, $SE = 0.234$, $CI = [0.303, 1.221]$). For each participant, an individual frontness effect was computed by subtracting their random effect for back vowels from that for front vowels. These effects are plotted against age in the right panel of Fig. 5.2. A positive association with age would indicate a developmental strengthening of frontness effects. However, the children's individual effects significantly decreased with age ($\tau = -0.206$, $z = -2.637$, $p < 0.01$).

When separate frontness-context models were fitted cross-sectionally, the frontness effect was positive in each age group, suggesting that peak ERB$_N$ was, on average, greater when /s/ occurred before a front vowel. As with rounding, the magnitude of the frontness effect decreased with increasing age. The effect was significant for the two- ($\hat{\beta} = 1.859$, $SE = 0.658$, $CI = [0.570, 3.148]$) and three-year-old children ($\hat{\beta} = 1.202$, $SE = 0.402$, $CI = [0.414, 1.990]$), but not for the four- ($\hat{\beta} = 0.278$, $SE = 0.397$, $CI = [-0.499, 1.056]$) or five-year-old children ($\hat{\beta} = 0.011$, $SE = 0.333$, $CI = [-0.642, 0.665]$). In the fitted vowel-context model for the adults, the estimated coefficients for the effects of /i/ and /e/ on the intercept were $\hat{\beta} = 0.435$ and $\hat{\beta} = 0.608$, respectively.

Height

In the English-speaking adults' productions of /s/, significant, positive effects of vowel height were found on cubic and quintic time, which reduced the magnitude of the negative main effects of these powers of time. A height-context model was built for all the children's productions of /s/. This model included the fixed effects listed on the italicized rows

Figure 5.3: The English-acquiring children's individual height effects on cubic and quintic time, for /s/, plotted against age.



in Table 3.3 and effects of vowel height on cubic and quintic time. In the fitted model, the effect of vowel height was not significant for either cubic ($\hat{\beta} = 0.039$, $SE = 0.527$, $CI = [-0.993, 1.071]$) or quintic time ($\hat{\beta} = 0.220$, $SE = 0.396$, $CI = [-0.557, 0.996]$). An individual height effect was computed by subtracting the random effect for high vowels from that for non-high vowels. The individual height effects on cubic and quintic time are shown in Fig. 5.3. Because the effects of cubic and quintic time are negative, and adjusted positively by an effect of height, a negative correlation between individual effects and age would signal the strengthening of height effects developmentally. The effects on cubic time significantly decreased with age ($\tau = -0.133$, $z = -1.695$, $p < 0.05$); those on quintic time decreased with age, though the association was not significant ($\tau = -0.096$, $z = -1.229$, $p > 0.1$).

When separate height-context models were built for each age group, the effect of vowel height on cubic time increased with age, but did not reach significance for any age group (two-year-olds: $\hat{\beta} = -1.214$, $SE = 1.014$, $CI = [-3.202, 0.774]$; three-year-olds: $\hat{\beta} = 0.135$, $SE = 1.095$, $CI = [-2.011, 2.281]$; four-year-olds: $\hat{\beta} = 0.459$, $SE = 1.162$, $CI = [-1.819, 2.738]$; five-year-olds: $\hat{\beta} = 0.549$, $SE = 0.934$, $CI = [-1.282, 2.380]$). In the adults'

Table 5.3: Fixed effects structure of the vowel-context model for English-speaking adults' peak ERB$_N$-number trajectories for /ʃ/.

	Likelihood ratio test		
Fixed effect term	*df*	χ^2 statistic	*p*-value (<)
Vowel	4	24.60	0.001
Time:Vowel	4	17.74	0.01
Time^2:Vowel	4	10.25	0.05
Time^5:Vowel	4	10.85	0.05

vowel-context model, the estimated coefficients for the effects of the high vowels, /i/ and /u/, on cubic time were $\hat{\beta} = 2.218$ and $\hat{\beta} = 1.971$, respectively; thus, the development of the effect of vowel height, across the children's age groups, seems to tend toward the community-norm.

On the other hand, there was an insignificant, negative effect of vowel height on quintic time for the two- ($\hat{\beta} = -0.883$, $SE = 1.001$, $CI = [-2.845, 1.080]$), three- ($\hat{\beta} = -0.115$, $SE = 0.866$, $CI = [-1.811, 1.582]$) and five-year-olds ($\hat{\beta} = -0.446$, $SE = 0.643$, $CI = [-1.707, 0.815]$). Thus, these age groups had vowel-height effects that were opposite in direction from those of the adults. For the four-year-olds, the effect of vowel height was significant and positive ($\hat{\beta} = 1.837$, $SE = 0.708$, $CI = [0.449, 3.224]$).

5.3.3 Community-norm vowel-context effects on English /ʃ/

The fitted vowel-context model for /ʃ/ comprised the fixed effects listed on the italicized rows of Table 3.3, as well as those listed in Table 5.3. The fixed-effects coefficients of the fitted model are listed in Table 5.4. The peak ERB$_N$-number trajectories, predicted on the fixed effects of this fitted model, are shown in Fig. 5.1. In the vowel-context model for /ʃ/, the vowel factor had five levels, one for each vowel category. As with the community-norm model for /s/, effects of vowel were estimated relative to /ɑ/.

The front vowels, /i/ and /e/, each had a significant positive effect on the intercept (/i/: $\hat{\beta} = 0.761$, $SE = 0.183$, $CI = [0.401, 1.120]$; /e/: $\hat{\beta} = 0.502$, $SE = 0.183$, $CI =$

Table 5.4: Fixed-effects coefficients of the vowel-context model fitted to the English-speaking adults' productions of /ʃ/. The reference levels for the vowel and sex factors were /ɑ/ and females, respectively. Shaded rows indicate fixed-effects coefficients whose estimate was not significantly different from zero, as determined by a 95% Wald confidence interval.

Fixed effect term	**Coeff. ($\hat{\beta}$)**	**Std. error**
Intercept	25.954	0.373
Vowel[/i/]	0.761	0.183
Vowel[/u/]	0.217	0.183
Vowel[/e/]	0.502	0.183
Vowel[/o/]	−0.032	0.183
Sex[Males]	−1.282	0.501
Time^{1}	−0.114	0.474
Time^{1} × Vowel[/i/]	1.248	0.485
Time^{1} × Vowel[/u/]	1.744	0.485
Time^{1} × Vowel[/e/]	0.681	0.485
Time^{1} × Vowel[/o/]	0.103	0.485
Time^{1} × Sex[Males]	−1.571	0.512
Time^{2}	−2.680	0.499
Time^{2} × Vowel[/i/]	1.053	0.496
Time^{2} × Vowel[/u/]	1.111	0.496
Time^{2} × Vowel[/e/]	0.660	0.496
Time^{2} × Vowel[/o/]	−0.078	0.496
Time^{3}	−0.660	0.229
Time^{3} × Sex[Males]	−0.694	0.324
Time^{4}	−0.703	0.177
Time^{5}	−0.123	0.334
Time^{5} × Vowel[/i/]	0.215	0.343
Time^{5} × Vowel[/u/]	0.541	0.343
Time^{5} × Vowel[/e/]	−0.459	0.343
Time^{5} × Vowel[/o/]	−0.300	0.343
Time^{5} × Sex[Males]	−0.916	0.360

$[0.143, 0.861]$), indicating that, across the time course of /ʃ/, peak ERB$_N$-number is higher on average before front vowels.

The high vowels, /i/ and /u/, had significant positive effects on linear (/i/: $\hat{\beta} = 1.248$, $SE = 0.485$, $CI = [0.297, 2.199]$; /u/: $\hat{\beta} = 1.744$, $SE = 0.485$, $CI = [0.793, 2.696]$) and quadratic time (/i/: $\hat{\beta} = 1.053$, $SE = 0.496$, $CI = [0.081, 2.024]$; /u/: $\hat{\beta} = 1.111$, $SE = 0.496$, $CI = [0.139, 2.082]$). The effect of vowel height on linear time indicates that before high vowels, the peak ERB_N-number trajectory of /ʃ/ is increasing across its midpoint, while before mid and low vowels it is decreasing. The effect of vowel height on quadratic time suggests that the peak ERB_N-number trajectory of /ʃ/ is less concave before high vowels. From the trajectories plotted in Fig. 5.1, this difference in concavity may be due to peak ERB_N-number exhibiting a relatively shallower drop across the second half of /ʃ/ before high vowels.

Despite the model fit being significantly improved by an interaction between vowel and quintic time, there were no significant effects of vowel on quintic time in the fitted model (/i/: $\hat{\beta} = -0.459$, $SE = 0.343$, $CI = [-1.131, 0.214]$; /e/: $\hat{\beta} = 0.215$, $SE = 0.343$, $CI = [-0.459, 0.887]$; /o/: $\hat{\beta} = -0.300$, $SE = 0.343$, $CI = [-0.972, 0.373]$; /u/: $\hat{\beta} = 0.541$, $SE = 0.343$, $CI = [-0.131, 1.214]$).

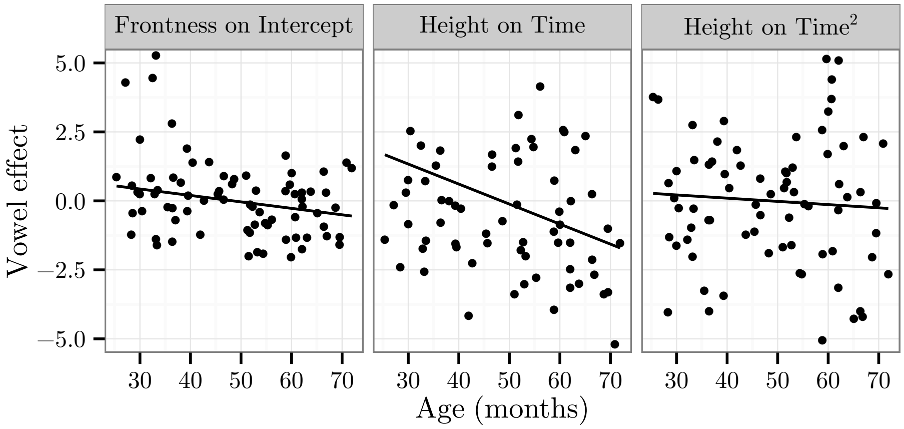
5.3.4 The development of vowel-context effects on English /ʃ/

Frontness

The English-speaking adults' productions of /ʃ/ were subject to an effect of vowel frontness on the intercept, with peak ERB_N being greater on average when the sibilant occurred before a front, as opposed to a back, vowel. When a frontness-context model was built for all the children's productions of /ʃ/, the fitted model included a significant, positive effect of vowel frontness ($\hat{\beta} = 0.705$, $SE = 0.204$, $CI = [0.305, 1.105]$). Individual frontness effects, shown in the left panel of Fig. 5.4, were computed in the same way as for /s/, by subtracting the random effects for back vowels from those of front vowels. The individual frontness effects were significantly, negatively associated with age ($\tau = -0.142$, $z = -1.812$, $p < 0.05$), indicating that the effect of vowel frontness weakened developmentally.

A separate frontness-context model was fitted to the data from each age group. The effect of frontness was positive for each age group, and its magnitude tended to decrease with increasing age; however, the effect only reached significance for the three-year-olds (two-year-olds: $\hat{\beta} = 1.316$, $SE = 0.719$, $CI = [-0.093, 2.726]$; three-year-olds: $\hat{\beta} = 1.144$, $SE =$

Figure 5.4: The English-acquiring children's individual vowel effects for /ʃ/, plotted against age.



0.300, $CI = [0.556, 1.731]$; four-year-olds: $\hat{\beta} = 0.163$, $SE = 0.264$, $CI = [-0.354, 0.680]$; five-year-olds: $\hat{\beta} = 0.406$, $SE = 0.248$, $CI = [-0.081, 0.893]$). As was found for the effect of frontness on /s/, the estimate of the effect was smaller in four- and five-year-olds than it was in the adults (/i/: $\hat{\beta} = 0.761$; /e/: $\hat{\beta} = 0.502$), suggesting that some groups of children exhibit smaller effects of vowel frontness than adults.

Height

In the adults' community-norm vowel-context effects model for /ʃ/, there were significant, positive effects on linear and quadratic time for the high vowels, /i/ and /u/. A height-context model, which included the fixed effects listed on the italicized rows in Table 3.3 as well as effects of vowel height on linear and quadratic time, was fitted to all the children's /ʃ/ productions. In this model, the effect of vowel height was not significant on linear time ($\hat{\beta} = -0.176$, $SE = 0.590$, $CI = [-1.333, 0.981]$), but was significant on quadratic time ($\hat{\beta} = 1.385$, $SE = 0.482$, $CI = [0.440, 2.329]$). Individual height effects, shown in the center and right panels of Fig. 5.4, were computed by subtracting the random effects for high vowels from those of non-high vowels; thus, a negative association with age would indicate that the effect of vowel height strengthens developmentally. The association with age was indeed negative for individual height effects on both linear ($\tau = -0.157$, $z = -2.001$, $p < 0.05$) and quadratic time ($\tau = -0.0267$, $z = -0.341$, $p > 0.36$), but it only reached significance for the former.

Separate height-context models were built to investigate how the magnitude of the fixed effect varied across the age groups. The effect of vowel height on linear time was not significant for any age group, but its coefficient estimate increased with age (two-year-olds: $\hat{\beta} = -2.669$, $SE = 1.796$, $CI = [-6.190, 0.851]$; three-year-olds: $\hat{\beta} = -0.049$, $SE = 0.956$, $CI = [-1.923, 1.825]$; four-year-olds: $\hat{\beta} = 0.277$, $SE = 0.973$, $CI = [-1.629, 2.183]$; five-year-olds: $\hat{\beta} = 1.256$, $SE = 0.868$, $CI = [-0.445, 2.957]$). On linear time, the adults' effects of /i/ and /u/ were estimated to be $\hat{\beta} = 1.248$ and $\hat{\beta} = 1.744$, respectively. Thus, in the two youngest groups of children, the effect of vowel height on linear time is in the opposite direction as the adults. The older children, however, show vowel height effects on linear time that are in the same direction as the adults' effects.

The effect of vowel height on quadratic time was positive for each age group, but did not show any apparent developmental trend (two-year-olds: $\hat{\beta} = 1.231$, $SE = 1.474$, $CI = [-1.658, 4.119]$; three-year-olds: $\hat{\beta} = 1.060$, $SE = 0.653$, $CI = [-0.220, 2.340]$; four-year-olds: $\hat{\beta} = 1.501$, $SE = 0.742$, $CI = [0.046, 2.956]$; five-year-olds: $\hat{\beta} = 1.435$, $SE = 0.875$, $CI = [-0.280, 3.150]$).

5.4 Discussion

Vowel-context effects on the peak ERB$_N$-number trajectories of /s/ and /ʃ/ were investigated in adults' and young children's productions. Each consonant was investigated separately given that previous research has found consonant-specific coarticulatory effects. The hypothesis that vowel context would affect both global and time-varying aspects of peak ERB$_N$-number trajectory was confirmed for the adults, in both consonants.

In the adults' productions of /s/, the intercept of the peak ERB_N-number trajectory was lower before rounded vowels and higher before front vowels. Because these effects are on the intercept, it is suggested that they derive from anticipatory coarticulatory effects that begin quite early on in the production of the sibilant-vowel syllable. For example, the effect of rounded vowels suggests a greater extent of lip protrusion across the full duration of /s/, in the context of /u/ and /o/, relative to /i/, /e/, and /ɑ/ (cf. Toda et al., 2002). Likewise, the effect of vowel frontness suggests a more anterior constriction from the onset of frication when /s/ precedes either /i/ or /u/ (cf. Iskarous et al., 2011, Fig. 5, p. 949).

Additionally, the adults' peak ERB_N-number trajectories for /s/ were subject to effects of vowel height on cubic and quintic time. Although these effects operate on powers of polynomial time, they seem to derive from a different source than what is typically denoted in the literature by a temporal coarticulatory effect. In particular, a temporal coarticulatory effect is usually identified by a point in time after which the position of an articulator differs according to the following phonetic context. From Fig. 5.1, though, the effects of vowel height on polynomial time seem to indicate differences at the onset of frication, where the increase in peak ERB_N was steeper before high vowels than before low vowels.

There are at least two possible explanations for the effect of vowel height on the shape of peak ERB_N-number trajectory. The first is durational. If /s/ is longer before high vowels, then the rise in peak ERB_N near frication onset would occur over a longer absolute duration of time. Thus, the time-normalization procedure may have artificially increased the slope of the peak ERB_N-number rise before high vowels. To explore this possibility, durations of the adults' productions of /s/ in high- and non-high-vowel contexts were compared with a linear mixed-effects model that included random adjustments to the intercept for each subject. It was found that /s/ was slightly shorter in duration before non-high vowels ($M = 183.936$, $SD = 46.625$) than before high vowels ($M = 186.963$, $SD = 41.696$), but the effect of height in the fitted model was not significant ($\hat{\beta} = 2.888$, $SE = 3.700$, $CI = [-4.363, 10.140]$). Thus, this effect does not seem to be due to durational differences of /s/ that are conditioned by the following vocalic context.

The second explanation concerns the coordination of the tongue and jaw movements during the production of /s/. The effects of vowel height could be due to the constriction aperture being narrower before high vowels, which would increase the velocity of the turbulent jet, causing a build-up of high-frequency noise at the incisors to occur sooner than it does in low-vowel contexts, but because the generation of the obstacle noise source at the incisors requires coordinated lingual and mandibular gestures, it is impossible to know for certain without articulatory data.

In the adults' productions of /ʃ/, the intercept of the peak ERB_N-number trajectory was subject to an effect of vowel frontness: across the sibilant's duration, peak ERB_N was on average higher before front vowels. Unlike /s/, no effect of vowel rounding was found, which was likely due to the fact that /ʃ/ is articulated with lip rounding even in isolation (Narayanan et al., 1995).

Vowel height was found to affect linear and quadratic time. These effects may derive from conventional temporal coarticulatory effects. From the predicted peak ERB_N-number trajectories in Fig. 5.1, the trajectories for the high-vowel contexts and those for the low-vowel contexts begin to diverge around halfway through /ʃ/, after which point peak ERB_N continues to rise in high-vowel contexts, but decrease in low-vowel contexts. This could be due to differences in the height of the jaw, which lowers more in low-vowel contexts (Iskarous et al., 2011).

The hypothesis that vowel-context effects would be found on global and dynamic aspects of peak ERB_N-number trajectory was not confirmed for the children, for either sibilant. In the vowel-rounding and -frontness context models built for the children's productions of /s/, pooled across age groups, the effect on the intercept of rounded vowels and front vowels, respectively, was significant and in the same direction as would be expected from the adults' model. The same was true for the effect of vowel rounding on the intercept of the peak ERB_N-number trajectory for /ʃ/. Vowel height, however, did not have a significant effect on any power of time for either consonant, indicating that the peak ERB_N-number trajectories for /s/ and /ʃ/ followed similar courses regardless of the following vowel.

The children's individual vowel-feature effects exhibited an interesting asymmetry in both consonants. For /s/, the effects of vowel rounding and vowel frontness on the intercept tended to weaken developmentally, while the effects of vowel height on cubic and quintic time tended to strengthen as children matured. For /ʃ/, the effects of frontness on the intercept weakened as age increased, but the effects of vowel height on linear and quadratic time became stronger with age. These results seem to point in different directions if vowel-context effects are considered indicative of the size of speech planning unit. That is, the effects on intercept suggest a developmental shift from syllabically-planned to segmentally-planned motor production, but the effects on time suggest a shift in the other direction. Thus, the vowel-context effects seem to be more an index of the maturation of articulator control over fluidly coordinated gestures that affect different articulations, as was suggested by Zharkova et al. (2014). The strengthening of the vowel height effect could then be understood as requiring a level of jaw-tongue coordination, the development of which is ongoing in children as they age.

Chapter 6

Sibilant contrast by children with cochlear implants

6.1 Introduction

6.1.1 Sibilant production deficits in pediatric CI users

Studies of the intelligibility rates of pediatric CI users' productions of /s/ and /ʃ/ suggest that their acquisition of these consonants lags behind that of children with normal hearing (NH). In a longitudinal study of phonological development in pediatric CI users, which tracked their consonant intelligibility from six months pre-implant to six years post-implant, it was found that target /ʃ/ was produced intelligibly at least 50% of the time by 48 months post-implant (Serry and Blamey, 1999); however, target /s/ was not produced with comparable intelligibility even after 72 months of implant use (Blamey et al., 2001). Similar results were found by Chin (2003), who rated the intelligibility of consonants produced by pediatric CI users, who had at least five years of experience with their implant. Using a threshold intelligibility rate of 75% to determine the acquisition of a consonant, it was found that /ʃ/ was acquired by five years post-implant, but /s/ was not. In their cross-sectional study of children with normal hearing, Smit et al. (1990) found that, when pooled across gender, the children's productions of /s/ and /ʃ/ were both intelligible 50% of the time by age three. The 75% intelligibility rate was reached by age five for /s/, and age four for /ʃ/. Thus, when comparing the duration of the CI users' experience to the chronological age of the children with normal hearing, the acquisition of /ʃ/ by the CI users' appears to lag that of NH children by approximately a year, but /s/ is delayed even longer.

Transcription and acoustic analyses have also indicated that the sibilant contrast is poorer in pediatric CI users than in their normal hearing peers. Uchanski and Geers (2003) recorded productions of word-initial target /s/ and /ʃ/ made by eight- and nine-year-old pediatric CI users with prelingual profound deafness and who had at least four years of experience with their prosthetic. Productions of the same words were elicited from a cohort of NH controls of the same age. The children's productions were phonetically transcribed and centroid frequency was estimated from each target production whose manner of articulation was a fricative. All of the NH children produced target /s/ and /ʃ/ as fricatives, but only 49% of the children with CIs did so. From each NH child's productions, the mean centroid was computed separately for /s/ and /ʃ/, as was the difference between these two means. These values were used to demarcate the normal limits of the spectral properties of the sibilant categories and the sibilant contrast. It was found that 63% of the children with CIs produced /s/ within normal limits, while 86% produced /ʃ/ within normal limits. In terms of the distance between the mean centroid of each sibilant, i.e. the cross-category difference, 71% of the pediatric CI users were found to be within normal limits.

One limitation of Uchanski and Geers's (2003) analysis is that the NH children were matched to the pediatric CI users in terms of chronological age; however, since pediatric CI users undergo a period of auditory deprivation before their hearing is restored through implantation, it is possible that the children with CIs were just delayed relative to their chronological-age matched peers due to this interval of auditory deprivation. To investigate this possibility, Todd et al. (2011) compared the /s/ and /ʃ/ productions of a group of children with CIs to a group of NH controls, chosen so that across groups the duration of the former's experience with the CIs matched the latter's chronological age. In other words, the two groups were matched on hearing age, this being identical to chronological age in children with normal hearing. Both groups of children's productions were transcribed phonetically, and accuracy was determined as a narrow phonetic match to the target sibilant. There was no significant difference between the CI and NH groups in terms of the transcribed accuracy of either sibilant. To compare the two groups' productions at the subphonemic acoustic level, peak frequency was computed from the middle 40 ms of each phonetically correct production. An interaction between group and consonant was found in the peak frequency measures: The cross-category distance between /s/ and /ʃ/ was smaller in the CI children's productions, indicating that their productions exhibited less acoustic contrast than the NH children's productions.

6.1.2 Purpose of analyses

The pediatric CI users analyzed in Todd et al. (2011) were between the age of four and nine years, and their duration of CI experience ranged from two to five years. Because that study's research questions pertained only to the performance of children with CIs relative to NH peers, the cohort of CI users was treated as a monolith. However, the results presented in chapter 4 suggest that the acoustic differentiation of /s/ and /ʃ/ develops on both static and dynamic aspects of peak ERB_N-number trajectory as normal-hearing children mature. Furthermore, a number of studies have found that the intelligibility of pediatric CI users' speech improves with prolonged experience with their prosthetic (e.g., Chin et al., 2003; Peng et al., 2004); thus, the spectral contrast between /s/ and /ʃ/ would be expected to improve with increased hearing age. The analyses in this chapter investigated whether the sibilant contrast likewise increased with hearing age, and if so, whether the CI children developed with respect to the same aspects of peak ERB_N-number trajectory as was observed in NH children.

Moreover, the investigation of the development of speech outcomes in pediatric CI users is mildly complicated by their absence of auditory input prior to implantation. The age at which a child with prelingual profound deafness receives a cochlear implant has been found affect the development of their speech and spoken language outcomes. For example, Hammes et al. (2002) used a battery of clinical instruments to assess the spoken language skills of a group of pediatric CI users who were implanted before their fourth birthday and who had at least six months of experience with their implant. The children's performance on the clinical assessments were used to determine their spoken-language age, and it was found that the earlier a child was implanted, the more likely that their spoken-language age was within one year of their chronological age. Similarly, Nicholas and Geers (2007) measured the number of different root words used spontaneously during a semistructured task, and found a significant effect of age at implantation over and above that of duration of experience with a CI. The extent of the acoustic differentiation of sibilant fricatives, though, is a much finer-grained speech outcome measure than has been explored in the literature; thus, the analyses in this chapter also explored if significant effects of age at implant are reflected in the spectral details of pediatric CI users' sibilant contrast.

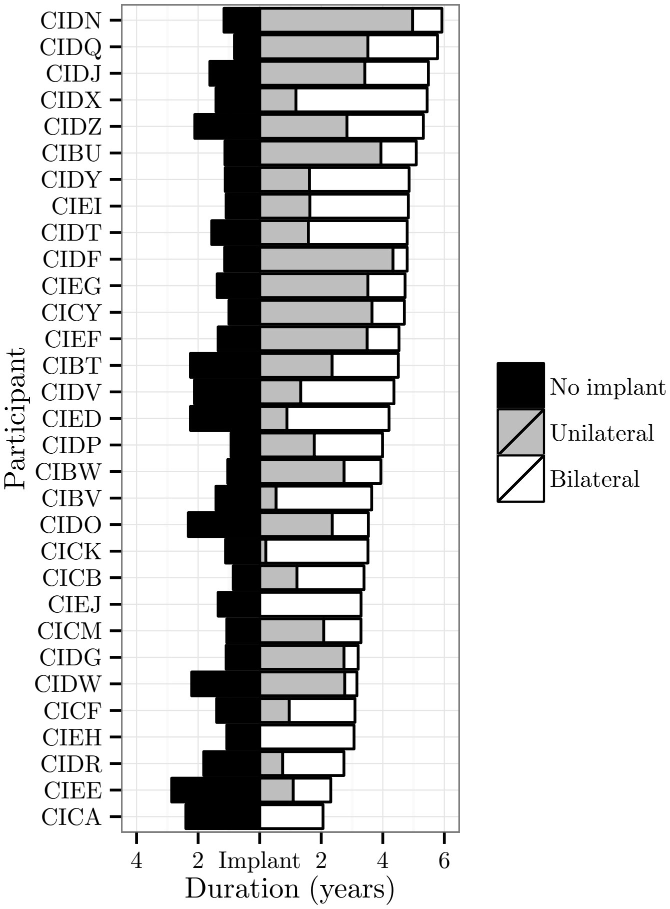
Finally, when pooled across age group, the English-acquiring children with normal hearing reported in chapter 4 were found to differentiate the sibilant fricatives in terms of both static (i.e. on the intercept) and dynamic (i.e. on quadratic and quartic time) aspects of peak ERB_N-number trajectory. However, Todd et al. (2011) examined group-related differences between hearing-age matched CI and NH children only in terms of a static measure of peak frequency. Thus, the analyses in this chapter extend their analysis to examine whether group-related differences also arise on how the spectral dynamic differences between /s/ and /ʃ/.

6.2 Method

6.2.1 Participants

Thirty-eight pediatric bilateral cochlear implant users, who were born congenitally deaf, participated in the picture-prompted word-repetition task. The participants were recruited from throughout the United States and tested at the University of Wisconsin–Madison. At the time of test, each child's speech, language, and developmental history was ascertained through parental report, from which, one child was excluded due to a suspected diagnosis

Figure 6.1: The implant histories of the children with bilateral cochlear implants, who participated in the word-repetition task.



of Usher syndrome, and two children were excluded because English was not their primary language. One child was excluded because they received their implant at age five, and two children were excluded because the duration of experience with their device exceeded five years. One final child was excluded because they were missing their incisors at the time of test. The implant histories of the remaining 31 participants are shown in Fig. 6.1.

6.2.2 Materials

The sibilant-initial target words that were used in the repetition task completed by the children with CIs were a subset of the ones listed in Table 3.1. In particular, the pediatric CI users were tested only on sibilant-initial words that preceded either /i/, /ɑ/, or /u/. The word list for the children with CIs also included stop-initial filler words, which were intermixed with the target words.

6.2.3 Elicitation and recording procedure

Testing took place inside a sound-attenuated room, in which the children were seated at a table, facing a computer screen, loud speakers, and a microphone. Prior to the task, the children were instructed that they would be shown images and played recordings of spoken words, and that they should repeat those words into the microphone. Furthermore, practice trials were completed before testing, so that the children would be familiar with the procedure.

The children completed the repetition task in the presence of an adult experimenter, who controlled the custom software program that presented the audiovisual stimulus of each trial (see §4.2.3). At the initiation of each trial, the visual stimulus was presented on the computer screen, and after a 300 ms delay, the audio prompt was played through Audix PH5-VS loud speakers, whose frequency range spanned 0.075 to 20 kHz. In some instances, a child's repetition overlapped the audio stimulus or was too quiet, in which case the prompt was replayed to elicit an audible, isolated production. The children's repetitions were recorded digitally at 44.1 kHz onto a Marantz PMD660 flash recorder for subsequent annotation and analysis.

6.2.4 Transcription and annotation of fricative events

A trained phonetician who was enrolled in a communication sciences and disorders graduate program, but who had no prior exposure to the speech of cochlear implant users, transcribed the initial sibilant of each target word using a custom Praat script. This script allowed the transcriber to listen to the auditory signal of a target repetition and to visualize its waveform and spectrogram before transcribing it.

For trials where more than one repetition of the target word was elicited, the transcriber first chose which repetition to transcribe. Here, the transcriber was instructed to choose the earliest audible attempt at the target word; a speech production error was not reason to skip an earlier repetition for a later, correct production.

Next, the transcriber judged the consonant's type as either a 'sibilant fricative', a 'sibilant affricate', a 'non-sibilant fricative', a 'non-sibilant plosive', or 'other'. This last category was a catch-all for any production that was not easily classified into any of the other categories, and was used on only one trial. If the consonant was judged to be either a 'sibilant fricative' or a 'sibilant affricate,' then the onset of frication and the fricative-vowel boundary were marked using the same criteria described in §3.2.4.

After transcription and annotation, two subjects (CIDW and CIEJ) were excluded because they produced no sibilant productions for target /s/. The remaining 29 subjects' age at implantation ranged from 0;9 to 2;10 years;months (M = 1;5). Their hearing age, defined as the sum of the durations of their unilateral and bilateral hearing experience, ranged from 2;0 to 5;11 (M = 4;2), and their chronological age ranged from 4;1 to 7;5 (M = 5;7).

Of the remaining 29 subjects' productions, 43 were excluded because they were judged to be non-sibilant fricatives, 30 were excluded because they were non-sibilant plosives, and one was excluded because its consonant type could not be determined. Fourteen sibilant productions were excluded because they were either shorter than 60 ms or longer than 500 ms. This left a total of 185 sibilant productions of target /s/ and 249 of target /ʃ/.

6.2.5 Normal-hearing age matches

Twenty-eight of the pediatric CI users were paired with English-acquiring children with normal hearing who completed the word-repetition task, reported in chapter 4. Each pair was matched on sex and hearing age, within three months. A Wilcoxon signed-rank test found no difference between the hearing ages of the pediatric CI users and the ages of NH controls ($W = 174.5$, $z = -0.649$, $p > 0.52$).

6.3 Results

6.3.1 Hearing experience and sibilant contrast in pediatric CI users

A consonant-contrast model, whose fixed effects included those listed in Table 3.3, was fitted to the productions of the pediatric CI users; the fitted model's fixed-effects coefficients are listed in Table 6.1. The predictions made on the fixed effects of the fitted model are shown in Fig. 6.2. The fitted model included significant, negative effects of each power of polynomial time. A significant effect of linear time ($\hat{\beta} = -3.092$, $SE = 0.826$, $CI = [-4.711, -1.472]$) indicated that peak ERB$_N$-number tended to decrease across the duration of the sibilant fricative. Significant effects of quadratic ($\hat{\beta} = -7.716$, $SE = 0.576$, $CI = [-8.844, -6.587]$) and quartic time ($\hat{\beta} = -3.700$, $SE = 0.577$, $CI = [-4.832, -2.569]$) indicated that peak ERB$_N$-number followed a concave trajectory. Significant effects of cubic ($\hat{\beta} = -4.021$, $SE = 0.593$, $CI = [-5.183, -2.859]$) and quintic time ($\hat{\beta} = -2.488$, $SE = 0.534$, $CI = [-3.536, -1.441]$) indicated that the right tail of the trajectory dropped more steeply than the right tail, and that there were local maxima and minima within the middle half of the trajectory.

Effects of consonant were evaluated relative to /s/. There was a significant, negative effect of consonant on the intercept ($\hat{\beta} = -3.414$, $SE = 0.520$, $CI = [-4.433, -2.395]$), which indicated that, across the duration of the sibilants, peak ERB$_N$-number was on average lower for /ʃ/ than for /s/. A significant, positive effect of consonant on quadratic time ($\hat{\beta} = 1.634$, $SE = 0.482$, $CI = [0.690, 2.579]$) indicated that the peak ERB$_N$-number trajectory for /ʃ/ was less concave than the one for /s/. Neither of the effects of consonant on cubic or quartic time was significant.

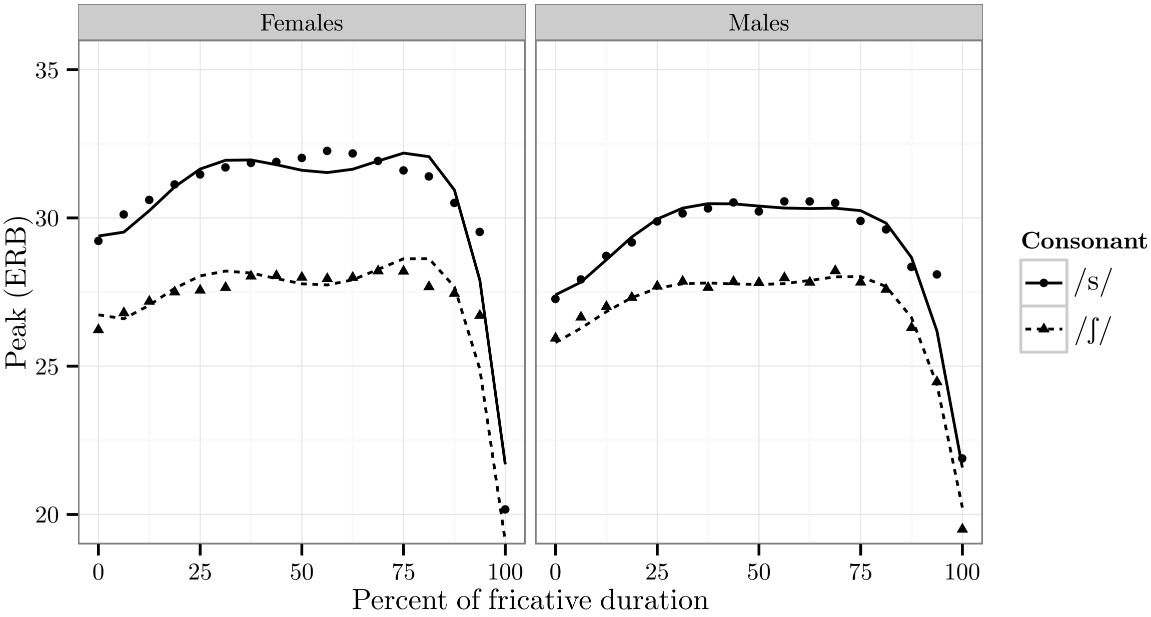
Table 6.1: Fixed-effects coefficients of the consonant-contrast model fitted to the sibilant productions of the English-acquiring children who have cochlear implants. The reference levels of the consonant and sex factors were /s/ and females, respectively. Shaded rows indicate fixed-effects coefficients whose estimate was not significantly different from zero, as determined by a 95% Wald confidence interval.

Fixed effect term	**Coeff. ($\hat{\beta}$)**	**Std. error**
Intercept	30.530	0.478
Consonant[/ʃ/]	−3.414	0.520
Sex[Males]	−1.571	0.688
Consonant[/ʃ/] × Sex[Males]	1.256	0.740
Time	−3.092	0.826
Time × Sex[Males]	0.385	1.046
Time^2	−7.716	0.576
Time^2 × Consonant[/ʃ/]	1.634	0.482
Time^3	−4.021	0.593
Time^3 × Consonant[/ʃ/]	−0.025	0.498
Time^3 × Sex[Males]	1.064	0.675
Time^4	−3.700	0.577
Time^4 × Consonant[/ʃ/]	0.114	0.636
Time^4 × Consonant[/s/] × Sex[Males]	1.310	0.794
Time^4 × Consonant[/ʃ/] × Sex[Males]	1.038	0.718
Time^5	−2.488	0.534
Time^5 × Sex[Males]	0.932	0.774
Time^5 × Consonant[/ʃ/] × Sex[Females]	0.127	0.625
Time^5 × Consonant[/ʃ/] × Sex[Males]	0.428	0.685

The random effects of the fitted consonant-contrast model were used to compute each participant's individual consonant effect on the intercept and on quadratic, cubic, and quartic time. These individual effects are shown, plotted against age at implant, chronological age, and hearing age, in Fig. 6.3. Greater values indicate stronger individual effects on intercept; whereas, smaller values indicate stronger effects on the powers of time.

Age at implant was found to have a very weak association with each of the effects of consonant (intercept: $\tau = -0.040$; quadratic: $\tau = 0.010$; cubic: $\tau = -0.010$; quartic: $\tau = -0.040$). Chronological age was significantly, positively associated with the effect on intercept ($\tau = 0.286$, $z = 2.160$, $p < 0.05$), none of its associations with the effects on

Figure 6.2: Predicted peak ERB$_N$-number trajectories from the English-acquiring pediatric CI users' consonant-contrast model.

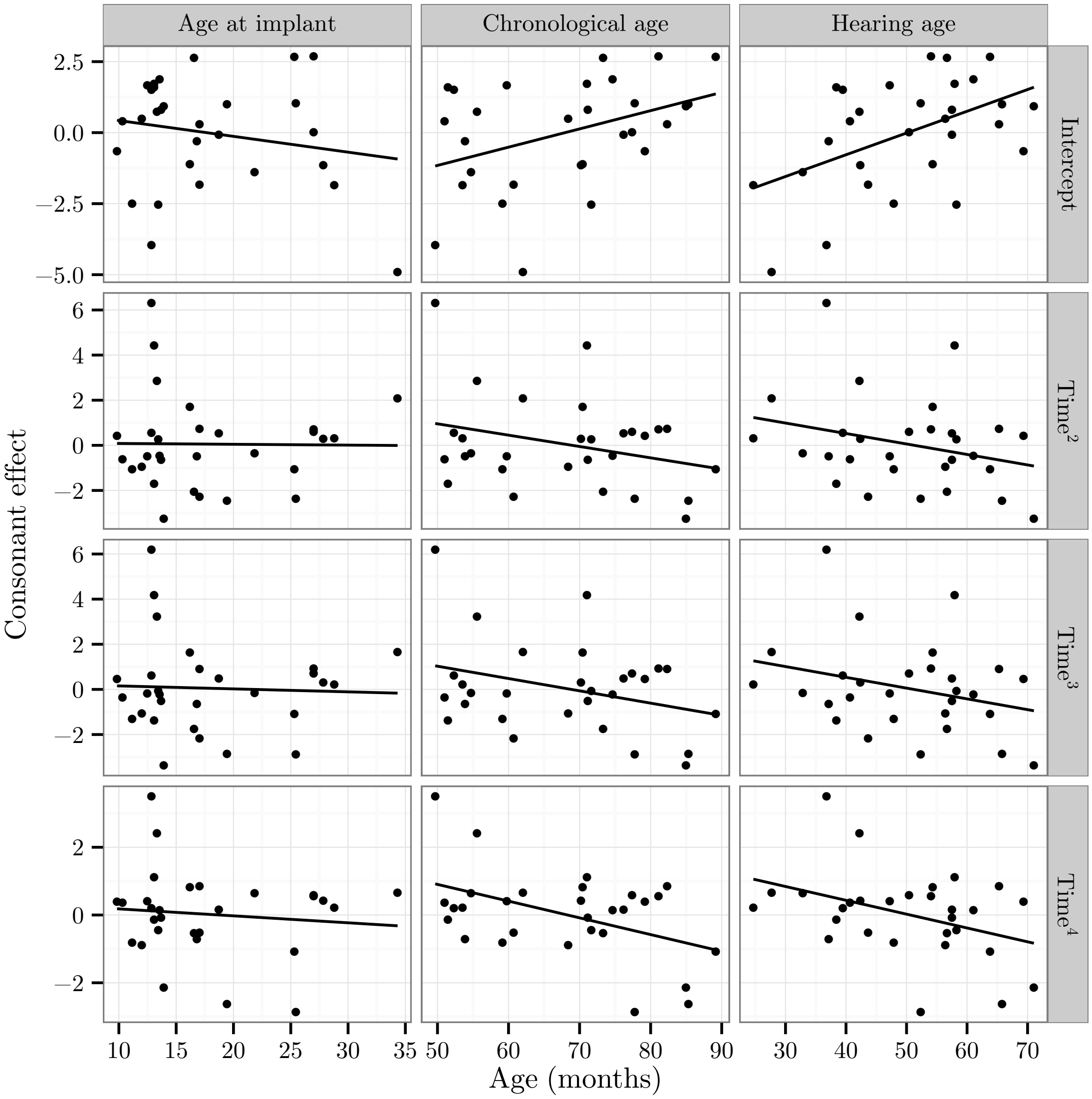


quadratic ($\tau = -0.123$), cubic ($\tau = -0.113$), or quartic time ($\tau = -0.177$) were significant ($p > 0.09$). Hearing age had a significant, positive association with the effect on intercept ($\tau = 0.284,\ z = 2.158,\ p < 0.05$). None of the associations between hearing age and the powers of time (quadratic: $\tau = -0.180$; cubic: $\tau = -0.160$; quartic: $\tau = -0.215$) were significant ($p > 0.051$).

6.3.2 Comparison with hearing-age matched normal-hearing peers

Differences in peak ERB$_N$-number trajectory between CI and NH children were investigated by fitting a consonant contrast model whose fixed effects included a main effect of and interactions involving a group factor. The fixed-effects structure of this model was built-up using a stepwise forward selection protocol. The base model involved all those fixed effects in the adults' community-norm consonant-contrast model that did not involve talker sex. The decision to exclude effects of talker sex was made to simplify the model structure and

Figure 6.3: The pediatric CI users' individual consonant effects on the intercept and on quadratic, cubic, and quartic time, plotted against age at implant, chronological age, and hearing age. In each panel, a linear regression model is fitted to the data.



because this factor was controlled across the two groups. The addition of a main effect of or interaction with the group factor was determined with a likelihood ratio test. Effects of group were first considered on the intercept and then on increasing powers of polynomial time. The interactions between group, consonant, and either the intercept or a power of time

Table 6.2: Fixed-effects structure of the consonant-contrast model fitted to the productions of the pediatric cochlear implant users and their hearing-age matched peers with normal hearing. Shaded rows denote the terms involving an interaction between consonant and group, none of which significantly improved the model fit.

	Likelihood ratio test		
Fixed effect	**Deg. freedom**	**χ^2 statistic**	***p*-value**
Group	1	4.488	< 0.05
Consonant × Group	1	0.015	> 0.90
Time^2 × Consonant × Group	2	5.350	> 0.06
Time^3 × Group	1	8.158	< 0.01
Time^3 × Consonant × Group	1	0.006	> 0.93
Time^4 × Consonant × Group	2	4.583	> 0.10
Time^5 × Group	1	8.176	< 0.01
Time^5 × Consonant × Group	2	1.260	> 0.53

were of particular interest as these would indicate a difference in the extent of consonant differentiation across groups.

The results of model building are summarized in Table 6.2. This table includes those terms that improved the model fit (unshaded rows) and terms that included an interaction between consonant and group (shaded rows), none of which significantly improved the model fit. The likelihood ratio test whose results are reported on a given row r compared: (1) the model whose fixed effects included those of the base model, the effect on row r, and any effects listed on unshaded rows above r; to (2) the model just described less the effect on row r. For example, the test reported on the third row compared Base Model + Group to Base Model + Group + (Time^2 × Consonant × Group). The fixed-effects coefficients of the fitted model are shown in Table 6.3.

The peak ERB$_N$-number trajectories predicted by the fixed effects of the fitted model are shown in Fig. 6.4. The fitted model included a significant, negative effect of quadratic time ($\hat{\beta} = -9.002$, $SE = 0.517$, $CI = [-10.016, -7.989]$, indicating that the peak ERB$_N$-number trajectory was concave for both groups. A significant, negative effect of cubic time ($\hat{\beta} = -1.358$, $SE = 0.480$, $CI = [-2.298, -0.418]$) indicated that the right tail fell more

Table 6.3: Fixed-effects coefficients of the hearing-status model comparing the sibilant productions of English-acquiring children with normal hearing (NH) and those with cochlear implants. The reference levels of the consonant and group factors were /s/ and NH, respectively. Shaded rows indicate fixed-effects coefficients whose estimate was not significantly different from zero, as determined by a 95% Wald confidence interval.

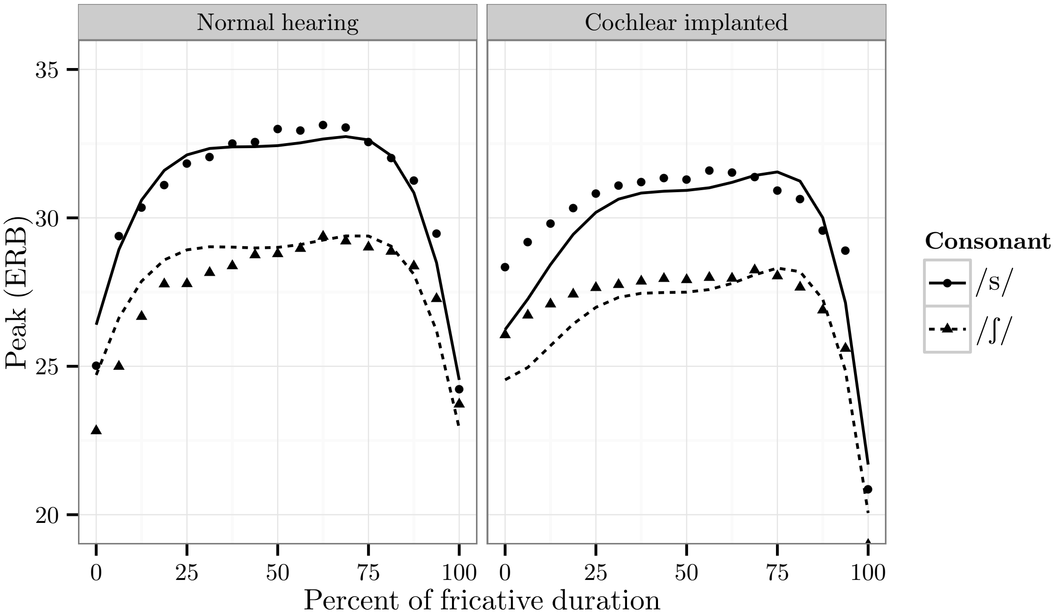
Fixed effect term	**Coeff. ($\hat{\beta}$)**	**Std. error**
Intercept	29.417	0.399
Consonant[/ʃ/]	−2.918	0.308
Group[CI]	−1.508	0.473
Time	−0.501	0.781
Time^2	−9.002	0.517
Time^2 × Consonant[/ʃ/]	2.352	0.469
Time^3	−3.368	0.477
Time^3 × Consonant[/ʃ/]	0.067	0.397
Time^3 × Group[CI]	−2.010	0.509
Time^4	−3.404	0.341
Time^4 × Consonant[/ʃ/]	0.470	0.403
Time^5	−1.708	0.319
Time^5 × Group[CI]	−1.342	0.412

steeply than the left tail rose. There was also a significant, negative effect on quadratic time ($\hat{\beta} = -3.404$, $SE = 0.341$, $CI = [-4.072, -2.736]$), which further indicated concave curvature, and local maxima across the middle half of the trajectory.

Effects of consonant were evaluated with /s/ as the reference level. A significant, negative main effect of consonant on the intercept ($\hat{\beta} = -2.918$, $SE = 0.308$, $CI = [-3.522, -2.314]$) indicated that across the duration of the sibilants, peak ERB$_N$-number was on average lower in /ʃ/ than in /s/. The fitted model also included a significant, positive effect of consonant on quadratic time ($\hat{\beta} = 2.352$, $SE = 0.469$, $CI = [1.433, 3.270]$), indicating that peak ERB$_N$-number trajectory was less curved for /ʃ/ relative to /s/.

Effects of group were evaluated with the normal-hearing children as the reference level. There was a significant, negative main effect of group on the intercept ($\hat{\beta} = -1.508$, $SE = 0.473$, $CI = [-2.434, -0.582]$). There were significant, negative effects of group on cubic ($\hat{\beta} = -2.010$, $SE = 0.509$, $CI = [-3.007, -1.012]$) and quintic time ($\hat{\beta} = -1.342$, $SE =$

Figure 6.4: Peak ERB$_N$-number trajectories for the pediatric cochlear implant users and their hearing-age matched peers, predicted by the fixed effects of the fitted consonant-contrast model. Data means are plotted as points.



0.412, $CI = [-2.150, -0.534]$). These effects on the odd powers of time indicated that the asymmetry between the left and right tails of the peak ERB$_N$-number trajectory was greater in the cochlear-implanted children than the normal-hearing children.

6.4 Discussion

6.4.1 The development of sibilant differentiation in pediatric CI users

The development of sibilant differentiation by pediatric bilateral cochlear implant users was investigated with respect to three variables related to their maturation and hearing experience: chronological age, hearing age, and age at first implantation. Chronological age and hearing age were similarly associated with each of the four types of individual consonant effects; however, for the consonant effects on quadratic, cubic, and quartic time, the association was marginally stronger for hearing age. This increased strength of the association suggests that the differentiation of sibilant fricatives is improved through the auditory feedback acquired through a cochlear implant.

The association between hearing age and the individual consonant effects on the powers of time was negative in each case, indicating a developmental trend toward greater differentiation of /s/ and /ʃ/ in terms of these dynamic aspects of peak ERB_N-number trajectory. However, none of these associations were statistically significant. Comparison of the CI users' τ coefficients with those of the normal-hearing children reported in §4.3.1 suggests that this absence of statistical significance is due to a lack of statistical power, a consequence of the smaller number of CI children analyzed. In particular, for the normal-hearing children, a coefficient of $\tau = -0.186$, between age and consonant effects on quadratic time, was significant at the $\alpha = 0.01$ level; whereas, for the cochlear implanted children, a coefficient of $\tau = -0.215$, between hearing age and consonant effects on quartic time, was not significant. The association between the cochlear-implanted children's hearing age and their individual consonant effects on cubic time ($\tau = -0.160$) was conspicuously larger than the analogous association for the normal-hearing children ($\tau = -0.030$, cf. §4.3.1). Potential reasons for this difference are discussed below in relation to the hearing-age comparison model.

The association between age at implant and the individual consonant effects was small in each case. Indeed, for the effects on cubic and quartic time, the association was even negative, which would suggest that younger implant ages were associated with poorer consonant contrast in terms of these spectral features. The absence of a significant effect could be related to either participants or the spoken language outcome measure studied here. The pediatric CI users analyzed here all received their implant between 9 and 34 months of age. While this range of ages is somewhat narrow, previous studies have found effects of age at implant within a comparable range (cf. Hammes et al., 2002). Earlier studies that have found an effect of age at implant, however, have typically measured such an effect on spoken language outcomes such as expressive vocabulary size, which are grosser than the size of acoustic contrast between sibilant fricatives; thus, the weak association between age at implantation and individual consonant effects may simply be due to the fact that this demographic variable does not have a strong influence on such a fine-grained outcome measure.

6.4.2 Group differences in peak-ERB$_N$ trajectory shape, but not consonant contrast

The hearing-age consonant-contrast model found significant effects of group on cubic and quintic time, which indicated that the peak ERB$_N$-number trajectories of the cochlear-implanted children's productions exhibited greater asymmetry across their tails. Inspection of the predicted peak ERB$_N$-number trajectories in Fig. 6.4 suggests that this group difference is located primarily within the first quarter of the peak ERB$_N$-number trajectories. On the one hand, in the normal-hearing children's trajectories the rise in peak ERB$_N$-number near frication onset resembles the fall in peak ERB$_N$-number near frication offset, suggesting that the linguapalatal aperture is not tightly constricted at the beginning of their productions, causing the back cavity to be partially coupled with the front. On the other hand, the rise in peak ERB$_N$-number across the first quarter of the CI children's trajectories is more gradual than its subsequent drop across the final quarter of frication. This pattern more closely resembled that of the English-speaking adults, suggesting that the cochlear-implanted children were better able to form a tight linguapalatal constriction at the onset of frication. This ability of the pediatric CI users may be due to their having more mature general motor skills, given their advanced age relative to the normal-hearing children.

Significant effects of consonant were found on the intercept and on quadratic time, suggesting that, for both groups of speakers, /s/ had a higher overall peak ERB$_N$-number and greater concavity to its peak ERB$_N$-number trajectory. None of the interactions between consonant and group were significant, although the interaction between consonant, group, and quadratic time nearly improved the model fit enough to be added during model building ($p > 0.06$). Thus, the finding, from Todd et al. (2011), that voiceless sibilants are better differentiated acoustically by normal-hearing than by cochlear-implanted children was not replicated. On this point, there are three salient innovations of the current study: a measure of peak psychoacoustic frequency (ERB_N) was used, rather than a measure of physical frequency (Hz); temporal variation in this measure was considered, rather than just its value at frication midpoint; phonemically incorrect but acoustically sibilant attempts were included, rather than just phonemically correct productions of /s/ and /ʃ/.

To investigate the marginal effect of the first of these innovations, the peak ERB_N-number data from this study were restricted to just those values computed from the ninth window (i.e. frication midpoint) of phonemically correct productions. To these data was fit a linear mixed-effects model that included fixed effects for intercept, consonant, group, and a consonant-by-group interaction. The model also included random effects on intercept by participant. Effects of consonant were evaluated with reference to /s/. The fitted model included a significant, negative effect of consonant ($\hat{\beta} = -4.985$, $SE = 0.377$, $CI = [-5.723, -4.247]$), indicating that peak ERB_N-number is lower in /ʃ/ than in /s/. Effects of group were evaluated relative to the normal-hearing children. In the fitted model, the effect of group was significant and negative ($\hat{\beta} = -2.163$, $SE = 0.527$, $CI = [-3.197, -1.130]$), indicating that the cochlear-implanted children's sibilants had lower peak ERB_N-number. The interaction between consonant and group was positive ($\hat{\beta} = 0.844$, $SE = 0.525$, $CI = [-0.185, 1.872]$), indicating that the CI children did produce less contrast between /s/ and /ʃ/; however, this interaction was not significant.

The absence of a significant interaction between consonant and group in peak ERB_N-number measured at frication midpoint of phonemically correct productions suggests that the result reported in Todd et al. (2011) may have been due to their use of an acoustic frequency measure. Given that the gammatone filter bank auditory model compresses the hertz-frequency scale, it is plausible that a significant difference in the acoustic domain disappears in the psychoacoustic domain, especially if that difference occurs in the high-frequency region, as is the case for the sibilant fricative contrast.

Chapter 7

Conclusion

7.1 Summary and implications of empirical findings

This dissertation investigated temporal variation in the spectral properties of sibilant fricatives. Toward this end, the spectral dynamics of a sibilant production were represented with a 17-point peak ERB_N-number trajectory, computed from intervals spaced evenly across the time course of the frication. Dynamic aspects of the peak ERB_N-number trajectories were modeled with fifth-order orthogonal polynomial growth curve models. Through these models, assumptions made previously in the literature about the irrelevance of the spectral dynamics of sibilant fricatives were found to be misplaced, and novel results regarding the acquisition of sibilant fricatives were found.

7.1.1 Language-specificity in sibilant productions

As was discussed in the introduction, the principal method for studying the acoustics of voiceless sibilants has been to compute a single spectral representation, e.g. from near the temporal midpoint of frication, and then to derive one (or more) summary measure(s) of the single spectral representation, e.g. the centroid or peak frequency. Thus, the prevailing methodology cannot capture any temporal variation in the spectral properties of a sibilant fricative, nor can it capture the differences in spectral dynamics that help differentiate the contrasting sibilants in a particular language or cross-linguistically. The growth curve models employed in this dissertation, however, were able to capture both aspects of spectral dynamics just mentioned because: first, time was explicitly incorporated as a factor in the model, and, second, because consonant type was allowed to interact with non-zero powers of polynomial time.

By fitting growth curve models to native adults' productions of English /s/ and /ʃ/ and Japanese /s/ and /ɕ/, in chapter 3, it was found that for none of these sibilants was peak ERB_N-number constant across its duration: For both languages, the main effects of polynomial time, up to the fifth power, were significant in the consonant contrast model. These findings question the ontology that underpins any analytical methodology that represents sibilant fricatives with only a single spectral representation. Two potential ontological commitments are considered:

1. First is the commitment that a single spectral representation is representative of the sibilant fricative from which is was computed because the spectral properties of that fricative are static. This commitment is clearly untenable, in general, since the English and Japanese sibilants have been found to be spectrally dynamic.

2. Second is the commitment that a single spectral representation simply represents the speech production "target" of the sibilant fricative. It may be possible to maintain this commitment in light of the spectral dynamical nature of sibilant fricatives; however, it is not clear what criteria should be used in determining this target in the frication.

In addition to finding evidence that the spectral properties of individual sibilants vary across the frication, the analyses of the adults' productions of English /s/ and /ʃ/ and Japanese /s/ and /ɕ/ also found that the contrast between the two sibilants within either language is indicated by the dynamic, as well as the static, properties of peak ERB_N-number trajectories. In other words, it was found not to be the case that in a given language, the peak ERB_N-number trajectories for contrastive sibilants are comparable in shape, one being merely the translation of the other along the peak ERB_N-number axis. Instead, peak ERB_N-number traced a unique trajectory for each sibilant.

While the effects of consonant on non-zero powers of polynomial time in the consonant-contrast models for the English- and Japanese-speaking adults suggested that the sibilant contrast in both languages is indicated by dynamic properties of peak ERB_N-number trajectory, it is logically possible that these results are reducible to general kinematic constraints on the articulators. Since it is well established from imaging studies that, at the temporal midpoint of frication, the articulators are positioned differently for the two sibilants, in either English or Japanese, it follows necessarily that the articulators move differently to form and release this midpoint posture. Thus, the differences in peak ERB_N-number trajectory observed between sibilants within a given language may only reflect the necessary kinematic differences in how the articulators move toward and away from distinct midpoint postures.

The results of the cross-linguistic comparison of English /s/ and Japanese /s/ make it difficult to maintain the position that differences in peak ERB_N-number trajectory-shape between consonants are due solely to general kinematic constraints on the articulators. Specifically, English /s/ and Japanese /s/ are articulated with comparable midpoint postures and were shown, in chapter 3, to have comparable peak ERB_N-number values across the middle 50% of frication. Despite this, the growth curve model comparing these two sibilants found a significant effect of language on each non-zero power of time that was included in the model. Thus, the cross-linguistic comparison of /s/ indicates language specificity in the dynamic aspects, i.e. the shape properties, of the peak ERB_N-number trajectories of sibilant fricatives.

Taken together, the consonant-contrast and cross-linguistic comparison models of the adults' productions of English and Japanese sibilants suggest that each sibilant has a unique peak ERB_N-number trajectory, and, by extension, unique spectral dynamics. The immediate question is, then: what factors determine, or at least influence, a sibilant's spectral dynamics? Here, two such factors are considered: First, peak ERB_N-number trajectory may be subject to language-specific patterns of coordination among individual articulators that support the maintenance of distinct gestures that instantiate that language's sibilant contrast. Second, peak ERB$_N$-number number trajectory may be affected by the prosodic organization of a language, which determines how a sibilant gesture is coproduced with that of a neighboring sound, e.g. a following vowel. Because both of these factors are language specific, it follows that they must be learned by children during language acquisition.

7.1.2 Development of sibilant differentiation

The development of the differentiation of contrastive gestures for English /s/ and /ʃ/, as evinced by peak ERB$_N$-number trajectories, was investigated in children with normal hearing and pediatric cochlear implant users. For these analyses, the structure of the adults' consonant-contrast model was used as a community-norm model, and the children's development was evaluated with respect to this community norm. Specifically, as concerns which aspects of peak ERB$_N$-number trajectory distinguish /s/ and /ʃ/, the children's sibilant contrast was investigated in terms of average peak ERB$_N$-number (i.e., effect of consonant on the intercept term in the model), as well as in terms of peak ERB$_N$-number trajectory-shape (i.e., the effects of consonant on the quadratic, cubic, and quartic time terms in the model); thus, the development of sibilant contrast was tracked in terms of both static and dynamic aspects of peak ERB$_N$-number trajectory.

The results for the English-acquiring children with normal hearing suggested that the contrast between /s/ and /ʃ/ developed in terms of differences both in average peak ERB$_N$-number and in the curvature of the consonants' peak ERB$_N$-number trajectories. Specifically, the two-year-olds' productions of /s/ and /ʃ/ were undifferentiated in terms of both overall peak ERB$_N$-number and trajectory curvature. As the children matured, the peak ERB$_N$-number trajectory for their productions of /ʃ/ lowered and its curvature flattened slightly, relative to the trajectory for their productions of /s/. Like their peers with normal hearing, the English-acquiring children with cochlear implants differentiated /s/ and /ʃ/ in terms of average peak ERB$_N$-number and the curvature of its trajectory across the time course of frication. Furthermore, the extent of the contrast between /s/ and /ʃ/ increased in terms of both average peak ERB_N-number and peak ERB_N-number trajectory-shape as the pediatric CI users aged and gained more experience with their implant. Thus, in both normal-hearing and cochlear-implanted children, the sibilant contrast developed in terms of both static and dynamic aspects of peak ERB_N-number trajectory.

Previous studies of the development of sibilant differentiation in children have focused on a limited number of spectral properties computed near the midpoint of frication, arguing that at that point the acoustic and spectral properties are relatively stable, the result of a maintained target articulatory posture. In this way, these previous analyses may be said to have conceptualized the sibilant contrast in terms of differences between articulatory postures. The results for the English-acquiring children presented in this dissertation, however, argue for analyzing sibilant fricatives such that they are recognized as gestures, which engender temporal variation in the spectral properties of the produced sibilant frication. While the English sibilant contrast did develop in terms of average peak ERB_N-number, it also developed in terms of peak ERB_N-number trajectory-shape, and this latter development was not necessarily synchronous with the former development. Thus, while there was evidence that children learn to differentiate English /s/ and /ʃ/ in terms of articulatory postures, they also become more fluent and adult-like in how the articulators move differently to form and release these distinctive target postures. In other words, the children seem to learn to refine articulatory gestures, rather than just learning to achieve target articulatory postures.

Further evidence for the claim that children learn gestures rather than just target postures is found in differences in the asymmetry of the left and right tails of the peak ERB_N-number trajectories. In particular, the normal-hearing, English-acquiring children's peak ERB_N-number trajectories were conspicuously more symmetric than the adults' due to there being a steeper increase in peak ERB_N across the initial quarter of either consonant. This symmetry suggested that, at the beginning of a sibilant-vowel syllable, the children have not yet formed a tight linguapalatal constriction so as to decouple the back cavity from the front. When the normal-hearing children were compared to hearing-age matched pediatric CI users, it was found that the peak ERB_N-number trajectories computed from the CI children's sibilant productions exhibited a greater amount of asymmetry than those computed from the NH children's productions. This difference was localized to the rise of peak ERB_N at the onset of frication, which was shallower in the CI children's productions, suggesting that they are better able to form a linguapalatal constriction that is narrow enough to decouple the back cavity from the onset of frication. Given that the CI children were on average a couple years older than their hearing-age matched NH peers, it is possible that this aspect of the sibilant gesture is more closely associated with chronological-age related motor development rather than with a gestural differentiation that relies on auditory feedback to guide it.

The results for the Japanese-acquiring children were complicated by an apparent developmental retrogression in the five-year-olds' differentiation of /s/ and /ɕ/ in terms of peak ERB_N-number trajectory. Between ages two and four, the children seemed to differentiate the sibilants to a greater and greater extent, but at age five, no contrast was detected. Due to the apparent developmental retrogression in the Japanese children, the sibilant contrast did not seem to develop in terms of the dynamic aspects of peak ERB_N-number trajectory, i.e. through effects on powers of polynomial time. A similar developmental retrogression was reported by Smit et al. (1990) for /s/ in English-acquiring children, which suggests that it is misplaced to assume that the acquisition of sibilant fricatives or of a sibilant contrast will always follow a monotonically increasing path from less contrast to more contrast.

7.1.3 Development of sibilant-vowel coproduction

In chapter 5, a second aspect of development was considered—the coproduction of sibilant and vowel gestures during the production of sibilant-vowel syllables. First, vowel-context effects on peak ERB_N-number trajectory were investigated in the English-speaking adults' productions of /s/ and /ʃ/ in order to determine the community-norm models. A following rounded vowel lowered peak ERB_N across the duration of the sibilant for /s/, but not for /ʃ/. A following front vowel raised peak ERB_N overall across the time course of both /s/ and /ʃ/. These effects of rounding and frontness indicated spatial coarticulatory effects arising from how articulators are positioned from the onset of the syllable. For example, the rounding effect indicated vowel-conditioned labial constriction for /s/; the absence of a rounding effect on /ʃ/ is consistent with this interpretation since /ʃ/ itself is articulated with a labial constriction. Similarly, the fronting effect indicated a more anterior constriction, from frication onset. Vowel height affected the dynamic aspects of the peak ERB_N-number trajectory of both sibilants. For /s/, the increase in peak ERB_N at frication onset was steeper before a high vowel. For /ʃ/, the rise in peak ERB_N across the time course of the sibilant was greater before a high vowel. Both of these effects seemed to indicate that the height of the following vowel modulates the coordination of the dynamic tongue and jaw gestures during the production of a sibilant.

The analysis of the children's data found that they exhibited the rounding and frontness vowel-context effects, but no significant effects of vowel height were found in the children. Additionally, the strength of the children's individual vowel-context effects was correlated with age in order to determine whether these effects increased or decreased in magnitude as the children matured and became more fluent in their speech production. The rounding and frontness effects were found to weaken as the children matured, but the height effects were found to strengthen. Thus, there is an asymmetry in how the magnitude of vowel context effects vary with age, and this asymmetry parallels the asymmetry in whether the vowel-context effect operates on static or dynamic aspects of peak ERB_N-number trajectory. In particular, rounding and frontness effects operated on the average peak ERB_N-number number, and these effects weakened with age; conversely, height effects operated on non-zero powers of polynomial time, and these effects strengthened with age.

A prevailing interpretation of vowel-context effects, advanced in Nittrouer et al. (1989);

Nittrouer (1995); Nittrouer et al. (1996), is that they indicate the size of the organizational unit used in speech motor planning. It is argued that children's speech is organized relatively more syllabically than adults' speech; thus, it is predicted that children would exhibit stronger vowel context effects than adults. In their studies, Nittrouer and her colleagues looked primarily at rounding effects on the spectral properties of sibilant fricatives at only a couple locations in the frication, e.g. the "steady-state" portion of frication and near the sibilant-vowel boundary. Moreover, vowel-context effects were analyzed for each of these locations independently of the other; hence, context effects on spectral dynamics were not considered.

In this dissertation, it was found that vowel height significantly affects the peak ERB_N-number trajectory shape in the adults' productions of English sibilants, and that English-acquiring children develop toward the adult-like pattern of vowel-height coarticulation, i.e. the strength of vowel-height effects increased with age in children. Thus, the investigation of vowel-context effects on the spectral dynamics of sibilant fricatives seems a legitimate enterprise. Moreover, when the developmental pattern of vowel-height effects in English-acquiring children challenges the interpretation of vowel-context effects, generally, as indices of the size of the organizational unit of speech motor planning. In particular, these effects strengthened with age, which would suggest that the organizational unit increases in size as children mature. An alternative interpretation, which is better supported by the data presented in this dissertation, is that vowel-context effects indicate only the maturation of the ability to coordinate gestures that involve different articulators.

7.2 Limitations of the study

While the studies presented in this dissertation offered strong evidence that the temporal variation in the spectral properties of sibilant fricatives are themselves indicators of the sibilant contrast within a language, and that these patterns of variation must be learned during language acquisition, the general methodology used to investigate these issues is not without limitations. Two of these limitations are discussed here.

7.2.1 Prosodic position of the elicited sibilant fricatives

First, sibilant fricatives were elicited only in word- and phrase-initial pre-vocalic position. Consequently, the results presented herein should be strictly circumscribed to only this prosodic position. The reason for this circumscription is that peak ERB_N is sensitive to the amplitude of the frication noise. Thus, the increase in the amplitude of the frication noise at the onset of the elicited productions may have contributed to the increase in peak ERB_N. Furthermore, the amplitude envelope is likely to differ according to whether the sibilant is produced word-initially or word-medially. Specifically, the latter position would likely exhibit a more uniform amplitude envelope. Moreover, medially produced sibilants may also be subject to perseveratory coarticulation with the preceding sound. The effect of prosodic position on the spectral dynamics of sibilant fricatives is thus an open empirical question; hence, for the moment, the present results should not be overgeneralized.

7.2.2 Subdivision of a sibilant into closure, maintenance, and opening movements

In the analyses of peak ERB_N-number trajectories, temporal variation was assessed by fitting a polynomial growth curve model to the full duration of the frication. However, in the introduction chapter, it was stated that the choice of some previous analyses to analyze the middle portion of the frication was motivated, not by an assumption that the spectral properties were reasonably constant across the full duration of the sibilant, but rather by an assumption of constancy across just the middle. This raises the possibility that temporal variation in a sibilant's spectral properties either could or should be examined across three intervals independently: the closure interval, during which the articulators move to form the linguapalatal constriction; the maintenance interval, during which the articulators remain relatively stable to maintain the degree and location of the constriction; and the opening interval, during which the articulators move to release the constriction. Unfortunately, none of these intervals are apparent in the acoustic waveform, but could be determined with appropriate articulatory and acoustic data that are temporally synchronized (cf. Mooshammer et al., 2006). Thus, this possibility of a spectrally stable middle portion would have to be investigated with qualitatively different data than were analyzed in this dissertation.

References

Akamatsu, T. (1997). *Japanese phonetics: Theory and practice*, volume 3. Lincom Europa, Newcastle, UK.

Apoux, F. and Healy, E. W. (2009). On the number of auditory filter outputs need to understand speech: Further evidence for auditory channel independence. *Hearing Research*, 255:99–108.

Atal, B. S. and Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50(2B):637–655.

Bailley, G., Laboissier, R., and Schwartz, J.-L. (1991). Formant trajectories as audible gestures. *Journal of Phonetics*, 19(1):9–23.

Bates, D., Mächoler, M., Bolker, B. M., and Walker, S. C. (2014). Fitting linear mixed-effects models using lme4. Submitted to *Journal of Statistical Software*, arXiv preprint arXiv:1406.5823.

Beerends, R. J., ted Morsche, H. G., van den Berg, J. C., and van de Vrie, E. M. (2003). *Fourier and Laplace Transforms*. Cambridge University Press, Cambridge, UK.

Behrens, S. J. and Blumstein, S. E. (1988). Acoustic characteristics of english voiceless fricatives: A descriptive analysis. *Journal of Phonetics*, 16:295–298.

Bierer, J. A. (2002). Auditory cortical images of cochlear-implant stimuli: Dependence on electrode configuration. *Journal of Neurophysiology*, 87:478–492.

Blacklock, O. S. (2004). *Characteristics of variation in production of normal and disordered fricatives, using reduced-variance spectral methods*. PhD thesis, University of Southampton.

Blamey, P. J., Barry, J. G., and Jacq, P. (2001). Phonetic inventory development in young cochlear implant users 6 years post-operation. *Journal of Speech, Language, and Hearing Research*, 44:73–79.

Browman, C. P. and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, 6:201–251.

Brownell, R. (2000). *Receptive One-Word Picture Vocabulary Test, Second Edition.* Academic Therapy Publication, Inc.

Brunner, J., Ghosh, S., Hoole, P., Matthies, M., Tiede, M., and Perkell, J. (2011). The influence of auditory acuity on acoustic variability and the use of a motor equivalence during adaptation to a perturbation. *Journal of Speech, Language, and Hearing Research*, 54:727–739.

Chin, S. B. (2003). Children's consonant inventories after extended cochlear implant use. *Journal of Speech, Language, and Hearing Research*, 46:849–862.

Chin, S. B., Tsai, P. L., and Gao, S. (2003). Connected speech intelligibility of children with cochlear implants and children with normal hearing. *American Journal of Speech-Language Pathology*, 12:440–451.

Cooke, M. P. (2003). Glimpsing speech. *Journal of Phonetics*, 31:579–584.

Cooke, M. P. (2006). A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America*, 119(3):1562–1573.

Edwards, J. R. and Beckman, M. E. (2008a). Methodological questions in studying consonant acquisition. *Clinical Linguistics & Phonetics*, 22(12):937–956.

Edwards, J. R. and Beckman, M. E. (2008b). Some cross-linguistic evidence for modulation of implicational universals by language-specific frequency effects in phonological development. *Language Learning and Development*, 4(2):122–156.

Fant, G. (1960). *Acoustic Theory of Speech Production.* Mouton, The Hague, Netherlands.

Fletcher, H. (1940). Auditory patterns. *Reviews of Modern Physics*, 12:47–65.

Fletcher, S. G. and Newman, D. G. (1991). [s] and [ʃ] as a function of linguapalatal contact place and sibilant groove width. *Journal of the Acoustical Society of America*, 89(2):850–858.

Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America*, 84(1):115–124.

Fowler, C. A. and Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech*, 36(2,3):171–195.

Fox, R. A. and Nissen, S. L. (2005). Sex-related acoustic changes in voiceless English fricatives. *Journal of Speech, Language, and Hearing Research*, 48(4):753–765.

Ghosh, S. S., Matthies, M. L., Maas, E., Hanson, A., Tiede, M., Ménard, L., Guenther, F. H., Lane, H., and Perkell, J. S. (2010). An investigation of the relation between sibilant production and somatosensory and auditory acuity. *Journal of the Acoustical Society of America*, 128(5):3079–3087.

Gibbon, F. (1999). Undifferentiated lingual gestures in children with articulation/phonological disorders. *Journal of Speech, Language, and Hearing Research*, 42:382–397.

Glasberg, B. R. and Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47:103–138.

Goffman, L., Smith, A., Heisler, L., and Ho, M. (2008). The breadth of coarticulatory units in children and adults. *Journal of Speech, Language, and Hearing Research*, 51:1424–1437.

Goldman, R. and Fristoe, M. (2000). *Goldman-Fristoe Test of Articulation, Second Edition*. Pearson, San Antonio, TX.

Greenwood, D. D. (1961). Critical bandwidth and the frequency coordinates of the basilar membrane. *Journal of the Acoustical Society of America*, 33(10):1344–1356.

Greenwood, D. D. (1990). A cochlear frequency-position function for several species—29 years later. *Journal of the Acoustical Society of America*, 87(6):2592–2605.

Gundersen, T., Skarstein, O., and Sikkeland, T. (1978). A study of the vibration of the basilar membrane in human temporal bone preparations by the use of the Mössbauer effect. *Acta Oto-laryngologica*, 86(3–4):225–232.

Haley, K. L., Seelinger, E., Mandulak, K. C., and Zajac, D. J. (2010). Evaluating the spectral distinction between sibilant fricatives through a speaker-centered approach. *Journal of Phonetics*, 38:548–554.

Hammes, D. M., Novak, M. A., Rotz, L. A., Willis, M., Edmondson, D. M., and Thomas, J. F. (2002). Early identification and cochlear implantation: Critical factors for spoken language development. *Annals of Otology, Rhinology, & Laryngology*, 111(5.2):74–78.

Hazan, V. and Baker, R. (2011). Is consonant perception linked to within-category dispersion or across-category distance? In *Proceedings of the International Congress of Phonetic Sciences (ICPhS) XVII*, pages 839–842.

Heinz, J. M. and Stevens, K. N. (1961). On the properties of voiceless fricatives. *Journal of the Acoustical Society of America*, 33(5):589–596.

Hughes, G. W. and Halle, M. (1956). Spectral properties of fricative consonants. *Journal of the Acoustical Society of America*, 28(2):303–310.

Iskarous, K., Shadle, C. H., and Proctor, M. I. (2011). Articulatory–acoustic kinematics: The production of American English /s/. *The Journal of the Acoustical Society of America*, 129(2):944–954.

Jongman, A., Wayland, R., and Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, 108(3):1252–1263.

Jordan, M. I. (1990). Motor learning and the degrees of freedom problem. In Jeannerod, M., editor, *Attention and Performance XIII*. Erlbaum, Hillsdale, NJ.

Katz, W. F. and Bharadwaj, S. (2001). Coarticulation in fricative-vowel syllables produced by children and adults: A preliminary report. *Clinical Linguistics & Phonetics*, 15(1 & 2):139–143.

Koenig, L. L., Shadle, C. H., Preston, J. L., and Mooshammer, C. R. (2013). Toward improved spectral measures of /s/: Results from adolescents. *Journal of Speech, Language, and Hearing Research*, 56(4):1175–1189.

Kral, A., Hartmann, R., Mortazavi, D., and Klinke, R. (1998). Spatial resolution of cochlear implants: The electrical field and excitation of auditory afferents. *Hearing Research*, 121:11–28.

Landau, H. J. and Pollak, H. O. (1961). Prolate spheroidal wave functions, Fourier analysis, and uncertainty—II. *Bell System Technical Journal*, 40:65–84.

Landau, H. J. and Pollak, H. O. (1962). Prolate spheroidal wave functions, Fourier analysis, and uncertainty—III. *Bell System Technical Journal*, 41:1295–1336.

Li, F. (2012). Language-specific developmental differences in speech production: A cross-language acoustic study. *Child Development*, 83(4):1303–1315.

Li, F., Edwards, J., and Beckman, M. E. (2009). Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in English and Japanese toddlers. *Journal of Phonetics*, 37:111–124.

Loizou, P. C. (2006). Speech processing in vocoder-centric cochlear implants. *Advances in Oto-Rhino-Laryngology*, 64:109–143.

Lyon, R. F., Katsiamis, A. G., and Drakakis, E. M. (2010). History and future of auditory filter models. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 3809–3812.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580.

Maniwa, K., Jongman, A., and Wade, T. (2009). Acoustic characteristics of clearly spoken english fricatives. *Journal of the Acoustical Society of America*, 125(6):3962–3973.

Markel, J. D. and Gray, A. H. (1976). *Linear Prediction of Speech*. Springer-Verlag, Berlin, Germany.

McCune, L. and Vihman, M. M. (1987). Vocal motor schemes. *Papers and Reports on Child Language Development*, 26:72–79.

McGowan, R. S. and Nittrouer, S. (1988). Differences in fricative production between children and adults: Evidence from an acoustic analysis of /ʃ/ and /s. *Journal of the Acoustical Society of America*, 83(1):229–236.

McLeod, S., Roberts, A., and Sita, J. (2006). Tongue/palate contact for the production of /s/ and /z/.

McMurray, B. and Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, 118(2):219–246.

Middlebrooks, J. C., Bierer, J. A., and Snyder, R. L. (2005). Cochlear implants: The view from the brain. *Current Opinion in Neurobiology*, 15:488–493.

Mirman, D. (2014). *Growth Curve Analysis and Visualization Using R.* CRC Press, New York, NY.

Mirman, D., Dixon, J. A., and Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, 59(4):475–494.

Moore, B. C. J. (1997). *An Introduction to the Psychology of Hearing.* Academic Press, London, UK, Fourth edition.

Moore, B. C. J. and Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74(3):750–753.

Moore, B. C. J. and Glasberg, B. R. (1996). A revision of Zwicker's loudness model. *Acta Acustica united with Acustica*, 82(2):335–345.

Moore, B. C. J., Glasberg, B. R., and Baer, T. (1997). A model for the prediction of thresholds, loudness and partial loudness. *Journal of the Audio Engineering Society*, 45(4):224–240.

Mooshammer, C., Hoole, P., and Geumann, A. (2006). Interarticulator cohesion within coronal consonant production. *Journal of the Acoustical Society of America*, 120(2):1028–1039.

Munson, B. (2001). A method for studying variability in fricatives using dynamic measures of spectral mean. *Journal of the Acoustical Society of America*, 110(2):1203–1206.

Munson, B. (2004). Variability in /s/ production in children and adults: Evidence from dynamic measures of spectral mean. *Journal of Speech, Language, and Hearing Research*, 47:58–69.

Nakamura, M., Nozaki, K., Takimoto, H., Nagamune, K., Fujigaki, M., and Wada, S. (2011). Simultaneous measurements of aeroacoustic sounds and wall vibration for exploring the contribution of tooth vibration in the production of sibilant sounds /s/. *Journal of Biomedical Science and Engineering*, 4:83–89.

Narayanan, S. S. and Alwan, A. A. (2000). Noise source models for fricative consonants. *IEEE Transactions on Speech and Audio Processing*, 8(2):328–344.

Narayanan, S. S., Alwan, A. A., and Haker, K. (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *Journal of the Acoustical Society of America*, 93(3):1325–1347.

Newman, R. S. (2003). Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report. *Journal of the Acoustical Society of America*, 113(5):2850–2860.

Newman, R. S., Clouse, S. A., and Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *Journal of the Acoustical Society of America*, 109(3):1181–1196.

Nicholas, J. G. and Geers, A. E. (2007). Will they catch up? The role of age at cochlear implantation in the spoken language development of children with severe to profound hearing loss. *Journal of Speech, Language, and Hearing Research*, 50:1048–1062.

Nissen, S. L. and Fox, R. A. (2005). Acoustic and spectral characteristics of young children's fricative productions: A developmental perspective. *Journal of the Acoustical Society of America*, 118(4):2570–2578.

Nittrouer, S. (1995). Children learn separate aspects of speech production at different rates: Evidence from spectral moments. *Journal of the Acoustical Society of America*, 91(1):520–530.

Nittrouer, S. and Studdert-Kennedy, M. (1987). The role of coarticulatory effects in the perception of fricatives by children and adults. *Journal of Speech and Hearing Research*, 30:319–329.

Nittrouer, S., Studdert-Kennedy, M., and McGowan, R. S. (1989). The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech and Hearing Research*, 32:120–132.

Nittrouer, S., Studdert-Kennedy, M., and Neely, S. T. (1996). How children learn to organize their speech gestures: Further evidence from fricative-vowel syllables. *Journal of Speech, Language, and Hearing Research*, 39:379–389.

Nozaki, K. (2010). Numerical simulation of sibilant [s] using the real geometry of a human vocal tract. In Resch, M., Benkert, K., Wang, X., Galle, M., Bez, W., Kobayashi, H., and Roller, S., editors, *High Performance Computing on Vector Systems 2010*, pages 137–148. Springer-Verlag, Berlin, Germany.

Patterson, R. D. (1976). Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society of America*, 59(3):640–654.

Peng, S. C., Spencer, L. J., and Tomblin, J. B. (2004). Speech intelligibility of pediatric cochlear implant recipients with 7 years of device experience. *Journal of Speech, Language, and Hearing Research*, 47:1227–1236.

Percival, D. B. and Walden, A. T. (1993). *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques.* Cambridge University Press, Cambridge, UK.

Perkell, J. S., Boyce, S. E., and Stevens, K. N. (1979). Articulatory and acoustic correlates of the /s/–/ʃ/ distinction. In Wolf, J. J. and Klatt, D. H., editors, *Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America*. American Institute of Physics, New York, NY.

Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., Stockmann, E., and Guenther, F. H. (2004). The distinctness of speakers' /s/–/ʃ/ contrast is related to their auditory discrimination and use of an articulatory saturation effect. *Journal of Speech, Language and Hearing Research*, 47(6):1259–1269.

Plummer, A. R. (2014). *The Acquisition of Vowel Normalization during Early Infancy: Theory and Computational Model*. PhD thesis, The Ohio State University.

Raggio, M. W. and Schreiner, C. E. (2003). Neuronal responses in cat primary auditory cortex to electrical cochlear stimulation: IV. Activation pattern for sinusoidal stimulation. *Journal of Neurophysiology*, 89:3190–3204.

Romeo, R., Hazan, V., and Pettinato, M. (2013). Developmental and gender-related trends of intra-talker variability in consonant production. *Journal of the Acoustical Society of America*, 134(5):3781–3792.

Saltzman, E. and Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1:333–382.

Schreiner, C. E., Read, H. L., and Sutter, M. L. (2000). Modular organization of frequency integration in primary auditory cortex. *Annual Review of Neuroscience*, 23:501–529.

Serry, T. A. and Blamey, P. J. (1999). A 4-year investigation into phonetic inventory development in young cochlear implant users. *Journal of Speech, Language, and Hearing Research*, 42:141–154.

Shadle, C. H. (1985). *The Acoustics of Fricative Consonants.* Technical Report 506, MIT Research Laboratory of Electronics, Cambridge, MA.

Shadle, C. H. (2006). Acoustic phonetics. In Brown, K., editor, *Encyclopedia of Language & Linguistics*, pages 442–460. Elsevier, Oxford, UK.

Shumway, R. H. and Stoffer, D. S. (2006). *Time Series Analysis and its Applications.* New York, NY, second edition.

Simpson, A. J. and Fitter, M. J. (1973). What is the best index of detectability? *Psychological Bulletin*, 80:481–488.

Slepian, D. (1964). Prolate spheroidal wave functions, Fourier analysis, and uncertainty—IV: Extensions to many dimensions; generalized prolate spheroidal functions. *Bell System Technical Journal*, 43:3009–3058.

Slepian, D. and Pollak, H. O. (1961). Prolate spheroidal wave functions, Fourier analysis, and uncertainty—I. *Bell System Technical Journal*, 40:43–64.

Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., and Bird, A. (1990). The Iowa Articulation Norms Project and its Nebraska replication. *Journal of Speech and Hearing Disorders*, 55:779–798.

Smith, A. and Goffman, L. (1998). Stability and patterning of speech movement sequences in children and adults. *Journal of Speech, Language, and Hearing Research*, 41:18–30.

Soli, S. D. (1981). Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation. *Journal of the Acoustical Society of America*, 70(4):976–984.

Stevens, K. N. (1971). Airflow and turbulence noise for fricative and stop consonants: Static considerations. *Journal of the Acoustical Society of America*, 50(4B):1180–1192.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory–acoustic data. In David, E. E. and Denes, P. B., editors, *Human Communication: A Unified View*, pages 51–66. McGraw-Hill.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*.

Stevens, K. N. (2000). *Acoustic Phonetics*. The MIT Press, Cambridge, MA.

Stone, M., Faber, A., Raphael, L. J., and Shawker, T. H. (1992). Cross-sectional tongue shape and linguapalatal contact patterns in [s], [ʃ] and [l]. *Journal of Phonetics*, 20(2):253–270.

Stone, M. and Lundberg, A. (1996). Three-dimensional tongue surface shapes of English consonants and vowels. *Journal of the Acoustical Society of America*, 99(6):3728–3737.

Thomson, D. J. (1982). Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70:1055–1096.

Toda, M. and Honda, K. (2003). An MRI-based cross-linguistic study of sibilant fricatives. In *Proceedings of the 6th International Seminar on Speech Production*.

Toda, M. and Maeda, S. (2006). Quantal aspects of non-anterior sibilant fricatives: A simulation study. In *Proceedings of the 7th International Seminar on Speech Production*, pages 573–580.

Toda, M., Maeda, S., Carlen, A. J., and Meftahi, L. (2002). Lip gestures in English sibilants: Articulatory–acoustic relationship. In *Proceedings of the 7th International Conference on Spoken Language Processing*, pages 2165–2168.

Toda, M., Maeda, S., and Honda, K. (2010). Formant-cavity affiliation in sibilant fricatives. In Fuchs, S., Toda, M., and Żygis, M., editors, *Turbulent Sounds: An Interdisciplinary Guide*, pages 343–374. De Gruyter Mouton, Berlin, Germany.

Todd, A. E., Edwards, J. R., and Litovsky, R. Y. (2011). Production of contrast between sibilant fricatives by children with cochlear implants. *Journal of the Acoustical Society of America*, 130(6):3969–3979.

Tsurutani, C. (2004). Acquisition of Yo-on (Japanese contracted sounds) in L1 and L2 phonology. *Second Language*, 3:27–47.

Uchanski, R. M. and Geers, A. E. (2003). Acoustic characteristics of the speech of young cochlear implant users: A comparison with normal-hearing age-mates. *Ear & Hearing*, 24(1S):90S–105S.

Umebayashi, N. and Takagi, S. (1965). Gakureimae no kodomo no koonnoryoku ni kansuru ichikenkyu. *Onseigengo Igaku*, 6:17–18.

Viemeister, N. F. and Wakefield, G. H. (1991). Temporal integration and multiple looks. *Journal of the Acoustical Society of America*, 90(2):858–865.

Vihman, M. M. and Keren-Portnoy, T. (2011). The role of production practice in lexical and phonological development. *Journal of Child Language*, 38(1):41–45.

von Békésy, G. (1947). The variation of phase along the basilar membrane with sinusoidal vibrations. *Journal of the Acoustical Society of America*, 19(3):452–460.

von Békésy, G. (1960). *Experiments in Hearing.* McGraw-Hill, New York, NY.

Weismer, G., Elbert, M., and Whiteside, J. (1980). [s] spectra in the speech of normally-articulating preschool children and adults. *Journal of the Acoustical Society of America*, 68(S1):S114.

Welch, P. D. (1967). The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73.

Williams, K. T. (1997). *Expressive Vocabulary Test.* American Guidance Service, Circle Pines, MN.

Yasuda, A. (1970). Articulatory skills in three-year-old children. *Studia Phonologica*, 5:52–71.

Zharkova, N., Hewlett, N., and Hardcastle, W. J. (2011). Coarticulation as an indicator of speech motor control development in children: An ultrasound study. *Motor Control*, 15:118–140.

Zharkova, N., Hewlett, N., and Hardcastle, W. J. (2012). An ultrasound study of lingual coarticulation in /sv/ syllables produced by adults and typically developing children. *Journal of the International Phonetic Association*, 42(2):193–208.

Zharkova, N., Hewlett, N., Hardcastle, W. J., and Lickley, R. J. (2014). Spatial and temporal lingual coarticulation and motor control in preadolescents. *Journal of Speech, Language, and Hearing Research*, 57:374–388.