SPEECH PERCEPTION OF CANADIAN ENGLISH SIBILANTS:
PROCESSING OF ACOUSTIC INFORMATION OR
UNDERLYING (ARTICULATORY) VOCAL TRACT CONFIGURATIONS?

SPEECH PERCEPTION OF CANADIAN ENGLISH SIBILANTS: PROCESSING OF ACOUSTIC INFORMATION OR UNDERLYING (ARTICULATORY) VOCAL TRACT CONFIGURATIONS?

By SAN-HEI KENNY LUK, B.E.L.S.(Hons.), M.A.

A Thesis Submitted to the School of Graduate Studies in
Partial Fulfillment of the Requirements for the Degree Master of Science

McMaster University
MASTER OF SCIENCE (2018)
Hamilton, Ontario (Cognitive Science of Language)

TITLE: Speech Perception of Canadian English Sibilants: Processing of Acoustic Information or Underlying (Articulatory) Vocal Tract Configurations?

AUTHOR: San-Hei Kenny Luk, B.E.L.S.(Hons.), M.A. (McMaster University)

SUPERVISORS: Dr. Daniel Pape, Dr. Elisabet Service

NUMBER OF PAGE: x, 68

Lay Abstract

Articulatory speech perception theories hypothesize that listeners would recover articulatory information when mapping acoustic signals to phonetic categories. To test this hypothesis, we manipulated the Canadian English sibilants /s/ and /ʃ/ such that the manipulated sibilants were articulatorily cued as the original sibilants but acoustically cued as the alternative sibilants (/s/ as /ʃ/ and /ʃ/ as /s/). In an identification task, our Canadian English listeners showed a categorical switch in the manipulated /ʃ/ stimuli but not the manipulated /s/ stimuli. This asymmetry between two sibilants can be explained by *an acoustic plus articulatory account* and *a purely acoustic account*, and we argue that a purely acoustic account is a more plausible. While our results cannot be used to reject the possibility of articulatory information recovery, acoustic information is showed to be more dominant than articulatory information during identification even if we assume that articulatory information is recovered.

Abstract

Acoustic and articulatory speech perception theories are proposed to explain how listeners map acoustic signals to phonetic categories. Different from acoustic theories, articulatory theories hypothesize that listeners would recover articulatory information during the mapping. To test this hypothesis, we altered the acoustic information of the Canadian English sibilants /s/ and /ʃ/ while keeping their articulatory information to signal places of articulation of the original sibilants. The manipulated sibilants were *articulatorily* cued as the original sibilants, but *acoustically* cued as the alternative sibilants (/s/ as /ʃ/ and /ʃ/ as /s/). We conducted an identification task to examine whether altering acoustic information would switch our Canadian English listeners' identification. The listeners identified acoustically /s/-like /ʃ/ completely as the alternative sibilant /s/, but the acoustically /ʃ/-like /s/ as 60% the alternative sibilant /ʃ/ and 40% the original sibilant /s/. There was a categorical switch in the /ʃ/ stimuli but not the /s/ stimuli. This asymmetry of identification between two sibilants can be explained by two accounts: *an acoustic plus articulatory account* would be that the listeners relied more articulatory information only when identifying /s/ but not /ʃ/; and *a purely acoustic account* would be that the asymmetry was only a result of still existing small acoustic differences. While the acoustic plus articulatory account cannot explain why articulatory information only influenced the /s/ identification of the /s/ stimuli even after adding a set of assumptions, the purely acoustic account allows us to explain our results consistently without additional assumptions. Although our results cannot be used as evidence to reject the possibility that listeners will recover articulatory information, the results do suggest that even if we assume that articulatory information is recovered, acoustic information is more dominant than articulatory information in the identification process, at least for Canadian English /s/ and /ʃ/.

Acknowledgements

I would like to thank my supervisors, Dr. Daniel Pape and Dr. Elisabet Service, for their helpful guidance throughout this project and their valuable comments on my thesis drafts.

I would also like to thank Sydney Crowdis and Olivier Mercier from the Phonetics Lab for their help in the experiment and Fareeha Rana for her help in statistics.

Lastly, I would like to thank my mum, dad and sister, my aunts, and the two Jerry's for their unconditional support and encouragement. I would like to dedicate this thesis to my grandma.

Contents

Lay Abstract ... iii

Abstract ... iv

Acknowledgements ... vi

1. Introduction ... 1

1.1. Speech perception theories ... 1

1.2. Articulation and acoustics of sibilants ... 6

1.3. Perception of sibilants ... 12

2. Methods ... 17

2.1. Participants ... 17

2.2. Stimuli ... 17

2.3. Acoustic properties of the stimuli ... 26

2.4. Experimental procedure ... 28

3. Results ... 30

3.1. Identification data ... 30

3.1.1. Isolated fricative noises ... 31

3.1.2. Long vs. short fricative noises ... 33

3.1.3. Original vs. alternative sibilant vowel contexts ... 34

3.2. Confidence data ... 37

3.2.1. Isolated fricative noises ... 37

3.2.2. Long vs. short fricative noises ... 40

3.2.3. Original vs. alternative sibilant vowel contexts ... 42

3.3. Reaction time data ... 43

3.3.1. Long vs. short fricative noises ... 44

3.3.2. Original vs. alternative sibilant vowel contexts ... 44

4. Discussion ... 46

4.1. Acoustic plus articulatory information vs. purely acoustic information....... 46

4.1.1. An acoustic plus articulatory account 48

4.1.2. A purely acoustic account 52

4.1.3. An artifact account 57

4.2. Effect of vowel context 59

4.3. Length of fricative noise 61

4.4. Speech vs. non-speech modes of perception.. 61

4.5. Future studies 62

4.6. Conclusion 63

References 66

List of Tables

Table 2.1. Summary of manipulation in stimulus steps. 20

Table 2.2. Spectral peak locations and spectral moments (mean, variance, skewness, kurtosis) of fricative noise in stimulus steps 27

Table 2.3. F2 onsets of vowel contexts 28

List of Figures

Figure 1.1. Schematic spectra of /s/ and /ʃ/ in Jongman et al. (2000). 9

Figure 1.2. Schematic spectra of normal and louder /ʃ/.. 10

Figure 2.1. Relevant and irrelevant frequencies of natural /s/ and /ʃ/.............................. 19

Figure 2.2. Manipulation of /ʃ/ in step 6.. 21

Figure 2.3. Spectra at 7 stimulus steps. 23

Figure 3.1. Proportions of /s/ responses to isolated fricative noise stimuli averaged across long and short fricative noise contexts.. 31

Figure 3.2. Proportions of /s/ responses to isolated long and short fricative noise stimuli. 34

Figure 3.3. Proportions of /s/ responses to fricatives in original and alternative sibilant vowel contexts. 35

Figure 3.4. Mean confidence scores average across isolated long and short fricative noise stimuli. 38

Figure 3.5. Means of confidence scores for long and short isolated fricative noise stimuli. 41

Figure 3.6. Mean confidence scores for fricatives in original and alternative sibilant vowel contexts. 42

Figure 3.7. Mean reaction times (in ms) for long and short fricative noise stimuli. 44

Figure 3.8. Mean reaction times (in ms) for fricatives in original and alternative sibilant vowel contexts. 45

Figure 4.1. Spectra of natural and manipulated stimuli with amplified irrelevant frequencies (step 7).. 53

Figure 4.2. Frequencies additionally amplified in the frequency range above 5 kHz...... 58

1. Introduction

To comprehend speech, listeners are required to perform a series of tasks. One of the initial tasks is that listeners have to map the acoustic signals they hear onto the phonetic or phonological categories that are assumed to exist in the listeners’ minds so that they can identify the phonetic or phonological categories. Speech perception theories have been proposed to explain how listeners accomplish this mapping, but these theories disagree on whether the information listeners employ is purely acoustic or ultimately articulatory in nature. This study examines if articulatory information is available to help listeners identify phonetic categories by looking at how listeners identify a subcategory of fricatives called sibilants. This chapter begins with (1.1) an overview of speech perception theories, and then discusses (1.2) articulation and acoustics of sibilants and (1.3) perception of sibilants.

1.1. Speech perception theories

Different speech perception theories have been proposed to explain how listeners map physical acoustic signals to the phonetic categories in their minds. There are two broad groups of theories: acoustic theories and articulatory theories. Both groups of theories have attempted to explain how listeners can still successfully perform the mapping even though speakers produce phonetic categories as physical tokens with tremendous variability originating from various sources like speaker variability (Peterson & Barney, 1952) and coarticulation (Magen, 1997). These theories agree that phonetic categories should comprise invariant representations of some kind to allow the mapping to happen. That is, listeners should have invariant mental representations for each phonetic category to which they can compare acoustic signals, and it is such invariant reference that makes mapping acoustic signals with variability onto corresponding phonetic categories possible. The fundamental difference between the two groups of theories lies in how they define the nature of invariant mental representations. They disagree on which type of information is primarily stored as the invariant mental representations and processed during the mapping. Acoustic theories propose that mental representations of phonetic categories are primarily acoustic, whereas articulatory theories argue that mental representations are ultimately articulatory, i.e. they assume that vocal tract targets and laryngeal settings are recovered from the perceived acoustic signals. The major theories from each group are summarized below (for an overview, see Perrier, 2005).

Two major acoustic theories are the *Acoustic Invariance Theory* (Blumstein, 1986; Blumstein & Stevens, 1979; Stevens & Blumstein, 1978) and the *Adaptive Variability Theory* (Lindblom, 1988, 1990). The *Acoustic Invariance Theory* (Blumstein, 1986; Blumstein & Stevens, 1979; Stevens & Blumstein, 1978) proposes that listeners can extract acoustic patterns representing different phonetic features from speech signals. The acoustic pattern which characterizes a specific phonetic feature is always consistent across different phonetic environments and different speakers. For example, Blumstein and Stevens (1979) analyzed the frequency spectra of labial, alveolar and velar stops produced in different vowel contexts and by different speakers. For each type of stops produced at the same place of articulation, they observed a consistent pattern of how amplitude changed over different frequencies. These consistent acoustic patterns are examples of the invariant acoustic information which physically exists in speech signals, and such acoustic information are proposed to enable listeners to identify phonetic categories. Similarly, the *Adaptive Variability Theory* (Lindblom, 1988, 1990) proposes that the information that listeners primarily rely on is also acoustic in nature. However, physical tokens of a phonetic category produced in acoustic signals are considered as inherently variable because they are produced in different contexts. Acoustic variability in physical tokens can be caused by a variety of contextual factors, including speaker variability, phonetic environment, speaking style and speaking tempo. When processing physical tokens in acoustic signals, listeners are able to interpret acoustic signals according to the context by determining how contextual factors influence the acoustic properties of a physical token. With this perceptual recovery from the context, physical tokens become invariant mental representations that are identified as their corresponding phonetic categories. Although acoustic signals are not physically invariant and require perceptual recovery, the information listeners try to interpret is still acoustic in nature. The two acoustic theories agree that the information primarily processed for identifying phonetic categories is acoustic, but they differ in where the invariant mental representations are to be found. Invariant acoustic patterns are directly derived from acoustic signals in the *Acoustic Invariance Theory*, whereas acoustic signals require perceptual recovery in the *Adaptive Variability Theory*.

Taking a different perspective, articulatory theories relate speech perception closely to speech production by suggesting that listeners perceive speech based on how it is produced. They propose that listeners can recover invariant articulatory information from acoustic signals with variability, and invariant mental representations are ultimately articulatory in nature. Two major articulatory theories are the *Motor Theory* (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985) and the *Direct-Realist Theory* (Fowler, 1986). According to the *Motor Theory* (Liberman et al., 1967; Liberman & Mattingly, 1985), listeners are able to recover what speakers intend to produce from the acoustic signals. More specifically, acoustic signals provide listeners with cues to recover the underlying motor commands that speakers need in order to move the corresponding articulators. These motor commands are invariant in nature and are performed to produce phonetic gestures that represent specific phonetic categories. Since continuous speech is produced from a variety of overlapping motor commands, the physical output in the form of acoustic signals is variable. Listeners, who know how motor commands overlap with each other, can still recover a set of underlying motor commands. The *Direct-Realist Theory* sees perception in general as a process to "recover events in the real world" (Fowler, 1986, p.3). In the case of speech perception, the events derive from how speakers configure their vocal tracts. The outputs of vocal tract configurations are acoustic signals. However, acoustic signals are merely a medium that carries information listeners use to recover the underlying vocal tract configurations. These vocal tract configurations that physically exist in the real world are invariant, and they are used to map onto phonetic categories. The two articulatory theories have in common the assumption that mental representations of phonetic categories are articulatory in nature. That is, acoustic signals only serve as a medium that delivers cues for recovering articulatory information, and it is this articulatory information that is ultimately used for identifying phonetic categories. However, the two articulatory theories differ in what articulatory information is actually recovered, i.e. motor commands in the *Motor Theory* and vocal tract configurations in the *Direct-Realist Theory*.

In summary, the major difference between acoustic and articulatory theories lies in whether mental representations of phonetic categories are acoustic or articulatory in nature. The nature of mental representations determines which type of information is used to map acoustic signals onto phonetic categories. For acoustic theories, acoustic information itself is sufficient for successful mapping. For articulatory theories, there is an intermediate process which recovers articulatory information from acoustic information. Acoustic signals are simply a medium which carries the necessary information listeners employ to recover articulatory information. After this process of recovery, listeners ultimately rely on articulatory information to identify phonetic categories.

There has been an ongoing debate on whether articulatory information is, indeed, recovered in the process of identifying phonetic categories. For example, Ohala (1996) tried to demonstrate that recovering articulatory information is unnecessary by investigating phonological inventories across languages. He argued that phonetic categories are distinguishable enough using only acoustic information (for a critique, see O'Shaughnessy, 1996). It has to be noted though that although many languages indeed show acoustically distinct categories only that many other languages successfully distinguish between acoustically very similar categories. In contrast, Galantucci, Fowler, and Turvey (2006) reviewed the *Motor Theory* contributions and suggested that although there are other possible accounts of the current findings of the debate, the account proposing that perceiving of sounds is accomplished through perceiving of articulatory gestures seems to provide the most coherent explanation for different empirical results. Moreover, the involvement of the motor system in speech perception is also supported by recent findings in cognitive neuroscience, such as mirror neurons (e.g., Fadiga, Craighero, Buccino, & Rizzolatti, 2002), although again it has to be noted that many researchers are not convinced that mirror neurons are used in speech perception (e.g., Rogalsky, Love, Driscoll, Anderson, & Hickok, 2011).

This study aims at adding empirical evidence to this ongoing debate by providing perceptual results for stimuli which are acoustically similar but articulatorily signal different vocal tract configurations. We manipulated acoustic and articulatory information in our stimuli and examined whether listeners only rely on acoustic information, or alternatively recover the underlying articulatory information from the acoustic signals.

1.2. Articulation and acoustics of sibilants

Sibilants are a subcategory of fricatives. Fricatives are a category of consonants characterized by frication noise. The frication noise is produced mainly by two sources. The first source is the turbulence generated by air passing through a narrow constriction in the vocal tract with high velocity. The second source is an obstruction in front of the narrow channel along the vocal tract, such as teeth. Based on their source of frication noise, fricatives are divided into two subcategories: non-sibilants with only narrow constrictions as the primary source, and sibilants with an additional obstruction as the primary source in addition to the posterior narrow constriction (Ladefoged & Maddieson, 1996). Note that the discussion below will focus on voiceless sibilants only since we used voiceless sibilants as our stimuli but not their voiced counterparts for the reason to exclude the influence of the laryngeal voicing source.

Languages have different inventories of voiceless sibilants that are contrastive in places of articulation. For example, English and French have a two-way contrast, and Polish and Mandarin Chinese have a three-way contrast. In this study, we are interested in how listeners perceive the two-way contrast in English. English has two voiceless sibilants: the alveolar /s/ produced by placing the tip or blade of the tongue against the alveolar ridge, and the postalveolar /ʃ/ produced with a wider constriction posterior (further back) from the alveolar ridge. In addition, English /ʃ/ can be produced with the secondary articulation of lip rounding but /s/ is not (Ladefoged & Maddieson, 1996).

Previous acoustic studies have tried to identify acoustic cues which can be used to distinguish fricatives produced with different places of articulation. These potential acoustic cues include frication noise, noise amplitude, noise duration and vocalic formant transition (for an overview, see Jongman, Wayland, & Wong, 2000). Here we will only focus on frication noise and vocalic formant transition because they are the only two measures that are relevant for the manipulation of our stimuli.

For frication noise, sibilants generally show clearer spectral properties than non-sibilants. More specifically, in a spectrum of a sibilant, there are usually clearer spectral peaks which represent higher energy levels concentrated at a specific range of frequencies (e.g., Hughes & Halle, 1956). Spectral properties are primarily determined by the shape and size of the cavity anterior to the narrow constriction in the vocal tract (Jongman et al., 2000). For the alveolar fricative, this cavity would be found between the narrow channel and the teeth.

There are two primary acoustic measures of the frication noise part of a sibilant: the location of the highest spectral peak(s) and spectral moments. The highest spectral peak location is a specific range of frequencies at which the highest primary spectral peak occurs, i.e. the frequencies where the highest acoustic energy can be found. This primary highest spectral peak is largely determined by the size of the anterior cavity to the narrow constriction in the vocal tract. The primary highest spectral peak occurs at a higher range of frequencies as the anterior cavity becomes shorter. An alveolar /s/ usually peaks above

5 kHz and a postalveolar /ʃ/ usually peaks at lower frequencies around 3 kHz. This is because /ʃ/ is produced by placing the tongue (and thus the narrow channel) further back, which forms a longer anterior cavity. However, the actual spectral peak location for each of these two sibilants is also affected by speaker variability and surrounding vowels (Jongman et al., 2000). For example, in Jongman et al. (2000), an alveolar /s/ peaks at 6839 Hz, and a postalveolar /ʃ/ at 3820 Hz. Spectral peak location is regarded as one of the major cues to acoustically differentiate /s/ and /ʃ/ and thus to differentiate place of articulation.

Another important measure of frication noise is spectral moments. Spectral moments reflect detailed statistical analyses of the distribution of energy spreading over all relevant frequencies. The first four spectral moments are mean, variance, skewness, and kurtosis. Here we will only discuss the first spectral moment, the spectral mean, because it is the only measure directly related to the manipulation of our stimuli. A spectral mean is the mean of the overall distribution of energy spread over all frequencies of a specified frequency range. Note that the spectral mean is also termed as center of gravity (COG) in some studies, such as Nowak (2006). COG is a phonetic term, whereas the spectral mean term is derived from statistics research. Similar to spectral peak location, COG is also influenced by the size of the anterior cavity in the vocal tract. A shorter anterior cavity will result in a higher spectral mean. For example, in Jongman et al. (2000), the spectral mean for an alveolar /s/ is 6133 Hz and that for a postalveolar /ʃ/ is 4229 Hz.

If we compare the two acoustic measures of frication noise, the difference between the location of the highest spectral peak and the location of COG of a given fricative token depends on how much energy spreads outside the frequency range of its primary spectral peak. Figure 1.1 shows the schematic spectra of /s/ and /ʃ/ in Jongman et al. (2000), and summarizes their spectral peak locations and COGs. Note that these spectra were produced schematically for illustrative purposes only and are not based on actual tokens. The COG of /s/ (6133 Hz in this example) is usually lower than its spectral peak location (6839 Hz) because there is significant energy at lower frequencies which drags the spectral mean to the left from the spectral peak. Similarly, the COG of /ʃ/ (4229 Hz) is usually higher than its spectral peak location (3820 Hz) because there is energy at higher frequencies which drags the COG to the right from the spectral peak.

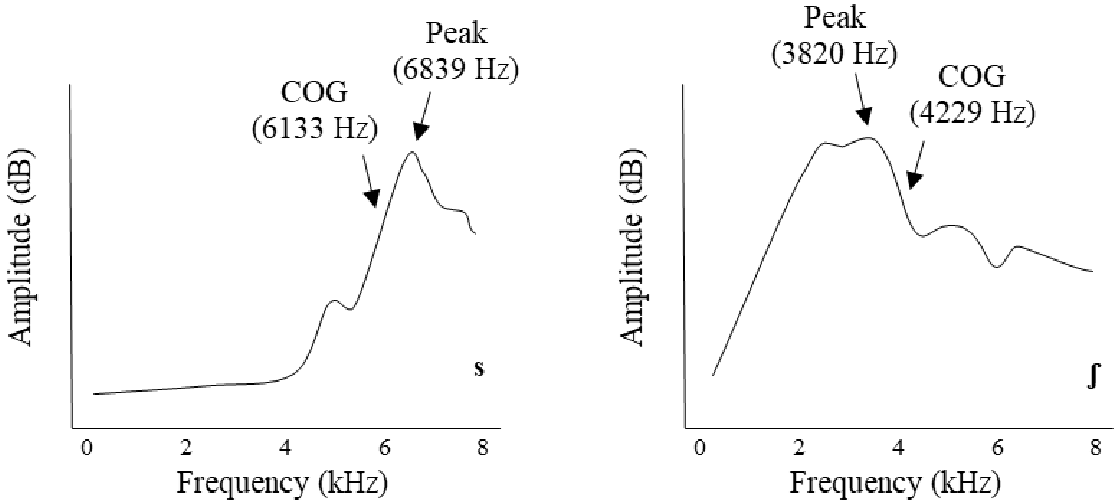

Figure 1.1. Schematic spectra of /s/ and /ʃ/ in Jongman et al. (2000).

Another example to illustrate the difference between these two measures is by comparing the same sibilant produced at different levels of loudness. Figure 1.2 shows the schematic comparison of the spectra of a modal (normal) and a louder /ʃ/. Note that again these schematic spectra are not based on actual tokens. Since the primary spectral peaks are largely determined by the anterior cavity, which is not altered even if the loudness increases, they remain the same in both the modal and louder /ʃ/. However, when a louder /ʃ/ is produced, the additional energy increase will mainly increase at higher frequencies only. Thus, the COG will be dragged to the right further away from the spectral peak.

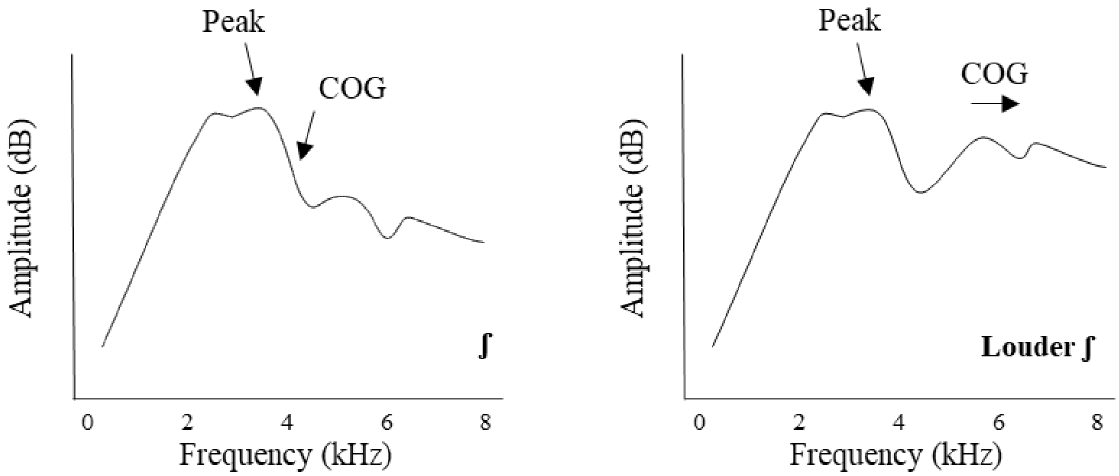

Figure 1.2. Schematic spectra of normal and louder /ʃ/.

If we compare the type of information each of the two measures represents, the location of the highest spectral peak is more related to articulatory information because it is primarily determined by the size of anterior cavity and thus by the articulatory setting. While COGs are partially determined by the size of anterior cavity, they are also determined by energy spreading outside the frequency range of the primary spectral peak. This gives a more general representation of the distribution of energy at different frequencies, thereby providing an overall picture of the available acoustic information. In this sense, we can regard spectral peak locations as a measure for articulatory information, and spectral means as a measure for acoustic information.

The second acoustic cue we are interested in is vocalic formant transition. The transition here specifically refers to the change of formants in the vowel immediately preceding or following the fricative. There are two primary acoustic measures of vocalic formant transition: locus equations and F2 (second formant) onset. Locus equations represent the dynamic changes of vocal tract configurations from a fricative to its proceeding vowel, by analyzing the onset and midpoint F2 values of the proceeding vowel. The changes of these F2 values can serve as a cue for identifying the place of articulation of a fricative. The second measure is F2 onset, the onset F2 value of the vowel immediately following the fricative. Note we will only use F2 onset as the measure of vocalic formant transition in our stimuli because previous studies suggest that it is a more reliable measure, as compared to locus equations which have been shown to be less consistent (e.g., Jongman et al., 2000). F2 onset values are determined by the size of the back resonance cavities in the vocal tract. As the constriction forming the fricative moves further back, which results in a shorter back resonance cavity, the F2 value of the proceeding vowel starts at higher frequencies. For example, in Jongman et al. (2000), the mean F2 onset following an alveolar /s/ is 1832 Hz, and that of /ʃ/ is 1982 Hz. The F2 onset of /ʃ/ is higher because the constriction is formed further back than /s/.

In summary, in this study, we are interested in two major acoustic cues to differentiate sibilants: frication noise and vocalic formant transition. To measure the manipulation of our stimuli, we will use spectral peak locations and COGs as the acoustic measures for frication noise, and F2 onset as the acoustic measure for vocalic formant transition. We want to see how changes in these measures by manipulating the acoustic properties of acoustic sibilant items are related to how listeners perceive these sibilants. In the following section, we will discuss how these acoustic cues have been shown to influence perception in previous studies.

1.3. Perception of sibilants

To form phonetic categories, listeners have to extract several cues available in acoustic signals. Previous studies of this area have attempted to build up the relationship between acoustic cues and perception by looking at how acoustic cues influence listeners' identification of different phonetic categories or different phonetic contrasts. There are two important aspects of perceiving acoustic cues: what acoustic cues listeners rely on, and how acoustic cues are weighted in terms of their importance in perception. The way how listeners use acoustic cues is influenced by various factors, for example linguistic experience (e.g., Polish vs. English listeners in Zygis & Padgett, 2010) and age (e.g., child vs. adult listeners in Nittrouer & Miller, 1997), among others, and is flexible depending on what acoustic cues are available (e.g., frication noise vs. vocalic formant transition in McGuire, 2007).

Previous studies of English sibilant perception have examined the acoustic cues English listeners use to distinguish between sibilants that are contrastive in their places of articulation. The two widely studied acoustic cues are spectral cues in the frication noise and the dynamic movement of formants in neighboring vowels. Most of the results have shown that the spectral cues in frication noise overrides vocalic formant transition as the primary acoustic cue. For example, Harris (1957) cross-spliced the frication noise of English sibilants /s/ and /ʃ/ with the vowel portions from their counterparts in CV syllables. Listeners' identification was primarily determined by the fricative noise alone without significant influence from the vowel portions. Similar results were observed in LaRiviere, Winitz, and Herriman (1975) with manipulated vocalic formant transitions, and Heinz and Stevens (1961) with synthesized stimuli: vocalic formant transitions did not influence how their listeners identified /s/ and /ʃ/. From a developmental perspective, Nittrouer and her colleagues (Nittrouer, 1992, 2002; Nittrouer & Miller, 1997) observed that children weighted vocalic formant transition more than adult listeners but the effect of vocalic formant transitions decreased as the age of listeners increased.

However, some studies have shown that vocalic formant transition can also have an influence on adult sibilant perception. For example, Delattre, Berman, and Cooper (1962) created a set of synthesized stimuli by using neutralized sibilants and strengthening vocalic formant transitions. Note that the voiced sibilants /z/ and /ʒ/ were used in their study because voiceless sibilants could not be synthesized for some technical reason. Their results showed that vocalic formant transitions could also serve as an acoustic cue for distinguishing /z/ and /ʒ/. Moreover, Whalen (1984) cross-spliced the frication noise of /s/ and /ʃ/ with matching and mismatching vocalic formant transitions. Although the final decisions made by his listeners were only based on frication noise, mismatching vocalic formant transitions increased the reaction time for making the judgments, suggesting that vocalic formant transitions were taken into account in the sibilant identification process. In a subsequent study, Whalen (1991) combined two frication noises of different lengths (e.g., combining 50 ms of /s/ with 150 ms of /ʃ/, or combining 150 ms of /s/ with 50 ms of /ʃ/). While identifications were usually based on the longer part of the fricative noise, listeners were also influenced by the proceeding vowels whose formant transition matched either the longer or shorter frication noise.

In summary, when English listeners identify English sibilants /s/ and /ʃ/, frication noise overrides vocalic formant transition as the primary acoustic cue. Although vocalic formant transition is likely to be processed, its contribution to the final decisions of identification is restricted.

Our study is particularly motivated by Polish listeners' remarkable identification of Polish sibilants that are acoustically extremely similar (see Nowak, 2006). Polish has a three-way contrast for voiceless sibilants: dental *s*, retroflex *sz* and alveolo-palatal *si*. Note that spelling conventions are used here as in Nowak (2006) because of the disagreement on the transcription of Polish sibilants. Acoustically, retroflex *sz* and alveolo-palatal *si* are extremely similar. From the acoustic analysis of his stimuli, COGs averaged across different vowel contexts were 5592 Hz for retroflex *sz* and 5619 Hz for alveopalatal *si*. There are two primary spectral peaks in retroflex *sz*, and the spectral peak locations were 3359 Hz and 5405 Hz. The single location for alveolo-palatal *si* was 3832 Hz. The difference in COGs between the two sibilants is extremely subtle, which suggests that their overall acoustics are extremely similar. This acoustic similarity was confirmed in Zygis, Pape, and Jesus (2012) where acoustic spectra and all acoustic parameters of Polish retroflexes and alveolo-palatals were strikingly similar. Despite this considerable acoustic similarity, Polish listeners were able to reliably distinguish between retroflexes and alveolo-palatals when only isolated fricative noise was presented (Nowak, 2006). This raises the question how listeners could accomplish this remarkable identification. If they, indeed, only relied on acoustic information capturing the overall acoustics, identification should be extremely challenging. However, if we also look at spectral peak locations, the difference between the two sibilants is greater. A solution to solve the discrepancy between the striking acoustic similarity and the robust discrimination of Polish listeners could be the use of additional articulatory information, such as articulatory settings, while parsing the acoustic signal. Such articulatory information representing how the sibilants were produced in the vocal tract might be beneficial in the process of identification. Our question is whether listeners restore articulatory information to help during the process of identification. As suggested in the earlier discussion on how we can acoustically measure frication noise of sibilants, COGs and spectral peak locations can be regarded as the representations of acoustic and articulatory information, respectively. This makes sibilants a very suitable type of phonetic categories which we can manipulate to study this question.

Previous studies have compared frication noise and vocalic formant transition as acoustic cues for sibilant identification. Nonetheless, none of them have focused on the internal acoustic properties of frication noise, i.e. the spectral distribution of energy of frication noise. Particularly, we are interested in how sibilant identification may be influenced by manipulations of acoustic and articulatory information available in the spectral distribution of frication noise. We manipulated the two sibilants in a way that the acoustic shape of the manipulated speech sound becomes very similar to the natural sibilant alternative (i.e. the natural alternative sibilant of the manipulated /s/ is a prototypical /ʃ/, and the natural alternative sibilant of the manipulated /ʃ/ is a prototypical /s/) to create an auditory confusion based on the acoustic information, while keeping the underlying articulatory information (based on the first highest spectral peak) constant. In other words, we manipulated, for example, /s/ so it would acoustically resemble a prototypical /ʃ/ and compared it to a manipulated /ʃ/ that acoustically resembled a prototypical /s/. By looking at the interaction between the manipulations and identification, we can get insight into whether listeners can recover articulatory information from the signals. Our hypothesis was that if listeners, indeed, only rely on acoustic information, they should switch identification (from /s/ to /ʃ/ and /ʃ/ to /s/) because the manipulated sibilants were acoustically similar (and thus confusable) to their natural alternative sibilants. However, if listeners are able to recover articulatory information of the unaltered primary spectral peaks, such a switch would not occur because the original sibilants, but not the manipulated part, is responsible for the building of the perceptual construct. We tested (Canadian) English listeners using (Canadian) English sibilants. Previous studies have shown that fricative noise is the primary acoustic cue English listeners use, which makes them ideal listeners for us to examine how the internal acoustic properties of frication noise may influence identification.

In addition, we were interested in whether vocalic formant transition becomes more important when listeners are faced with gradually increased difficulty when identifying sibilants. For this aim, we cross-spliced natural and manipulated sibilants with vowel contexts that favor either the original or alternative sibilants. Our research question is if listeners rely more on vocalic formant transitions when frication noise is acoustically confusing (manipulated /s/ resembling natural /ʃ/ versus manipulated /ʃ/ resembling natural /s/). The first possibility is that listeners interpret the manipulated sibilants only or largely based on vowel contexts, suggesting that greater importance is given to vocalic formant transitions the more confusing the frication noise becomes. The second possibility is that the two types of vowel contexts do not lead to different identification in the manipulated sibilants, suggesting that frication noise still serves as the primary cue to determine identification, thus disregarding the additional acoustic information available in vocalic formant transitions.

2. Methods

2.1. Participants

Thirty-six native speakers of Canadian English were recruited from the Linguistic Research Participation System administrated by the Department of Linguistics and Languages at McMaster University. They were undergraduate students and received course credit for participating in the experiment. All of them reported normal vision and hearing. Four subjects who had knowledge of Mandarin Chinese or Polish were excluded. The reason for excluding them was that Mandarin Chinese and Polish have a three-way contrast in voiceless sibilants, and such linguistic experience may result in perception different from that of other participants (e.g., Polish vs. English listeners in Zygis & Padgett, 2010). The data from the other 32 participants were further analyzed.

2.2. Stimuli

The stimuli were recorded by a female native speaker of Canadian English in an acoustically treated (silent) room. The recording was made on a fanless notebook computer using a Shure SM-27 microphone connected to a Tascam US-122MKII audio interface with a sampling frequency of 44,100 Hz and 16 bits. Several pseudowords in a VCV structure, [asa] and [aʃa], were recorded, and the most prototypical items were selected for further processing. The recorded stimuli were processed and manipulated in the phonetic software Praat (Boersma & Weenink, 2018) and in the audio processing software Sound Forge Pro with the WAVES Linear Equalizer (Waves LinEQ) for spectral manipulation. The Waves equalizer allows for very precise frequency manipulation with high sideband suppressions. The three components of each token, the initial vowel, medial sibilant consonant, and final vowel, were divided by slicing at the (zero crossing) time points where the first traces of aperiodic waveforms of frication noise started and ended. The two sibilants were extracted from their vocalic context and both normalized to the length of 336 ms.

These normalized tokens of /s/ and /ʃ/ were manipulated with the goal to alter acoustic information while keeping underlying articulatory information of the front cavity resonances as identical as possible. Overall, the aim was to manipulate the frequency spectrum of a manipulated sibilant so it would resemble the prototypical acoustic spectrum of the other place of articulation, while the primary spectral peak representing the anterior cavity in the vocal tract configuration remained the same. That is, the manipulated /s/ would have an identical spectrum as the prototypical /ʃ/, and the manipulated /ʃ/ would have an identical spectrum of a prototypical /s/. In other words, just based on spectral (acoustic) information the prototypical and manipulated stimuli were made to be as identical and thus perceptually confusable as possible, whereas the underlying articulatory information was contrary.

The whole frequency range was divided into two subranges at the center frequency of 5 kHz, the midpoint between the primary spectral peaks of /s/ and /ʃ/. This 5 kHz point represents the division between *relevant frequencies* and *irrelevant frequencies* based on the articulatory first cavity resonance. Relevant frequencies represent the frequency ranges where primary spectral peaks appear, and irrelevant frequencies represent the other frequency ranges. For /s/, relevant frequencies are above 5 kHz and irrelevant frequencies below 5 kHz. The ranges are the opposite for /ʃ/: relevant frequencies are below 5 kHz and irrelevant frequencies above 5 kHz. Figure 2.1 shows the division of relevant and irrelevant frequencies with natural /s/ and /ʃ/ (relevant frequencies in grey). Note that the frequency range above 5 kHz also includes frequencies above10 kHz. This is the reason why, in the figure, the right frequency ranges (above 5 kHz) are wider than the left frequency ranges (below 5 kHz). The spectra presented here will show frequencies up to 16 kHz.

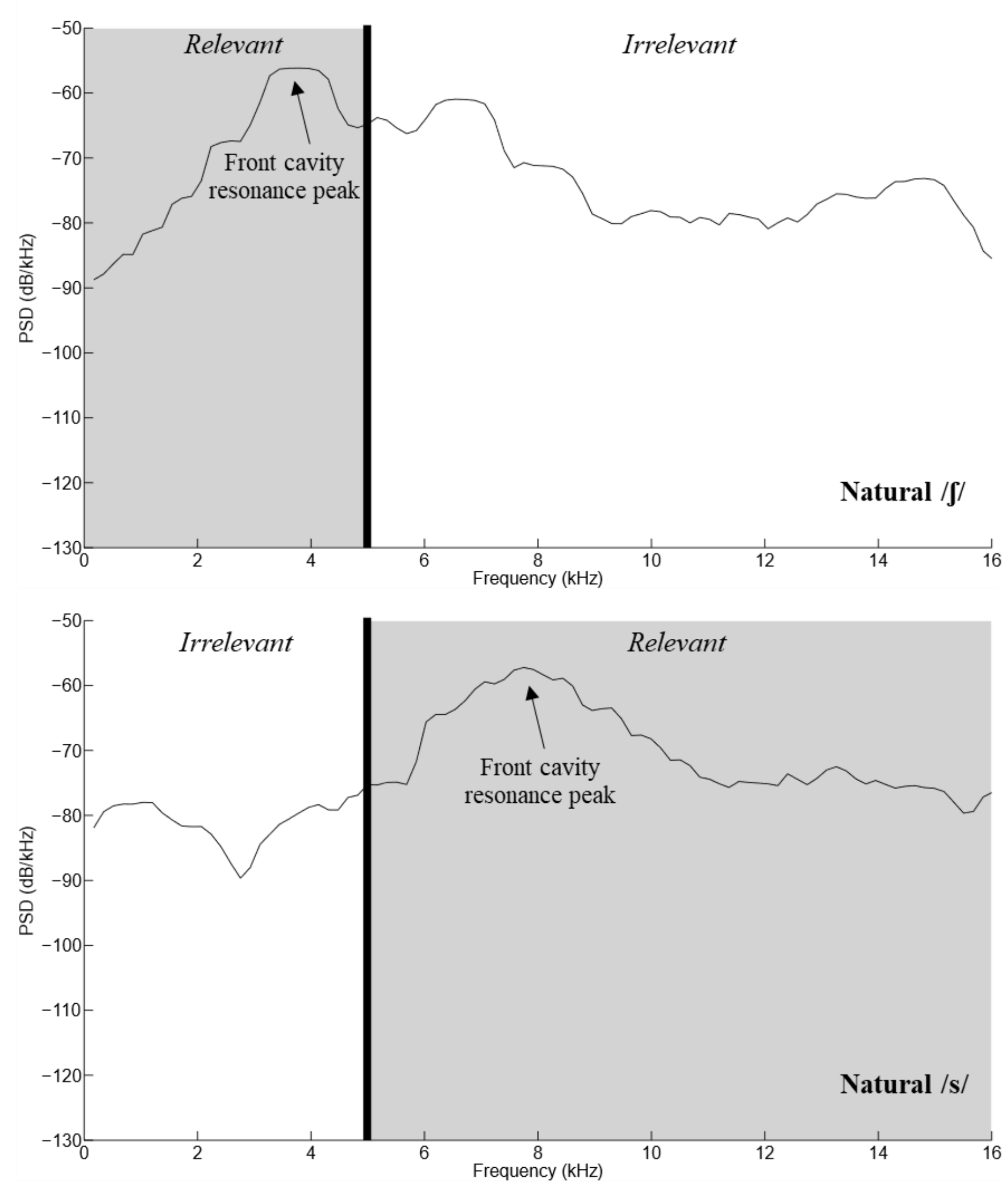

Figure 2.1. Relevant and irrelevant frequencies of natural /s/ and /ʃ/.

For each sibilant, seven steps were created by amplifying either relevant or irrelevant frequencies by different magnitudes in dB (decibel). Frequencies to be amplified were filtered with the Waves LinEQ Filter using high-accuracy shelving filters (V-Slope high-shelf and V-Slope low-shelf). All the steps were then normalized using European Broadcasting Union's (EBU) loudness standards to ensure that perceived loudness was identical across the different manipulated stimuli. For steps 1 and 2, *relevant* frequencies were amplified by 24 and 12 dB respectively. We expected that the enhancement of relevant frequencies should make identification even easier. Only two steps were created because a strong ceiling effect for the listeners was predicted. Step 3 employed the *natural tokens* (prototypes) without manipulation. For steps 4, 5, 6 and 7, *irrelevant* frequencies were amplified by 12, 24, 36, 48 dB respectively. As amplification of irrelevant frequencies increases towards the end of the continuum at step 7, the overall spectral shapes of the manipulated sibilants became increasingly similar to their natural alternative sibilants. Table 2.1 summarize the manipulation in each step.

Table 2.1.

Summary of manipulation in stimulus steps

Step	**1**	**2**	**3 (Natural; Prototypes)**	**4**	**5**	**6**	**7**
Amplified Frequencies	Relevant /s/: above 5 kHz /ʃ/: below 5 kHz		--	Irrelevant /s/: below 5 kHz /ʃ/: above 5 kHz			
Degree of Amplification (dB)	24	12	--	12	24	36	48

Figure 2.2 illustrates how a manipulated /ʃ/ in step 6 was created from a natural /ʃ/.

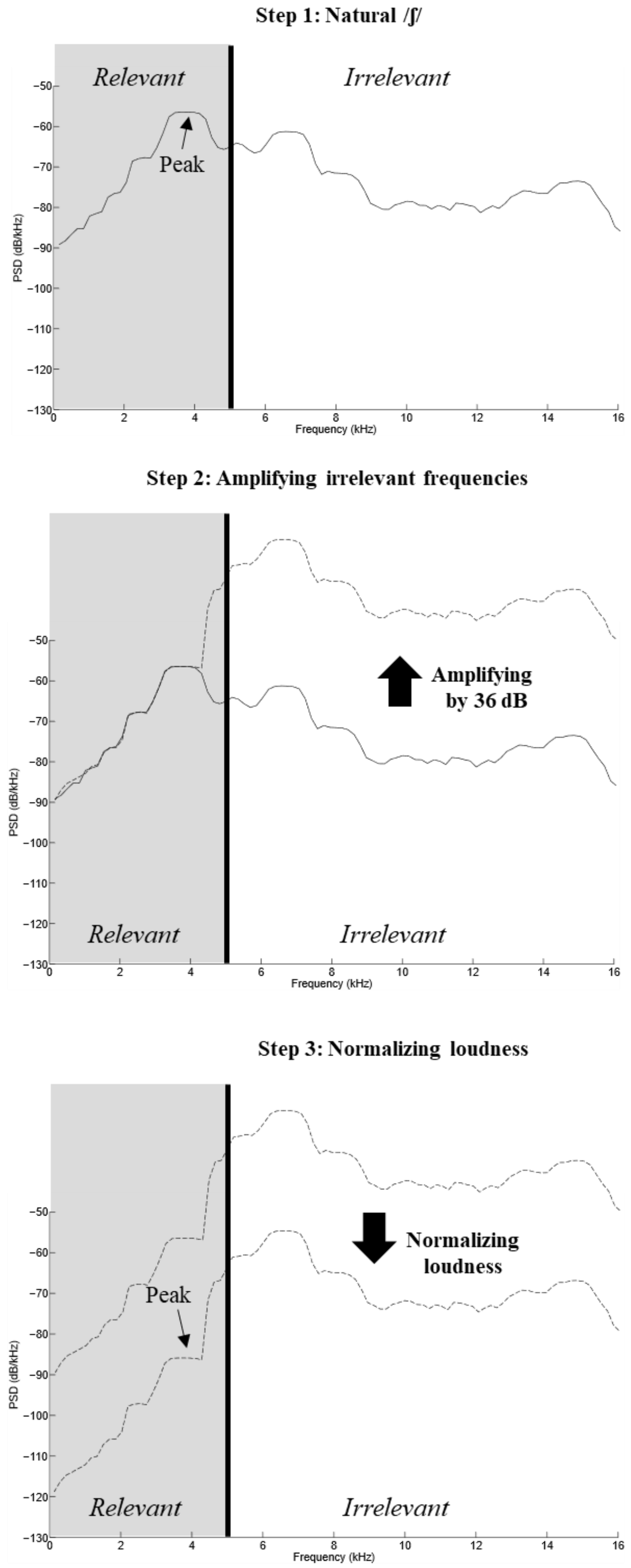

Figure 2.2. Manipulation of /ʃ/ in step 6.

The irrelevant frequencies (above 5 kHz) of a natural /ʃ/ were first amplified by 36 dB, and then the whole token was normalized in loudness (to be identical to all other stimuli). Note that a difference in loudness could significantly change listeners' identification which we wanted to avoid. Also, note that the primary spectral peaks determined by the anterior cavity are the same before and after manipulation. Figure 2.3 shows the spectra of each stimulus step.

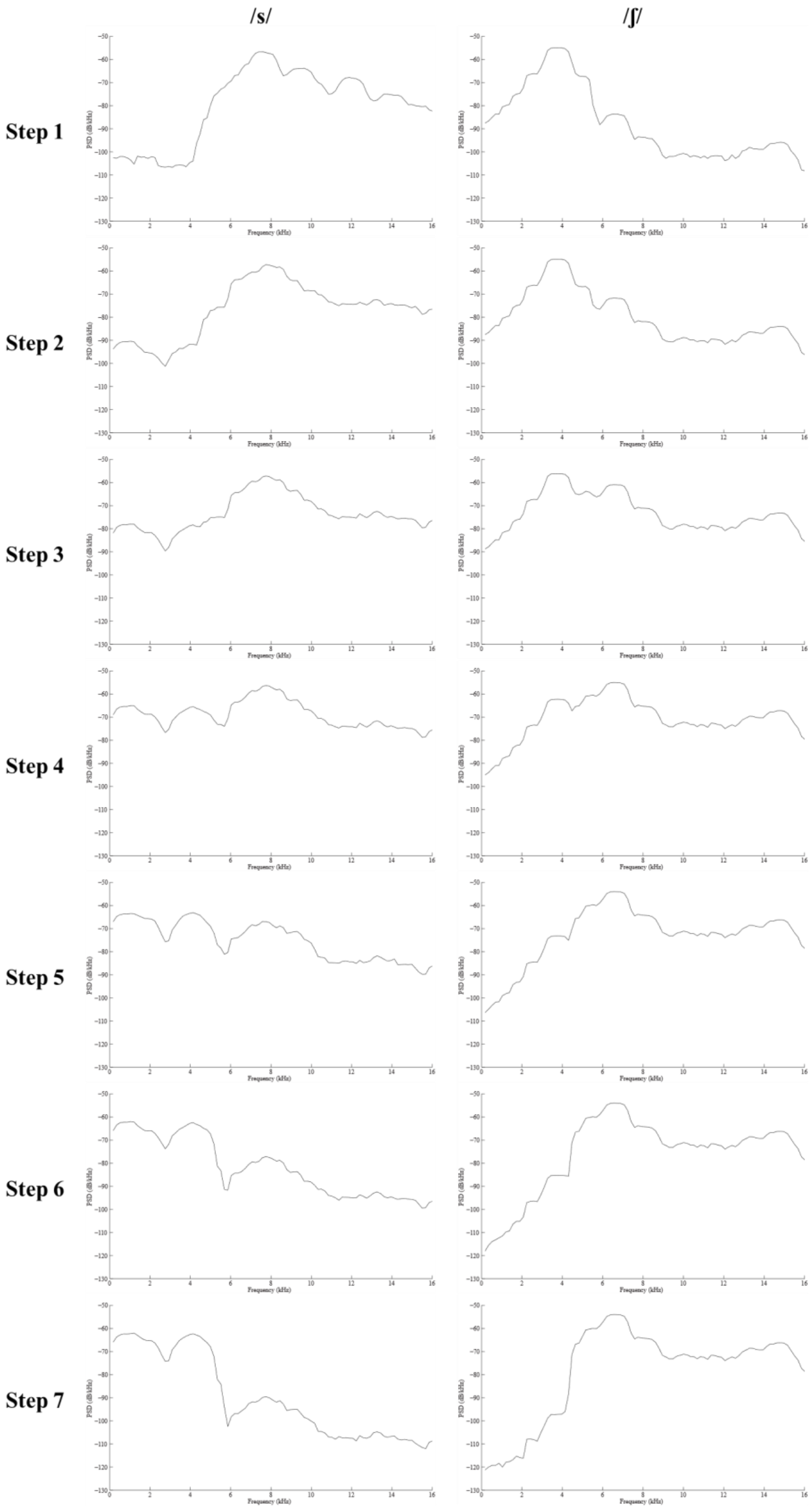

Figure 2.3. Spectra at 7 stimulus steps.

Figure 2.4 compares the natural sibilants with the manipulated stimuli of their alternative sibilants in step 6.

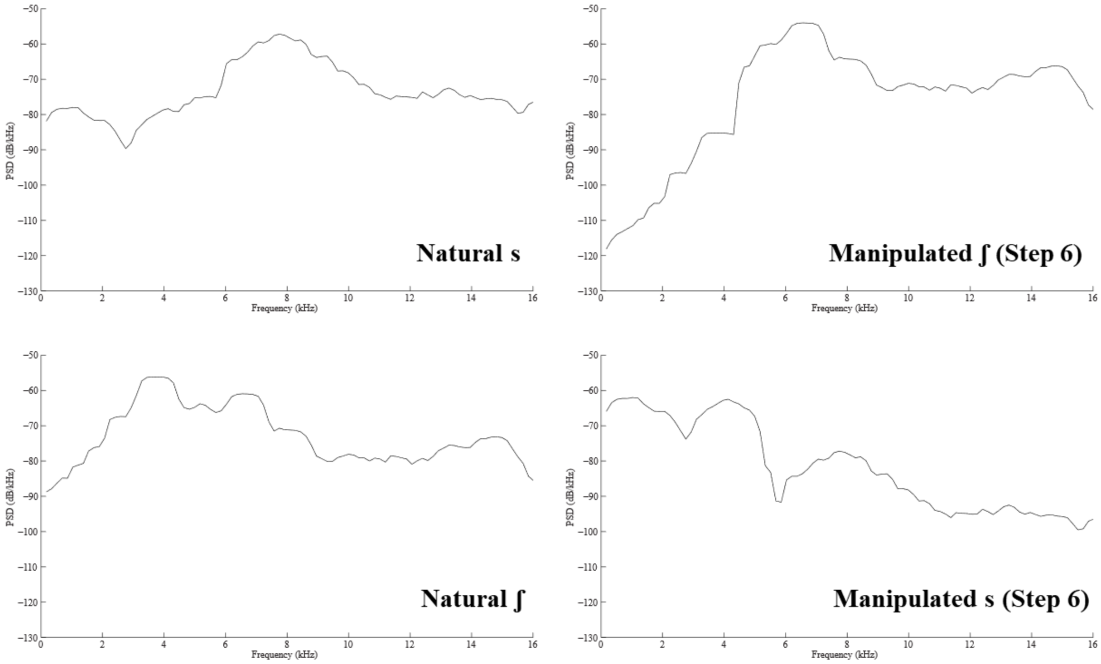

Figure 2.4. Comparison between natural stimuli and manipulated stimuli in steps 6.

The spectral distribution of energy of a natural /s/ and a manipulated /ʃ/ are very similar. Similarly, the spectral distribution of energy of a natural /ʃ/ and a manipulated /s/ are very similar. Figure 2.5 compares the manipulated /s/ and /ʃ/ in step 6 (in dash lines) with their natural alternative sibilants (in solid lines). Note that the overall spectral shapes of the manipulated stimuli highly resemble those of their natural alternative sibilants (illustrated by the thick grey lines).

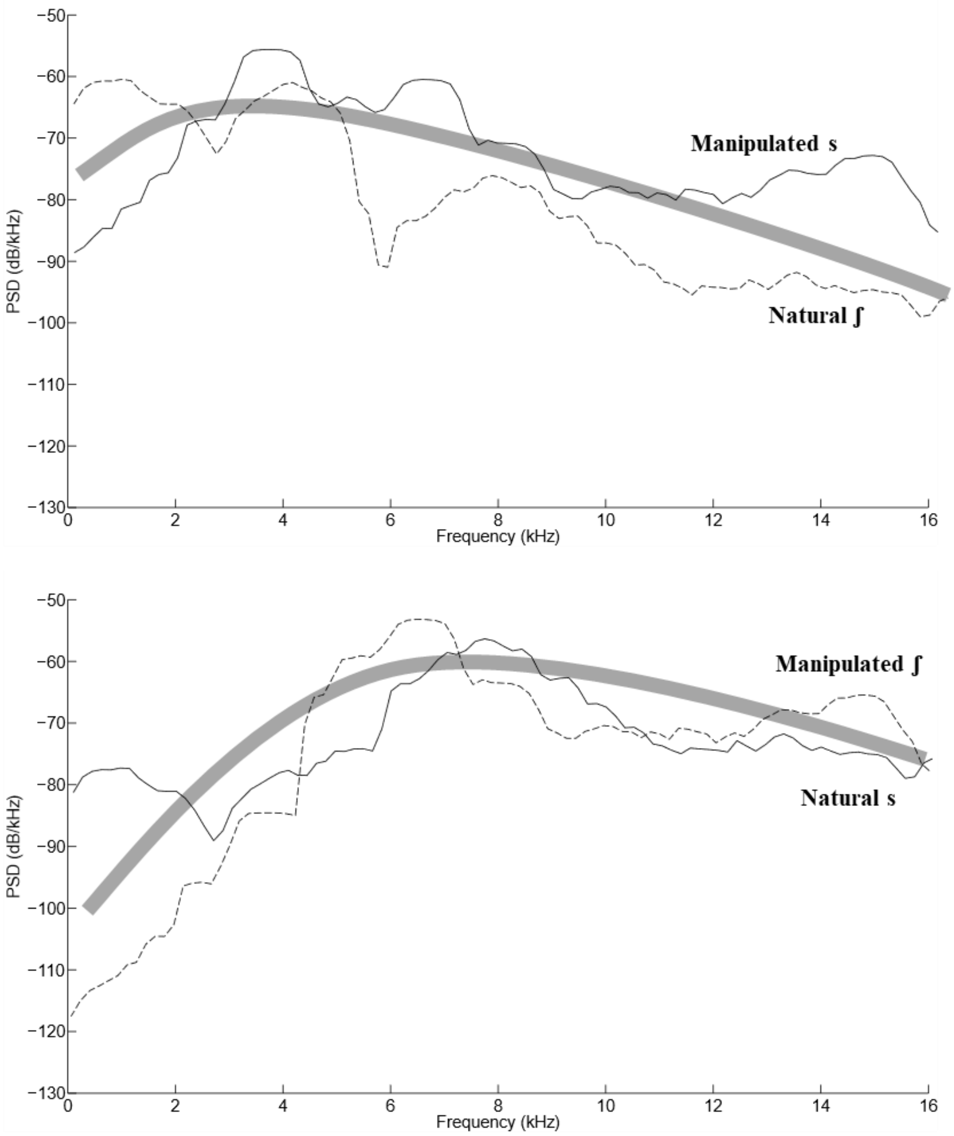

Figure 2.5 Comparison between natural stimuli and manipulated stimuli in step 6.

The steps were then used to create three different conditions: long isolated fricative noise, short isolated fricative noise, and fricative in a vowel context (VCV structure). For long fricative noise, the steps were used directly without further manipulation, and they are 336 ms in length. For short fricative noise, a shorter version with 150 ms in length was created for each step. This duration resembles more the length of prototypical speech sounds (100-200ms for running speech phonetic categories). Thus, the length of the short versions is closer to the length of a fricative that is usually produced in normal continuous streams of speech. The short versions were cut from the medial part of the original steps. These conditions were created to examine whether the length of the stimuli influences perception because Jongman (1989) has shown that duration can influence perception of frication of different places of articulation. For the VCV vowel context condition, the long versions of the steps were spliced into the original vowel contexts (vowels from [asa] for /s/ steps and [aʃa] for /ʃ/ steps) and the alternative sibilant vowel contexts (vowels from [aʃa] for /s/ steps and [asa] for /ʃ/ steps). This condition was created to examine whether vowel contexts have a stronger effect on steps with amplified irrelevant frequencies.

2.3. *Acoustic properties of the stimuli*

The acoustic properties of the stimuli are presented below. The acoustic measures for fricative noise include spectral peak locations and the four spectral moments, namely mean (first moment, also known as Center of Gravity, COG), variance (second moment), skewness (third moment), and kurtosis (fourth moment). These values were calculated using a 256-point Multitaper window at the midpoint of each step with a frequency range between 20 and 16000 Hz. The spectral moments were obtained based on Forrest, Weismer, Milenkovic, and Dougall (1988). Table 2.2 summarizes these values for each /s/ or /ʃ/ step.

Table 2.2.

Spectral peak locations and spectral moments (mean, variance, skewness, kurtosis) of fricative noise in stimulus steps

Fricative	**Step**	**Spectral peak location [Hz]**	**Spectral moments**			
			Mean [Hz]	**Variance [Hz]**	**Skewness**	**Kurtosis**
s	1	7579	8180	1594	2.15	6.42
	2	7752	8193	1629	2.40	8.41
	3	7752	8081	1713	1.75	8.63
	4	7752	7290	2562	-0.34	2.43
	5	4134	3982	2822	0.90	0.71
	6	1034	2700	1728	0.83	1.98
	7	1206	2579	1507	0.19	-1.16
ʃ	1	3790	3746	560	0.81	21.38
	2	3790	3837	852	4.77	51.05
	3	3790	4801	2037	2.52	8.91
	4	6546	6757	2373	1.92	4.74
	5	6546	7172	2232	2.43	5.73
	6	6546	7219	2219	2.50	5.82
	7	6546	7225	2218	2.50	5.82

The spectral peaks for /s/ stimuli occurred at approximately 8 kHz for the first four steps and moved to 4 kHz in step 5. The peaks further moved to 1 kHz in steps 6 and 7. This further decrease to lower frequencies is probably because lower frequencies were further apart from the frequency division point of the low-shelf filter at 5 kHz and thus further amplified. The lower frequencies were amplified more than the frequencies that were closer to 5 kHz, which made the original lower peak at 1 kHz become the highest peak. The highest spectral peaks for /ʃ/ occurred at approximately 4 kHz for the first three steps and moved to 7 kHz in the next four steps. For spectral means (or COGs), the values for steps with amplified relevant frequencies slightly increased in /s/ stimuli and decreased in /ʃ/ stimuli in comparison with those for natural stimuli. In steps with amplified irrelevant frequencies, the spectral means for /s/ stimuli gradually decreased becoming like a natural /ʃ/, and those for /ʃ/ stimuli gradually increased becoming like a natural /s/.

F2 onsets were used to measure vocalic formant transition in the vowel following each sibilant. The values were extracted at 10 ms after the end of the preceding fricative noise. Table 2.3 summarizes the F2 onsets of the two vowel contexts.

Table 2.3.

F2 onsets of vowel contexts

Fricative	**F2 Onset**
After /s/	1502 Hz
After /ʃ/	1735 Hz

The F2 of /s/ is 233 Hz lower than that of /ʃ/.

2.4. *Experimental procedure*

The experiment was conducted in a soundproof room located in the Phonetics Lab of McMaster University’s Department of Linguistics and Languages. The software used to present the experiment was Alvin 3 (Hillenbrand, Gayvert, & Clark, 2015). Stimuli were presented through a pair of Sennheiser HD 598 headphones with linear frequency response, connected to a Focusrite Scarlett 2i2 audio interface using a laptop computer. Instructions were given verbally with a printed guideline.

The experiment was an identification task with confidence ratings for each presented stimulus. For each trial, participants were first presented with a stimulus and then asked to make two responses. The first response was whether the stimulus was /s/ or /ʃ/, thus asking for a forced-choice identification. The second response was how confident their identification was by selecting a score from the Likert scale where 5 represented the most confidence and 1 represented the least confidence. Responses were given by clicking the buttons on the computer screen using a mouse. After clicking the identification and confidence responses, listeners had to click a button labeled with 'Okay' to submit their responses. There was no time limit for each trial. After responses had been made to a trial, a new trial began 1500 ms later.

Each participant completed eight blocks of the experiment, except one that completed only six blocks due to a technical problem. In each block, fifty-six items from the three different contexts were mixed together (long frication noise: 2 sibilants × 7 steps, short frication noise: 2 sibilants × 7 steps, and vowel context: 2 sibilants × 7 steps × 2 vowel contexts). In vowel contexts, participants were instructed to identify the medial consonants in the VCV structures. Items were presented in an online randomized order within each block. Preceding the main experiment, participants completed a practice block with ten items including only stimuli from steps 1 and 2. Responses (identification and confidence) and reaction times (measured from the time when the presentation of a stimulus started) were recorded.

3. Results

This chapter presents the identification, confidence, and reaction time data. Since all three types of data show a similar picture, we will focus on the identification data and only highlight the observations in the confidence data which are not shown in the identification data. The reaction time data will be presented but not discussed in detail.

Only trials with both identification and confidence responses were further analyzed. The data were averaged for each stimulus item by each participant for statistical tests. The figures below present grand means averaged across all the participants.

3.1. Identification data

In our identification task, the listeners were first asked to identify a stimulus as either /s/ or /ʃ/ in each trial. This section presents the proportions of /s/ responses to the /s/ and /ʃ/ stimuli, first the results for the isolated fricative noise contexts and then those for the VCV contexts. The proportions of /s/ responses were analyzed to examine the effect of our stimulus manipulation on the listeners' identification.

If our manipulation to amplify irrelevant frequencies have induced a phonetic categorical switch in the listeners' perception, the proportions of /s/ responses would change across the stimulus steps and show an S-shaped identification curve which is often observed in categorical perception. That is, listeners would perceive the natural stimuli and manipulated stimuli with amplified *relevant* frequencies as the original sibilants (steps 1 to 3, the /s/ stimuli as /s/ and the /ʃ/ stimuli as /ʃ/) and would switch to perceive the manipulated stimuli with amplified *irrelevant* frequencies, first more often and later always, as the alternative sibilants (steps 4 to 7, the /s/ stimuli as /ʃ/ and the /ʃ/ stimuli as /s/).

3.1.1. Isolated fricative noises

The results for the long and short isolated fricative noise contexts were first combined for analyses. Figure 3.1 shows the proportions of /s/ responses to the isolated fricative noise stimuli averaged across long and short fricative noise contexts.

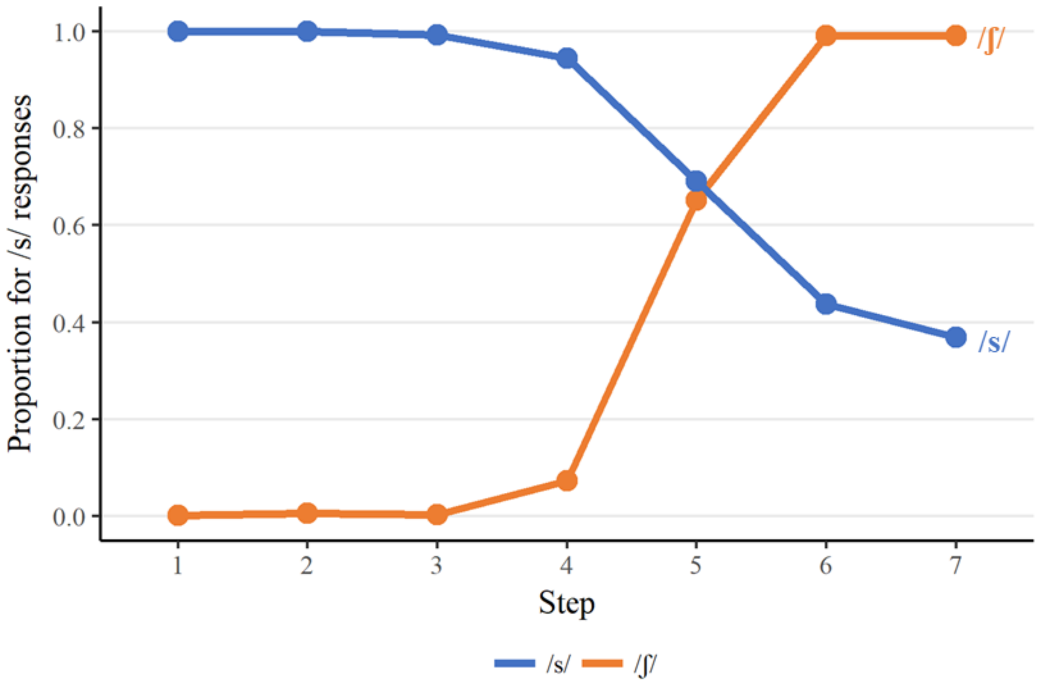

Figure 3.1. Proportions of /s/ responses to isolated fricative noise stimuli averaged across long and short fricative noise contexts.

The natural stimuli (step 3) and manipulated stimuli with amplified *relevant* frequencies (steps 1 and 2) of the two sibilants /s/ and /ʃ/ were well perceived as their original sibilants, i.e. the /s/ stimuli as /s/ and the /ʃ/ stimuli as /ʃ/. Our manipulation to amplify *irrelevant* frequencies at steps 4 to 7 led the listeners to identify the manipulated stimuli as the alternative sibilants. As the amplification of irrelevant frequencies increased in magnitude from steps 4 to 7, the stimuli were increasingly perceived as their alternative sibilants, i.e. the manipulated /s/ stimuli as /ʃ/ and the manipulated /ʃ/ stimuli as /s/. A marked difference between two sibilants was observed at steps 6 and 7: while the manipulated /ʃ/ stimuli were almost always identified as the alternative sibilant /s/, only approximately 60% of the manipulated /s/ stimuli were identified as the alternative sibilant /ʃ/. In other words, while there was a phonetic categorical switch in the /ʃ/ stimuli from complete /ʃ/ responses to complete /s/ responses, the /s/ stimuli did not show a phonetic categorical switch and remained confusing to the listeners even at the last two steps with greatest amplification of irrelevant frequencies. Therefore, although our manipulation to amplify irrelevant frequencies influenced the identification of both sibilants, the effect on the two sibilants was not completely symmetric.

Two sets of planned paired sample t-tests were performed to investigate the effect of our manipulation to amplify relevant and irrelevant frequencies on the listeners' identification. The Bonferroni-adjusted *p*-value for each set of tests is 0.025 (.05 / 2 = 0.025). The natural stimuli (step 3) were compared with first the manipulated stimuli with amplified *relevant* (steps 1 and 2) and then those with amplified *irrelevant* frequencies (steps 4 to 7).

To examine whether amplifying *relevant* frequencies can benefit identification, the first set of tests compared the proportions of /s/ responses for the natural stimuli (step 3) with those for the manipulated stimuli with amplified relevant frequencies (step 1) of the same sibilant. The comparison between the two types of /s/ stimuli did not show a significant difference, *t*(31) = -2.104, *p* = 0.044, nor did the comparison between the two types of /ʃ/ stimuli, *t*(31) = 0.571, *p* = 0.572. These test results suggest that our amplification of *relevant* frequencies did not change identification significantly.

To examine whether the manipulated sibilants with amplified *irrelevant* frequencies were perceived similarly as the natural stimuli of their alternative sibilants, the second set of tests compared the proportions of /s/ responses for the natural stimuli (step 3) of one sibilant to those for the manipulated stimuli with amplified irrelevant frequencies (step 7) of the alternative sibilant, i.e. the natural /s/ stimuli vs. the manipulated /ʃ/ stimuli, and the natural /ʃ/ stimuli vs. the manipulated /s/ stimuli. There was no significant difference between the natural /s/ and manipulated /ʃ/ stimuli, *t*(31) = 0.329, *p* = 0.745, suggesting that the manipulated /ʃ/ stimuli were perceived similarly as the natural /s/ stimuli. In contrast, the comparison between the natural /ʃ/ and manipulated /s/ stimuli showed a significant difference, *t*(31) = -6.211, *p* < .001. The listeners identified the manipulated /s/ stimuli as /ʃ/ less often than the natural /ʃ/ stimuli.

3.1.2. Long vs. short fricative noises

The results for the long and short fricative noise contexts were then compared to examine whether the length of fricative noise influenced the listeners' identification. Figure 3.2 presents the proportions of /s/ responses to the isolated long and short fricative noise stimuli separately.

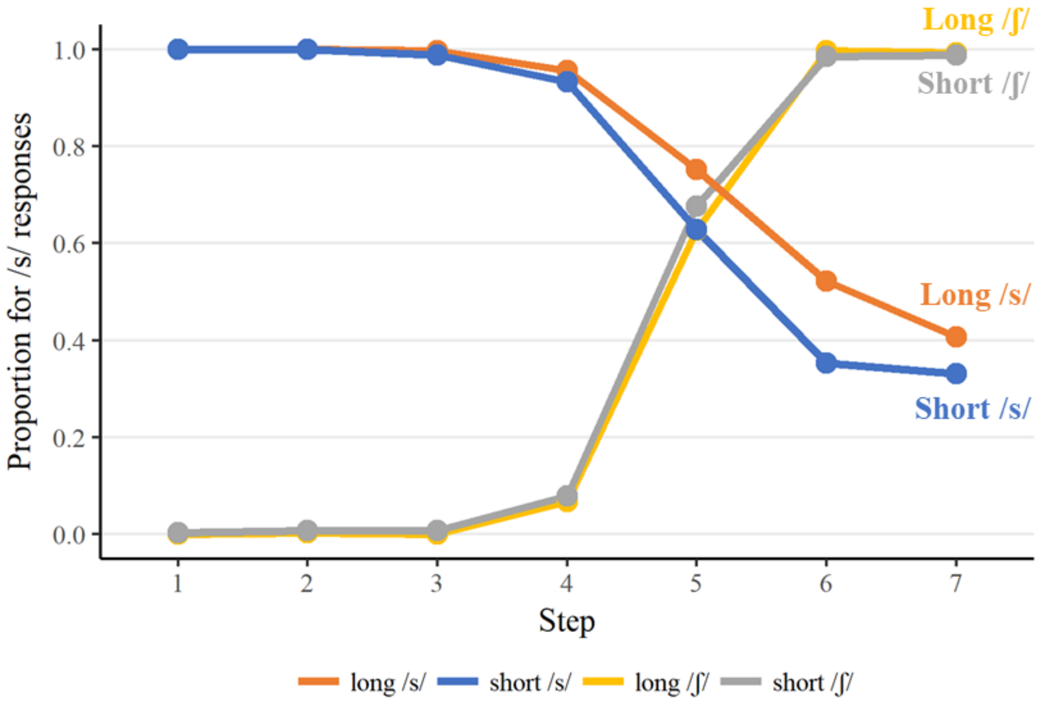

Figure 3.2. Proportions of /s/ responses to isolated long and short fricative noise stimuli.

These proportions for the /s/ and /ʃ/ stimuli were analyzed with two separate two-way repeated measures ANOVAs (one for each sibilant) using length (long vs. short fricative noises) and step (steps 1 to 7) as factors. There was a main effect of length for the /s/ stimuli, $F(1, 31) = 14.00$, $p < .001$, $\eta^2_p = 0.311$, but not for the /ʃ/ stimuli, $F(1, 31) = 1.66$, $p = 0.207$, $\eta^2_p = 0.051$. Planned paired sample t-tests were performed to compare the long and short /s/ stimuli at steps 4 to 7. The Bonferroni-adjusted *p*-value is 0.0125 (.05 / 4 = 0.0125). There was a significant difference at step 6, $t(31) = 3.16$, $p = 0.003$, where the long stimulus was less often identified as the alternative sibilant /ʃ/ than the short stimulus.

3.1.3. Original vs. alternative sibilant vowel contexts

The results for VCV stimuli were analyzed to examine the effect of vowel context on the listeners' identification. There were two types of vowel contexts: *original sibilant vowel context* (/s/ fricative noise in the /s/ vowel context, and /ʃ/ fricative noise in the /ʃ/ vowel context) and *alternative sibilant vowel context* (/s/ fricative noise in the /ʃ/ vowel context, and /ʃ/ fricative noise in the /s/ vowel context). Figure 3.3 shows the proportions of /s/ responses to the two fricatives in their original and alternative sibilant vowel contexts.

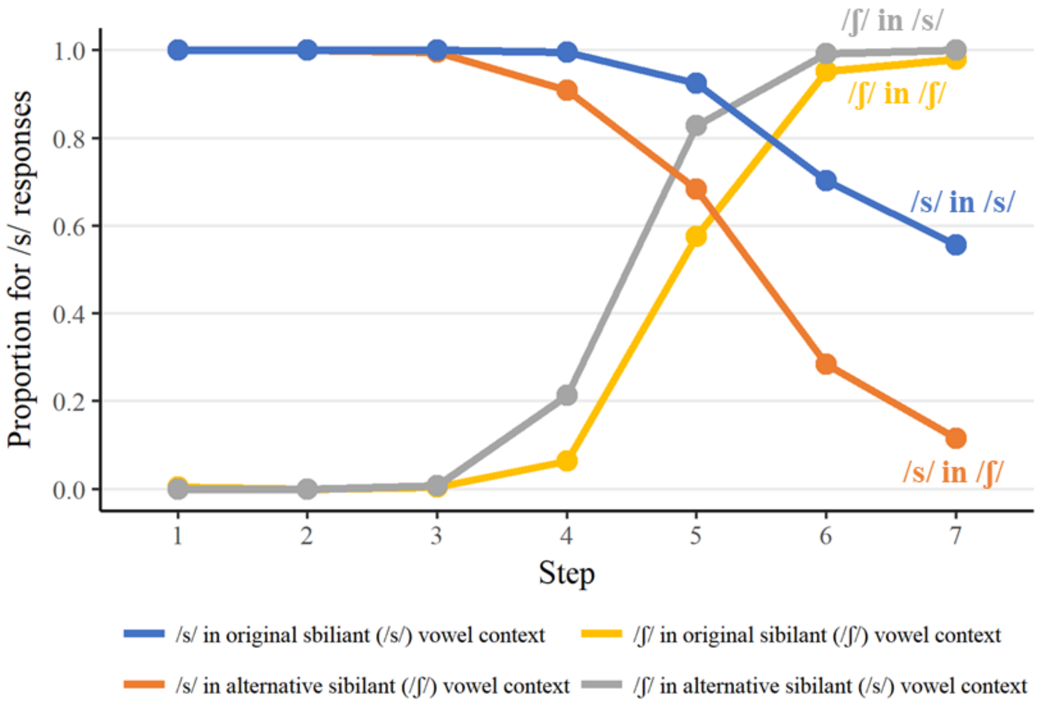

Figure 3.3. Proportions of /s/ responses to fricatives in original and alternative sibilant vowel contexts.

The natural stimuli (step 3) and the manipulated stimuli with amplified *relevant* frequencies (steps 1 and 2) of /s/ and /ʃ/ were well perceived as their original sibilants in either vowel context. For the manipulated stimuli with amplified *irrelevant* frequencies (steps 4 to 7), the vowel context biased the listeners to choose the sibilant matching the vowel context, i.e. the original sibilant in an original sibilant vowel context and the alternative sibilant in an alternative sibilant vowel context. The size of the effect of vowel context differed between sibilants and across steps. The size differences were examined by ANOVAs and t-tests presented below.

The proportions of /s/ responses for the /s/ and /ʃ/ stimuli were analyzed in two separate two-way repeated measures ANOVAs (one for each sibilant) using vowel context (original vs. alternative sibilant vowel contexts) and step (steps 1 to 7) as factors. For the /s/ stimuli, there were main effects of vowel context, $F(1, 31) = 87.1, p < .001, \eta^2_p = 0.738$, and step, $F(6, 186) = 126.7, p < .001$, $\eta^2_p = 0.803$, and an interaction between vowel context and step, $F(6, 186) = 31.5, p < .001, \eta^2_p = 0.504$. The results for the /ʃ/ stimuli were similar: there were main effects of vowel context, $F(1, 31) = 75.3, p < .001$, $\eta^2_p = 0.708$, and step, $F(6, 186) = 436.6, p < .001$, $\eta^2_p = 0.934$, and an interaction between vowel context and step, $F(6, 186) = 16.5, p < .001$, $\eta^2_p = 0.348$.While there was an effect of vowel context on both sibilants, the size of the effect was stronger on the /s/ stimuli ($\eta^2_p = 0.738$) than the /ʃ/ stimuli ($\eta^2_p = 0.708$).

For each sibilant, two sets of planned paired sample t-tests (one set for each sibilant) were performed to compare the proportions of /s/ responses for the two vowel contexts at steps 4 to 7. The Bonferroni-adjusted *p*-value for each set of tests is 0.0125 (.05 / 4 = 0.0125). For the /s/ stimuli, these proportions between two vowel contexts were significantly different at all four steps: step 4: $t(31) = -2.871, p = 0.007$, step 5: $t(31) = -5.228, p < .001$, step 6: $t(31) = -7.464, p < .001$ and step 7: $t(31) = -7.403, p < .001$. For the /ʃ/ stimuli, the difference was observed in step 4, $t(31) = 4.507, p < .001$ and step 5, $t(31) = 5.962, p < .001$. While the effect of vowel context on the /s/ stimuli continued to become increasingly stronger from steps 4 to 7, the effect on /ʃ/ were strongest at steps 4 and 5 and disappeared at steps 6 and 7. If we map these results for the VCV stimuli onto those for isolated fricative noise stimuli, the steps showing the strongest effect of vowel context were also the steps at which isolated fricative noise stimuli were the most confusing to the listeners, i.e. where identification was closest to chance level (50% /s/ responses and 50% /ʃ/ responses).

3.2. Confidence data

After identifying a stimulus, the listeners were asked to indicate how confident the identification of that stimulus was by choosing from 1 (the least confidence) to 5 (the most confidence). When analyzing the confidence data, we were particularly interested in the possible qualitative difference of the listeners' identification, specifically how confident the identifications were, which were not revealed in their final identification decisions.

3.2.1. Isolated fricative noises

The confidence scores for the long and short isolated fricative noise stimuli were first combined and analyzed to examine whether our manipulation influenced the listeners' confidence. Figure 3.4 presents the mean confidence scores averaged across the isolated long and short fricative noise stimuli.

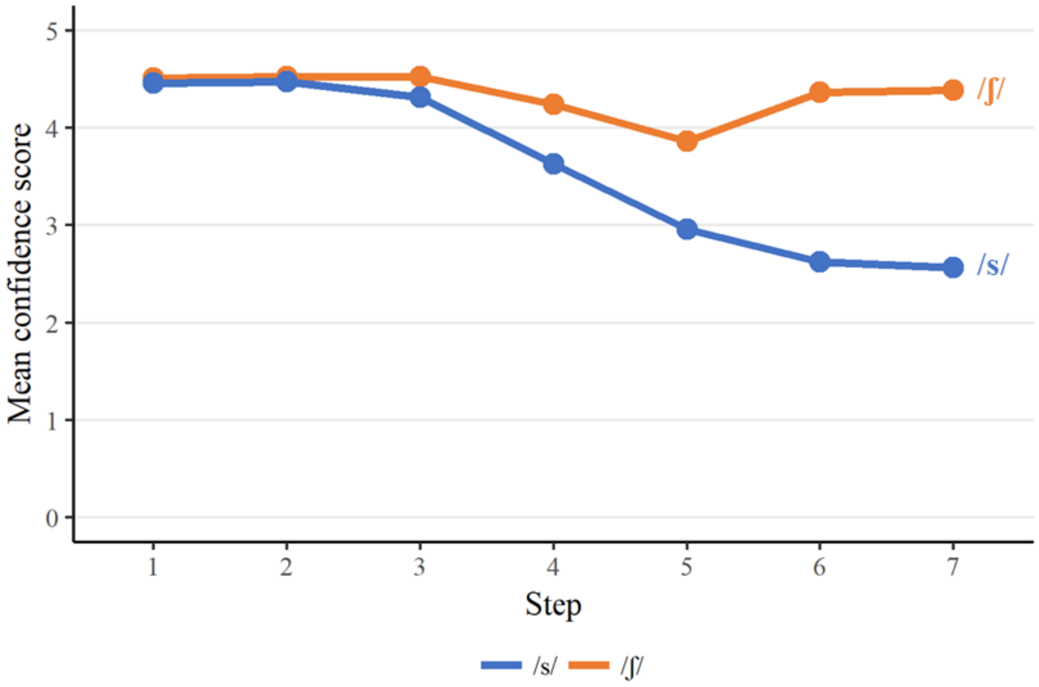

Figure 3.4. Mean confidence scores average across isolated long and short fricative noise stimuli.

The confidence scores dropped at the steps that were confusing to the listeners, i.e. where identification was close to chance level (50% /s/ responses and 50% /ʃ/ responses). The confidence scores for the natural stimuli (step 3) and the manipulated stimuli with amplified *relevant* frequencies (steps 1 and 2) were high in both sibilants. For the manipulated stimuli with amplified *irrelevant* frequencies (steps 4 to 7), while the confidence scores for the /ʃ/ stimuli dropped at the first two steps and returned to high scores at the last two steps, the confidence scores for the /s/ stimuli continuously dropped across the steps.

The difference between two sibilants in confidence scores was analyzed in a two-way repeated measures ANOVA using fricative identity (/s/ vs /ʃ/) and step (steps 1 to 7) as factors. There were a main effect of fricative identity, $F(1, 31) = 134.0$, $p < .001$, $\eta^2_p =$

0.812, and an interaction between fricative identity and step, $F(6, 186) = 48.8$, $p < .001$, $\eta^2_p = 0.611$. Planned paired sample t-tests were performed to compare the two sibilants at each step. The Bonferroni-adjusted *p*-value is 0.01 (.05 / 5 = 0.01). The /s/ stimuli elicited significantly lower confidence scores at steps 3 to 7, step 3: $t(31) = -3.593$, $p = 0.001$, step 4: $t(31) = -6.592$, $p < .001$, step 5: $t(31) = -6.319$, $p < .001$, step 6: $t(31) = -10.360$, $p < .001$, and step 7: $t(31) = -9.774$, $p < .001$. One obvious reason would be that the /s/ stimuli remained confusing to the listeners at steps 4 to 7, which in turn would elicit lower confidence scores. However, even if we compare the natural stimuli of the two sibilants, the confidence scores of the natural /s/ stimuli were still significantly lower than those of the natural /ʃ/ stimuli.

Such a difference between two sibilants might have occurred as a task effect. Since the manipulated /s/ stimuli never reached a categorical switch like the /ʃ/ stimuli did, the listeners were constantly exposed to exemplars of /s/ stimuli that they were not able to reliably identify. This might have led to a decrease in overall confidence of perceiving any /s/ stimuli. To test if this is the case, we compared the confidence scores for the /s/ trials in the first two blocks (early blocks) with those for the /s trials in the last two blocks (late blocks) at each step. The confidence scores for the late blocks were slightly lower than those for the early blocks (except step 5). However, in a two-way repeated measures ANOVA using block (early vs. late blocks) and step (steps 1 to 7) as factors, the main effect of block was not significant, $F(1, 30) = 0.827$, $p = 0.370$, $\eta^2_p = 0.027$.

Planned paired sample t-tests were performed to investigate the effect of our stimulus manipulation of amplifying relevant and irrelevant frequencies on the listeners' confidence. Two sets of the results were presented below.

Since the manipulated /ʃ/ stimuli with amplified *irrelevant* frequencies at step 7 were well perceived as /s/, we were interested in whether there was a qualitative difference in the listeners' confidence between the natural /s/ stimuli and the acoustically /s/-like manipulated /ʃ/ stimuli. The confidence scores for the manipulated /ʃ/ stimuli with amplified irrelevant frequencies (step 7) were compared to those for the natural /s/ stimuli (step 3). Although, interestingly, the manipulated /ʃ/ stimuli scored higher than the natural /s/ stimuli (step 3), the difference was not significant, $t(31) = -1.687$, $p = 0.102$, failing to support the idea that the listeners were differentially confident of identifying the two types of stimuli as /s/.

Moreover, since the benefit of amplifying *relevant* frequencies was not shown in the identification data, we were interested in whether there was an effect on the listeners' confidence. For each sibilant, the confidence scores for the manipulated stimuli with amplified relevant frequencies (step 1) were compared to those for the natural stimuli (step 3). The Bonferroni-adjusted *p*-value is 0.025 (.05 / 2 = 0.025). The confidence scores for the /s/ manipulated stimuli were significantly higher than those for the natural /s/ stimuli, $t(31) = -3.430$, $p = 0.002$, but there was no significant difference between the two types of /ʃ/ stimuli, $t(31) = 0.318$, $p = 0.753$. This suggests that although the effect of amplifying relevant frequencies was not observed in the listeners' final identification decisions, the amplification might have benefited the listeners' confidence.

3.2.2. Long vs. short fricative noises

While the identification data showed that the length of fricative noise did not in general influence identification (except some steps of the /s/ stimuli), the confidence score data revealed that the listeners were more confident of identifying long fricative noise stimuli than the short ones. Figure 3.5 shows the mean confidence scores of long and short fricative noise stimuli separately.

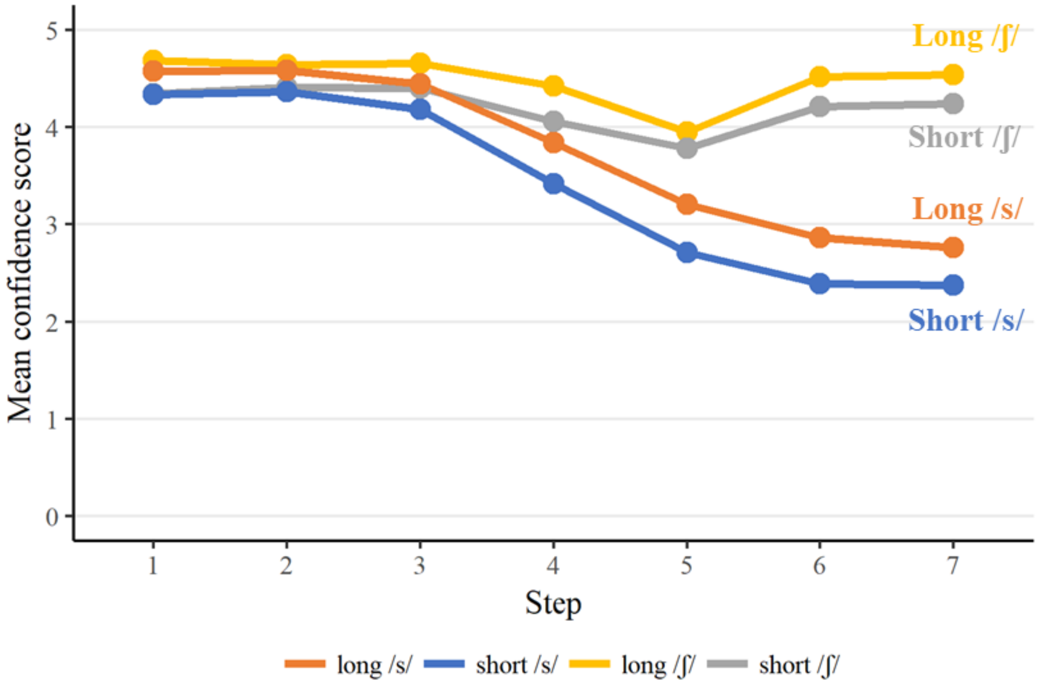

Figure 3.5. Means of confidence scores for long and short isolated fricative noise stimuli.

The difference between the long and short fricative noise stimuli was confirmed by two separate two-way repeated measures ANOVAs (one for each sibilant) using length (long vs. short) and step (steps 1 to 7) as factors. The main effect of length was significant in both tests, the /s/ stimuli: $F(1, 31) = 57.68$, $p < .001$, $\eta^2_p = 0.650$, and the /ʃ/ stimuli: $F(1, 31) = 22.63$, $p < .001$, $\eta^2_p = 0.422$.

3.2.3. Original vs. alternative sibilant vowel contexts

The results for the original and alternative vowel contexts of the VCV stimuli were compared to examine whether vowel contexts influenced the listeners' confidence. Figure 3.6 presents the mean confidence scores for the fricatives in the original and alternative sibilant vowel contexts.

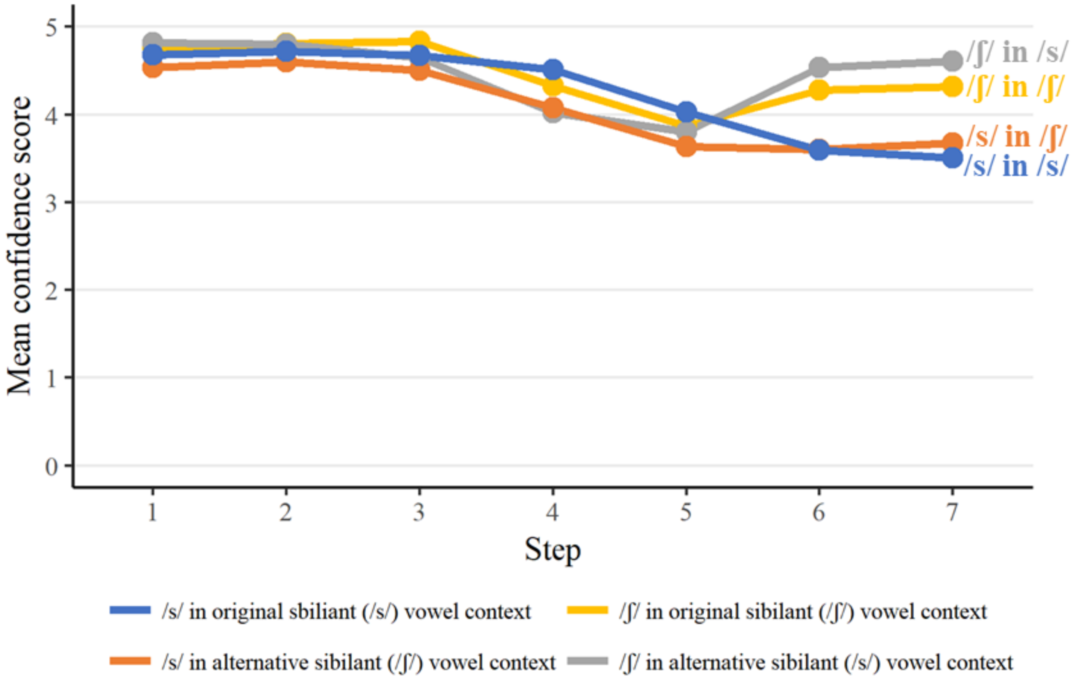

Figure 3.6. Mean confidence scores for fricatives in original and alternative sibilant vowel contexts.

For each sibilant, a set of paired sample t-tests were performed to compare the results for the two types of vowel contexts at each step. Although the confidence scores for the same sibilant tended to be higher for the original sibilant vowel contexts at steps 1 to 5 and for the alternative sibilant vowel contexts at steps 6 and 7, the differences were only significant in some of comparisons.

To examine whether vowel contexts improved the listeners' confidence when the listeners were highly confused by the fricative noise, another set of paired sample t-tests were performed to compare the confidence scores for the /s/ VCV stimuli with those for the /s/ long fricative noise stimuli at steps 6 and 7. The Bonferroni-adjusted p-value is 0.0125 (.05 / 4 = 0.0125). Both types of the /s/ VCV stimuli elicited higher confidence scores than the /s/ long fricative noise stimuli at these two steps, step 6 (long fricative vs original sibilant vowel context): $t(31) = -5.94$, $p < .001$, step 6 (long fricative vs alternative sibilant vowel context): $t(31) = -5.69$, $p < .001$, step 7 (long fricative vs original sibilant vowel context): $t(31) = -6.02$, $p < .001$, and step 7 (long fricative vs alternative sibilant vowel context): $t(31) = -6.28$, $p < .001$.

3.3. Reaction time data

Reaction times (RTs) for each trial were measured from the beginning of the stimulus until the 'Okay' button was clicked to submit the responses. The RTs were first log-transformed. Means and standard deviations were then calculated for each of the three contexts, namely the long fricative noise, short fricative noise, and VCV contexts. Trials with a RT that was longer than the mean plus two standard deviations or shorter than the mean minus two standard deviations of the corresponding context were removed. The percentage of removed trials were 2.68%. Note that the figures below show the raw RTs before log-transformation.

In general, the RTs increased as the stimuli became more confusing to the listeners, i.e. where identification was closer to chance level (50% /s/ responses and 50% /ʃ/ responses). Since the RT data do not reveal any new observations, only the figures are presented here.

3.3.1. *Long vs. short fricative noises*

Figure 3.7 presents mean RTs for the long and short fricative noise stimuli.

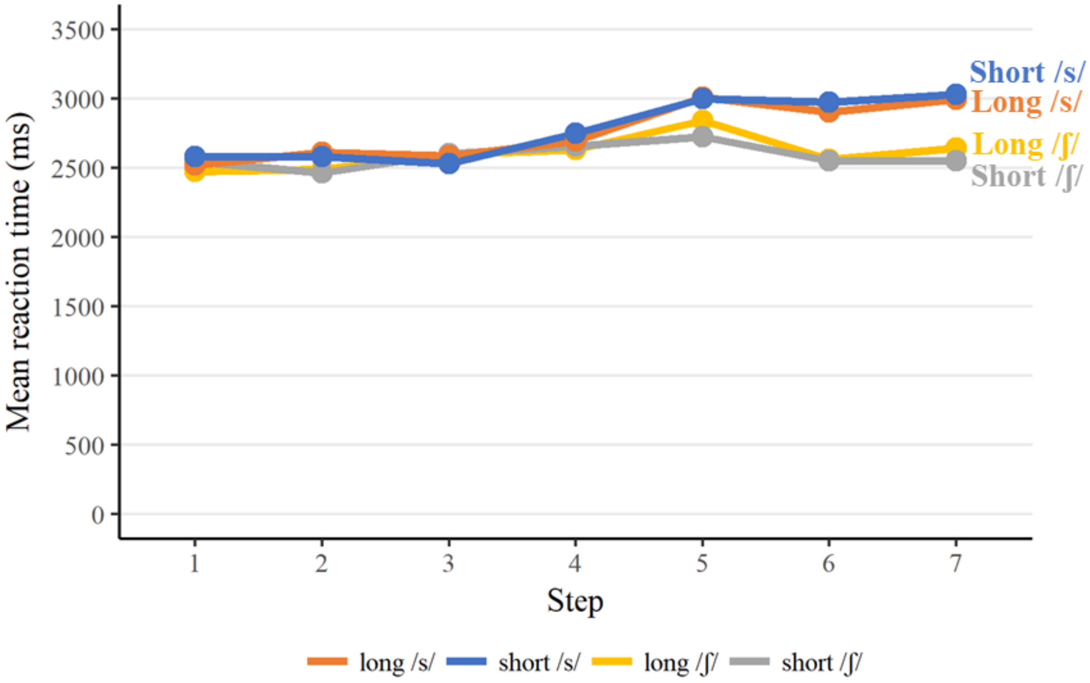

Figure 3.7. Mean reaction times (in ms) for long and short fricative noise stimuli.

3.3.2. *Original vs. alternative sibilant vowel contexts*

Figure 3.8 presents mean RTs for the fricatives in original and alternative sibilant vowel contexts.

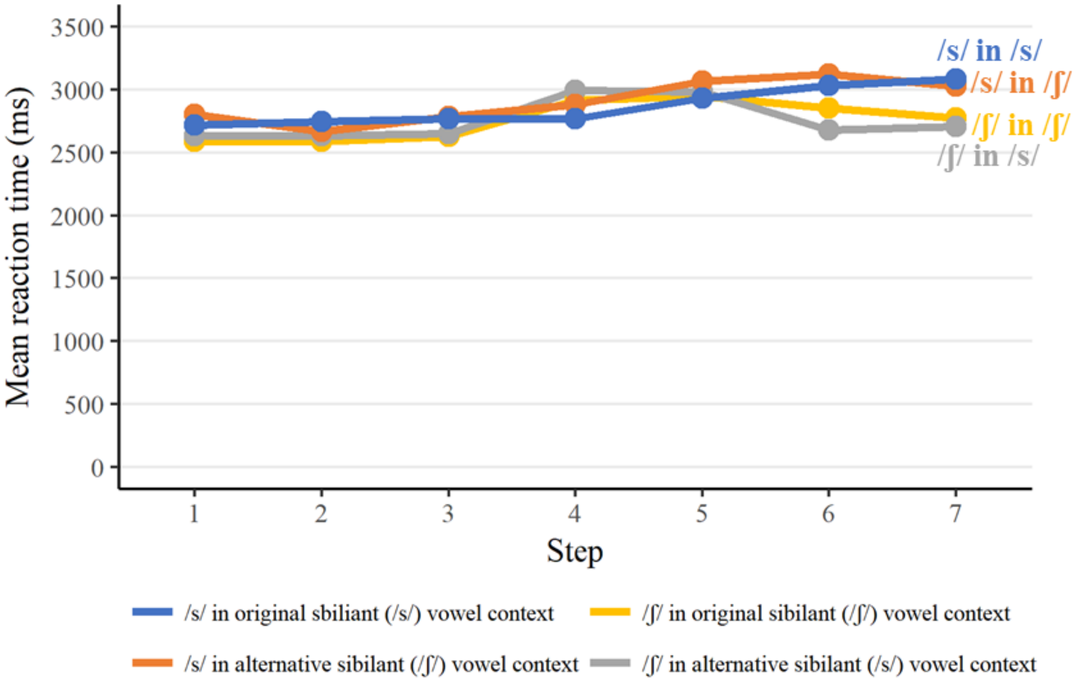

Figure 3.8. Mean reaction times (in ms) for fricatives in original and alternative sibilant vowel contexts.

4. Discussion

4.1. Acoustic plus articulatory information vs. purely acoustic information

Different speech perception theories have been proposed to explain how listeners map acoustic signals onto phonetic categories. Two major groups of theories are acoustic theories and articulatory theories. The primary difference between these two groups of theories is that articulatory theories suggest that there is an intermediate process in which articulatory information is recovered from acoustic signals, and after this recovery, articulatory information is primarily used for determining the perceived phonetic categories.

Our experiment was designed to examine whether articulatory information is, indeed, recovered in the process of phonetic category identification. In our stimulus manipulation, we amplified *irrelevant* acoustic frequencies outside the frequency range of the main characteristic spectral peaks by different amplitudes. After such amplification for both of the /s/ and /ʃ/ prototypes, the overall spectral shape of the manipulated /s/ was *acoustically* almost identical to that of the prototype /ʃ/, and the overall spectral shape of the manipulated /ʃ/ was *acoustically* almost identical to that of the prototype /s/. In other words, from an acoustic point of view, listeners should perceptually confuse the manipulated stimuli with the prototypical ones. Critically, the relevant main characteristic spectral peak frequencies, which signal the front cavity configurations of the speaker, were maintained so that articulatory information for recovering the articulatory configurations is still available to the listeners in the acoustic signals.

Our hypothesis was that if the listeners would recover articulatory information, their identification of the sibilants should primarily be determined by the articulatory configurations of the original sibilants available in the acoustic signals (the /s/ articulatory configurations in all the /s/ stimuli and the /ʃ/ articulatory configurations in all the /s/ stimuli), which means that listeners would identify the acoustically manipulated stimuli as their original sibilants (the /s/ stimuli as /s/ and the /ʃ/ stimuli as /ʃ/). However, if listeners primarily rely on acoustic information (i.e. the overall spectral shape of the stimulus), they would perceptually be confused and identify the acoustically manipulated stimuli as their alternative sibilants (the /s/ stimuli as /ʃ/ and the /ʃ/ stimuli as /s/).

Our manipulation to amplify *irrelevant* frequencies, which made the two sibilants /s/ and /ʃ/ acoustically resemble their alternative sibilants, led to a phonetic categorical switch in the identification of the /ʃ/ stimuli but not in that of the /s/ stimuli. The change in identification was influenced by the degree of amplification of irrelevant frequencies. The effect of weak amplification of irrelevant frequencies by 12 dB was limited. The manipulated s/ and /ʃ/stimuli were still mostly identified as the original sibilants because the overall spectral shapes of these manipulated stimuli were still relatively similar to those of their natural original sibilants. The effect became stronger with 36 dB and 48 dB amplification. The manipulated /ʃ/ stimuli were now only identified as /s/, and roughly 60% of the manipulated /s/ stimuli were identified as /ʃ/. The results show an asymmetry between the /s/ and /ʃ/ stimuli. The listeners made a categorical switch for the /ʃ/ stimuli to completely identify them as /s/. However, there was no such categorical switch for the /s/ stimuli. The identification of the manipulated /s/ stimuli as /ʃ/ was merely above chance level, i.e. 50% of the time the stimulus in question was identified as /s/ and 50% as /ʃ/. We will discuss three possible accounts for this asymmetry below: (1) an acoustic plus articulatory account, (2) a purely acoustic account, and (3) an artifact account.

4.1.1. *An acoustic plus articulatory account*

According to our original research hypothesis, the manipulation of acoustic information while maintaining articulatory information would allow the listeners to recover articulatory configurations in phonetic category identification. In line with this hypothesis, *an acoustic plus articulatory account* to explain the results for the /s/ stimuli would be that listeners, indeed, recovered articulatory information.

In our manipulation, we maintained the main spectral peaks that were primarily characterized by the anterior cavity of the vocal tract configuration. If the hypothesis was right, the listeners would be able to extract this articulatory information and recover the vocal tract configurations which signaled that the speaker had the vocal tract setting of a /s/ place of articulation instead of the /ʃ/ place of articulation (signaled by the size of the front cavity and thus, by the frequency of the main spectral peak). Since /s/ cued by articulatory information would compete with /ʃ/ cued by acoustic information, the identification would become mixed. Our results seem to agree with this hypothesis since the listeners had mixed identification of the acoustically /ʃ/-like /s/ stimuli as both /s/ and /ʃ/. However, if this is the case, such mixed identification should also be observed for the /ʃ/ stimuli. That is, since /ʃ/ cued by articulatory information would compete with /s/ cued by acoustic information, the identification in the /ʃ/ stimuli would mix with both /ʃ/ and /s/. Our results do not show such mixed identification. Rather, the acoustically /s/-like /ʃ/ stimuli was always identified as /s/.

To accommodate the results that the /ʃ/ stimuli did not show mixed identification, we could propose that while the listeners did recover articulatory information, they only employed the articulatory information to identify /ʃ/. That is, we could hypothesize that the weighting between acoustic and articulatory information could be different in the identification of the two sibilants, with articulatory information becoming more important in the /s/ identification than the /ʃ/ identification. This hypothesis seems plausible based on our results: the acoustically /ʃ/-like /s/ stimuli were identified as both /s/ and /ʃ/, and the acoustically /s/-like /ʃ/ stimuli always as /s/. However, if we look more closely at the results for the /ʃ/ stimuli, we could realize that we would still not be able to completely explain the results for the /ʃ/ stimuli. Since the articulatory information of the manipulated /ʃ/ stimuli did not signal a /s/ place of articulation but instead a /ʃ/ place of articulation, there would be no articulatory information supporting an /s/ identification of the manipulated /ʃ/ stimuli. If we follow the hypothesis that listeners rely more on articulatory information when identifying /s/, we could expect that the manipulated /ʃ/ stimuli would not be perceived, or at least not completely, as /s/, given that the manipulated /ʃ/ stimuli lacked the articulatory information signaling a /s/ place of articulation. Our identification data show the opposite: the acoustically /s/-like /ʃ/ stimuli without the articulatory information signaling the /s/ place of articulation were still always perceived as /s/ (which was cued by acoustic information). Therefore, although this hypothesis would be able to explain part of the fact that the /s/ stimuli were identified partially as /s/ when the articulatory information of /s/ place of articulation was available, it would not be able to explain the fact that the /ʃ/ stimuli were still always identified as /s/ when the information of /s/ place of articulation was absent to support /s/ identification and the articulatory information of /ʃ/ place of articulation was there to compete.

One could argue that the listeners additionally draw on the articulatory information of /s/ place of articulation only when it is available (as in the /s/ stimuli) and would primary rely on acoustic information when the articulatory information of /s/ is not available (as in the /ʃ/ stimuli). We could hypothesize the identification process works this way: first, if there is articulatory information specifically signaling a /s/ place of articulation (as in the /s/ stimuli), the listeners would immediately identify the stimuli as /s/; then, if there is no articulatory information of /s/, the listeners would make decisions based on the acoustic information (as in the /ʃ/ stimuli and assuming that they would specifically ignore the available articulatory information of the /ʃ/ place of articulation). An identification process like this would mean that /s/ would still be reliably identified using only acoustic information without the additional help from the articulatory information of /s/ as we observed the /ʃ/ stimuli. This would suggest that acoustic information itself is enough for reliable identification. The question, then, would be why it is necessary for the listeners to occasionally draw on the articulatory information of /s/ (i.e. when identifying the /s/ stimuli) when acoustic information itself is sufficient for reliable identification. More importantly, if this is really how the identification process works, the listeners should have only identified the /s/ stimuli as /s/. This is because the first step would be always to use the articulatory information of /s/ (which is available in the /s/ stimuli), and this would lead to a /s/ identification. It would be unnecessary to take the second step to identify the /s/ stimuli based on acoustic information which would lead to a /ʃ/ identification. Our results for the /s/ stimuli show mixed identification of both /s/ and /ʃ/, suggesting that the listeners did employ acoustic information.

While based on our results, we could definitely attempt to hypothesize that each of the two sibilants was identified with a different weighting between acoustic and articulatory information, it would be difficult to explain how a different weighting could be actually used for identifying the two sibilants, i.e. how this could be accomplished mechanically in the identification process, so that the results predicted by the hypothesis would consistently match our observed results. One fundamental question we need to answer is how the listeners would know which weighting to use before they identify the sibilant. We run into a problem in suggesting that the listeners would use the corresponding weighting based on the specific sibilant they were perceiving to identify that specific sibilant *before* they actually knew the identity of that specific sibilant. After all, the listeners would not know the sibilant they were perceiving was /s/ beforehand and thereby, choosing to weight articulatory information more heavily. Similarly, they would not know the sibilant they were perceiving was /ʃ/ beforehand and thereby, choosing to weight articulatory information less heavily.

It could also be argued that the comparisons between the natural stimuli and stimuli with amplified *relevant* frequencies seem to support the hypothesis that the listeners have relied more on articulatory information specifically when identifying /s/. Although these two types of stimuli are equally well identified as /s/ in the identification data, the confidence score results revealed a subtle difference between the natural /s/ stimuli and the manipulated /s/ stimuli with amplified *relevant* frequencies: listeners were slightly more confident in identifying the manipulated /s/ stimuli with amplified relevant frequencies (where the main spectral peaks signaling the articulatory configurations occurs). It seems that this could be used as evidence to show that articulatory information was recovered and employed. However, if this increased confidence was actually an effect of articulatory information, we would expect a difference between the natural /s/ stimuli and the acoustically /s/-like /ʃ/ stimuli since the natural /s/ stimuli provided the additional articulatory information of a /s/ place of articulatory and the acoustically /s/-like /ʃ/ stimuli did not. Our test shows that the confidence scores for these two types of stimuli were not significantly different (although, interestingly, the scores were higher for the acoustically /s/-like /ʃ/ stimuli than the natural /s/ stimuli). Therefore, we would argue that the confidence data do not provide strong evidence to support the hypothesis that the /s/ identification was helped by articulatory information.

In short, if we would like to use the articulatory hypothesis to explain our data, we will need to add these assumptions to explain why articulatory information influences specifically the /s/ identification but not the /ʃ/ identification, and even if we have added a set of assumptions, it would still be difficult to explain why articulatory information influences the /s/ identification specifically of the /s/ stimuli, but not in the /ʃ/ stimuli. More importantly, it would be difficult to explain how listeners could, indeed, use a different weighting between acoustic and articulatory information to identify a sibilant before they identify the sibilant.

4.1.2. A purely acoustic account

Alternatively, our results can be explained by a purely acoustic account without adding the set of assumptions required by the acoustic plus articulatory account. A purely acoustic account would suggest that an asymmetry of identification between two sibilants was only a result of acoustic differences. We compared the spectra of the natural and manipulated stimuli of the same sibilant to identify acoustic differences which may result in an asymmetry of the identification. Our observations match well with our identification data.

Figure 4.1 shows the spectra of the natural stimuli (step 3) and the stimuli with amplified *irrelevant frequencies* (step 7) for the two sibilants.

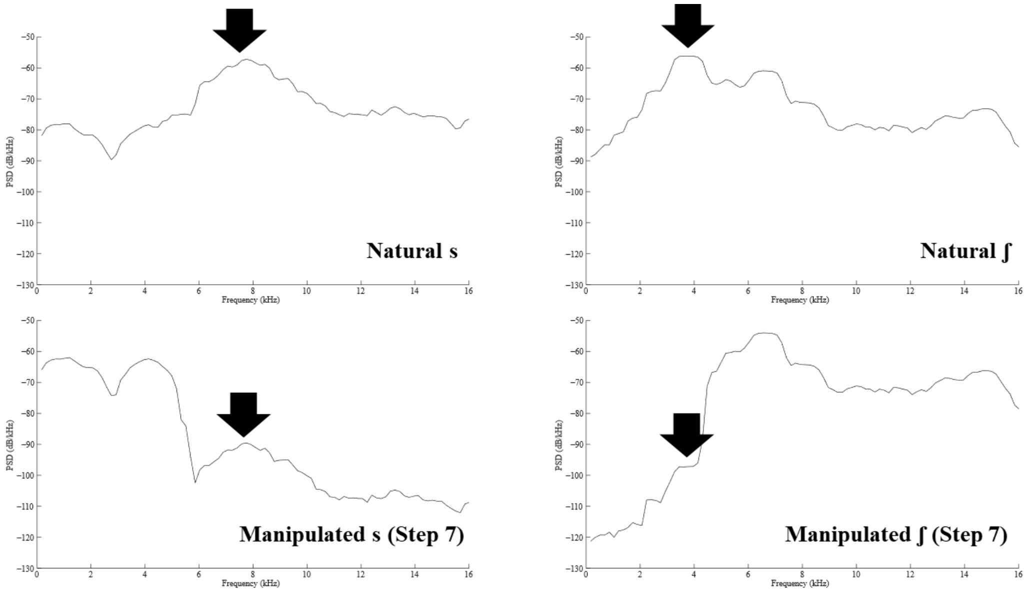

Figure 4.1. Spectra of natural and manipulated stimuli with amplified irrelevant frequencies (step 7).

Note: The arrows point to the frequencies in which the main spectral peaks occur in the natural stimuli

The black arrows point to the spectral peaks which are of interest here. First, if we look at the natural /s/ spectrum, the main spectral peak occurs at approximately 8 kHz. After amplifying the *irrelevant frequencies* below 5 kHz, this peak at around 8 kHz of the manipulated /s/ spectrum is not as high as that of the natural /s/ spectrum., but critically, there is still an amount of energy in this specific frequency range around 8 kHz, (even though there are also higher peaks in the lower frequency range). Then, if we look at the natural /ʃ/ spectrum, the main spectral peak occurs at approximately 4 kHz. After amplifying the *irrelevant frequencies* above 5 kHz, the peak at 4 kHz of the manipulated /ʃ/ spectrum has become only a small peak which has been merged into the large amount of energy at around 6 to 7 kHz.

These observations match well with our identification data. If we compare the two manipulated stimuli, the peak at around 8 kHz of the manipulated /s/ is still relatively high while the peak at around 5 kHz of the manipulated /s/ is only part of a larger peak. Although the peak at 8 kHz in the manipulated /s/ would obviously not be as perceptually salient as the peaks in the lower frequency range, this peak at 8 kHz of the manipulated /s/ should still be more perceptually salient than the peak at 4 kHz of the manipulated /ʃ/. Following our observations, we could predict the possible difference in the listeners' identification. For the manipulated /s/, the largest spectral peaks at the lower frequencies would be perceptually most salient and largely favor a /ʃ/ identification, but at the same time, the spectral peak at around 8 kHz would still be perceptually relatively salient and partially favor a /s/ identification. For the manipulated /ʃ/, the largest spectral peak at the higher frequencies would be perceptually most salient and favor a /s/ identification, and the spectral peak at 4 kHz is not salient enough to favor an /ʃ/ identification. This prediction matches our identification data well: our listeners identified a larger proportion of the manipulated /s/ as /ʃ/ and a smaller proportion of them as /s/, and all the manipulated /ʃ/ as /s/. Our comparison shows that spectral peaks of different degrees of salience, depending on their amplitudes relative to the overall energy distributions, result in different identification results.

It should be noted that although we are explaining our data with the main spectral peaks characterized by the front cavity in the vocal tract configurations, we are only discussing these spectral peaks as the physical acoustic characteristics that can be observed in our stimuli. That is, it was the physical acoustic differences of energy distribution between the natural and manipulated stimuli of the two sibilants /s/ and /ʃ/ in the spectra, which include the spectral peaks, that led to an asymmetry of identification. Critically, it is not the possible recovery of articulatory information from these spectral peaks, as hypothesized by an acoustic plus articulatory account, that has led to an asymmetry of identification between two sibilants.

The results that the remaining small spectral peak at around 8 kHz of the manipulated /s/ stimuli was enough to induce a /s/ identification can be supplemented by other speech perception studies of the /s/ sibilant. These studies often show an asymmetry between /s/ and other phonetic categories. In studies on perceptual learning, listeners were trained to adjust their phonetic category boundary to accept less typical tokens of a phonetic category. Several studies showed that the training effect on /s/ was less significant than on the other sibilant /ʃ/ (e.g., Kraljic & Samuel, 2005) and also other phonetic categories (e.g., Burchfield, Luk, Antoniou, & Cutler, 2017; Norris, McQueen, & Cutler, 2003; Sjerps & McQueen, 2010). This can have two different reasons. The first reason could be that the phonetic category boundary of /s/ is more solid than that of other phonetic categories. That is, /s/ has a boundary that is harder to shift and has a higher threshold for a token to be identified as its member. As a result, it would be more difficult to produce the same degree of effect on /s/ than other categories. This explanation does not seem compatible with our data. In fact, the manipulated /ʃ/ tokens that were acoustically extremely similar to /s/ were well identified as /s/.

The second reason for the asymmetry could be that /s/ has a more flexible boundary and accepts greater acoustic variability in its physical tokens. If this was the case, the manipulated /s/ tokens would still be identified as /s/ even though the overall spectral shapes of the manipulated /s/ stimuli deviated from a natural /s/ but still preserved some perceptual salient cues of /s/. This explanation seems more plausible since roughly 40% of manipulated /s/ tokens were still identified as /s/ even though their overall spectral shapes were much more similar to a natural /ʃ/ rather than a natural /s/, and merely the small spectral peak at 8 kHz was able to favor an /s/ identification. This also agrees with the results that the manipulated /ʃ/ was completely switched to be identified as /s/ because /s/ allows greater acoustic variability. However, the reasons for such a category-specific difference remains unknown.

Comparing the acoustic plus articulatory and purely acoustic accounts, we argue that the purely acoustic account is a more plausible explanation for the asymmetry of identification between the /s/ and /ʃ/ stimuli. If we follow an acoustic plus articulatory account, we would need to add the assumption that listeners weight articulatory information more heavily when identifying /s/, and even if we have added the assumption, we would not be able to explain why listeners would rely more heavily on articulatory information in some circumstances but not the others. More importantly, it would be extremely difficult to explain how listeners could actually employ a different weighting between acoustic and articulatory information to identify a sibilant before knowing the sibilant. Using a purely acoustic account, we can explain our data without any assumptions. The asymmetry of identification can be explained well by the actual acoustic differences between the natural and manipulated stimuli of the same sibilants. In addition, if we accept the acoustic and articulatory account, we need to assume that listeners would use two different ways of weighting acoustic and articulatory information (i.e. articulatory information is more important when identifying /s/ than /ʃ/) when they are perceiving the same type of speech sounds, i.e. sibilants. The purely acoustic account provides an advantage here because the listeners' processing of the two sibilants would be fundamentally identical, and the asymmetry of identification is only a result of acoustic differences.

The original research hypothesis we wanted to test was whether listeners will recover articulatory information in the process of identifying phonetic categories. If we follow *a purely acoustic account*, our results show that it is acoustic information rather than articulatory information which has influenced the listeners' identification. If we follow *an acoustic plus articulatory account*, our results suggest that articulatory information was possibly recovered. However, the use and influence of articulatory information seemed to be restricted to a specific context. That is, articulatory information only influenced how the listeners identified /s/ in the /s/ stimuli, and acoustic information was more dominant than articulatory information (60% acoustically cued /ʃ/ and only 40% articulatorily cued /s/). Our results do not provide conclusive evidence to reject the possibility of articulatory information recovery. However, even if we assume that listeners will, indeed, recover articulatory information, our results do show that, at least for identifying the two English sibilants /s/ and /ʃ/, acoustic information is more dominant in the identification process.

4.1.3. An artifact account

The third possible explanation for the asymmetry of identification between the two sibilants is related to the possible artifact effect due to our stimulus manipulation. We divided the frequency range at 5 kHz to create one range below 5 kHz and the other above 5 kHz. When the frequency range above 5 kHz was amplified, the amplification was not only applied to the frequencies between 5 to 10 kHz but also the frequencies above 10 kHz (as highlighted by the grey box in Figure 4.2). As a result, the amplified frequency range above 5 kHz was broader than that below 5 kHz.

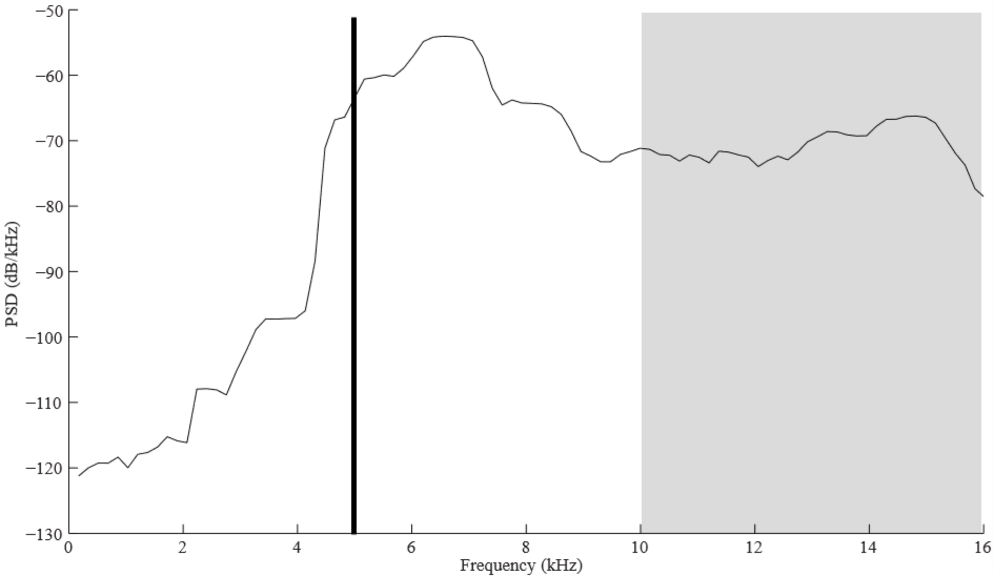

Figure 4.2. Frequencies additionally amplified in the frequency range above 5 kHz (in the grey box).

When for the /ʃ/ stimuli the irrelevant frequencies above 5 kHz were amplified, there was an amplification of energy of a broader frequency range. Meanwhile, for the /s/ stimuli the irrelevant frequencies below 5 kHz were amplified, but the energy of a much narrower frequency range was amplified here. The manipulation in the /ʃ/ stimuli may in turn have favored /s/ identification more than the manipulation in the /s/ stimuli favoring /ʃ/ identification.

However, since the human auditory system is less sensitive to differences at the higher frequency range than the lower frequency range in perception, we argue that the effect of amplifying the frequency range above 10 kHz on the listeners' identification would be very limited. An additional experiment using stimuli in which frequencies over 10 kHz were filtered out can eliminate the possibility that the asymmetry of identification is an artifact effect of stimulus manipulation.

4.2. Effect of vowel context

Previous studies have shown that fricative noise is the primary acoustic cue for English listeners to determine sibilant identities, and the influence of vocalic formant transition is restricted. We are interested in whether the reliance on vocalic formant transition increases when the fricative noise itself is confusing and therefore, unreliable for identification. The natural and manipulated fricative noises were cross-spliced into two types of vowel contexts, one favoring the original sibilant and another favoring the alterative sibilant.

Our results show that the effect of vowel context on identification was stronger when the fricative noise was confusing to the listeners. There were no noticeable differences between the two vowel contexts for the natural stimuli and manipulated stimuli with amplified *relevant* frequencies probably because a ceiling effect was observed in both contexts. The difference occurred for the manipulated stimuli with amplified *irrelevant* frequencies and varied between the two sibilants. For the /ʃ/ stimuli, the effect of vowel context was strongest at step 5 where the isolated /ʃ/ fricative noise stimuli were most confusing to the listeners, i.e. when the identification was close to chance level (50% /s/ responses and 50% /ʃ/ responses) among all the steps. When the fricative noise itself is reliable enough for identification at steps 6 and 7, the influence of vowel context is minimized, which also means that the original vowel context was not strong enough to make the identification resist the phonetic categorical switch to /s/. Similar results were observed for the /s/ stimuli. The effect of vowel context became the strongest at steps 6 and 7 where the isolated /ʃ/ fricative noise stimuli were the most confusing to the listeners. Since the /s/ stimuli were in general more confusing to the listeners, i.e. there were three steps at which the identification was close to chance level (compared to only one step for the /ʃ/ stimuli), the effect of vowel context was stronger for the /s/ stimuli than the /ʃ/ stimuli.

The effect of vowel context on the listeners' confidence was not as obvious as observed in the identification data. Although the listeners tended to be more confident for the stimuli with original sibilant vowel contexts at steps 1 to 5 (where the fricative noise itself was more often identified as the original sibilant) and for the stimuli with alternative sibilant vowel contexts at step 6 and 7 (where the fricative noise itself was more often identified as the alternative sibilant), the difference between two types of vowel contexts was not consistently significant. However, when we compared the /s/ VCV stimuli with the isolated /s/ long fricative noise stimuli, the vowel context, either the original or alternative sibilant vowel contexts, significantly improved the listeners' confidence at steps 6 and 7 where the isolated /s/ fricative noise stimuli were most confusing to the listeners. This again shows that the effect of vowel context became stronger when the fricative noise itself was more confusing.

Our results agree with previous studies that for English sibilant identification the weighting of frication noise is much higher than vocalic formant transition. If the fricative noise itself is reliable for identification, as in the natural stimuli and manipulated stimuli with amplified relevant frequencies, the listeners largely rely on fricative noise. However, when fricative noise was unreliable or insufficient for identification, the listeners would draw on the additional available information of vocalic formant transition to make their decisions.

4.3. Length of fricative noise

To test the possible effect of fricative noise duration on perception as shown in Jongman (1989), we created two lengths of the isolated fricative noise stimuli, 336 ms for the long stimuli and 150 ms for the short stimuli. Our results have shown some interesting differences between the long and short fricative noise stimuli. For identification, the only significant difference occurred for one step of the /s/ stimuli with amplified *irrelevant* frequencies. The short /s/ stimulus was more often identified as the alternative sibilant /ʃ/ than the long /s/ stimulus, suggesting that identification switched more towards the alternative sibilant /ʃ/ when shorter /ʃ/-like /s/ tokens were presented. The listeners were also generally less confident of identifying the long fricative noise stimuli than the short fricative noise stimuli even though the short fricative noise stimuli (150 ms) were already considerably longer than fricatives in normal continuous speech (between 100 ms and 200 ms).

4.4. Speech vs. non-speech modes of perception

It could be argued that listeners were using a non-speech mode to perceive the sibilants given that sibilants are frication noise. We argue that this is unlikely to be the case based on our experimental design for three reasons. First, instead of using a discrimination task (e.g., an ABX task where it is very easy to switch into a non-speech mode), we employed an identification task in which listeners were always required to compare a given stimulus to their existing mental representations of the two sibilants, which suggests that the decisions were most likely to be linguistic in nature. Secondly, the experiment was contextualized to be a speech experiment in which listeners were clearly instructed to identify the stimuli as two sibilants. Thirdly, one half of all stimuli were syllabic VCV stimuli. Identifying the sibilants in a VCV structure is a highly language-oriented task, and the stimuli of this CVC condition were randomly mixed with the isolated fricative noise stimuli. Such a design created a task environment in which listeners were less likely to switch their perceptual strategy from a speech mode to a non-speech mode, or alter their way of identification to differentiate between stimuli with a VCV structure and isolated fricative noise.

4.5. Future studies

This is an exploratory study on sibilant perception which examines how the intrinsic spectral structure of fricative noise influences perception. There are two aspects that can be further studied.

Firstly, while our identification and confidence score data showed that the natural /s/ and /s/-like manipulated /ʃ/ were similar, it would also be interesting to compare the two stimuli in terms of similarity and typicality. Listeners can be presented a natural stimulus and a manipulated stimulus, and be asked to rate the similarity between the two stimuli, or an individual stimulus and be asked to rate the goodness of the stimulus. This would provide insights into the possible perceptual differences between a natural sibilant and its manipulated counterpart.

Secondly, our results can be explained by both *an acoustic plus articulatory account* and *a purely acoustic account* (although we argue that the purely acoustic account would be more plausible). The reason why our Canadian English listeners show a limited use of articulatory information (for an acoustic plus articulatory account) or no use of articulatory information (for a purely acoustic account) could be that the two-way contrast in English is already highly distinguishable in perception which makes acoustic information sufficient and articulatory information highly redundant. In other words, there is not really a need to rely on articulatory information to guarantee a robust perceptual outcome. It would be interesting to test languages which have a three-way contrast with very similar acoustic properties to see if the increase of perceptual difficulty may encourage the employment of articulatory information.

4.6. Conclusion

Articulatory theories hypothesize that listeners recover articulatory information when identifying phonetic categories. To test this hypothesis, we altered the acoustic information while keeping the articulatory information in the two sibilants /s/ and /ʃ/ to see whether the listeners would switch identification. We manipulated the acoustic information by making the overall spectral shapes of natural /s/ and /ʃ/ like those of their alternative sibilants, and critically, we preserved articulatory information by keeping the main spectral speaks characterized by the location of the anterior cavity of the vocal tract configuration. The manipulated sibilants were articulatorily cued as the original sibilant but acoustically cued as the alternative sibilants. In our identification experiment, the listeners identified acoustically /s/-like /ʃ/ completely as the alternative sibilant /s/, but the acoustically /ʃ/-like

/s/ as 60% of the alternative sibilant /ʃ/ and 40% of the original sibilant /s/. There was a categorical switch in the /ʃ/ stimuli but not the /s/ stimuli.

This asymmetry of identification between two sibilants can be explained by two accounts: *an acoustic plus articulatory account* and *a purely acoustic account*. An acoustic plus articulatory account would be that the listeners employed articulatory information or weighted articulatory information more heavily only when identifying /s/ but not /ʃ/. That is, the listeners identified some of the /ʃ/-like /s/ as /s/ because the articulatory information signaled /s/ placed of articulation. A purely acoustic account would be that the asymmetry was only a result of acoustic differences between the natural and manipulated stimuli of the same sibilant. Specifically, the spectral peak favoring /s/ identification was still relatively salient in the manipulated /s /stimuli and triggered the listeners to identify the manipulated /s/ stimuli as /s/, as compared to the less salient spectral peak favoring a /ʃ/ identification in the manipulated /ʃ/ stimuli. Comparing the *acoustic plus articulatory* and *purely acoustic* accounts, we argue that the purely acoustic account is a more plausible explanation for the observed asymmetry for two reasons. First, for an acoustic plus articulatory account, even if we could hypothesize that /s/ is identified with the help of articulatory information, it would be difficult to explain how it could possibly happen in the identification process. Second, a purely acoustic account allows us to consistently explain our data and to avoid adding a set of assumptions as required by an acoustic plus articulatory account.

Our results cannot be used as evidence to reject the possibility that listeners recover articulatory information from acoustic signals to extract the vocal tract configurations when perceiving phonetic categories. However, even if we follow *the acoustic plus articulatory* *account* and assume that articulatory information is recovered, our results do suggest that the perceptual system largely relies on acoustic information or weights acoustic information more when performing the mapping, at least for the two Canadian English phonetic categories tested here. Our results, thus, provide evidence to support acoustic speech perception theories but not articulatory speech perception theories like the *Motor Theory* and the *Direct Realist Theory*. Since the importance of acoustic information is likely to be higher than that of articulatory information, it would be more plausible to regard mental representations as representations of phonetic categories that are acoustic in nature rather than ultimately articulatory.

References

Blumstein, S. E. (1986). On acoustic invariance in speech processes. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 178–193). Hillsdale, NJ: Lawrence Erlbaum.

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America*, *66*(4), 1001–1017.

Boersma, P., &Weenink, D. (2018). Praat: doing phonetics by computer [Computer program]. Retrieved from http://www.praat.org/

Burchfield, L. A., Luk, S.-H. K., Antoniou, M., & Cutler, A. (2017). Lexically Guided Perceptual Learning in Mandarin Chinese. *Interspeech 2017*, 576–580.

Delattre, P. C., Berman, A. M. L., & Cooper, F. S. (1962). Formant transitions and loci as acoustic correlations of place of articulation in American fricatives. *Studia Linguistica*, *16*(1–2), 104–122.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specically modulates the excitability of tongue muscles: a TMS study. *European Journal of Neuroscience*, *15*, 399–402.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *The Journal of the Acoustical Society of America*, *84*(1), 115–123.

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

Galantucci, B., Fowler, C., & Turvey, M. T. (2006). “The motor theory of speech perception reviewed”: Erratum. *Psychonomic Bulletin & Review*, *13*(4), 742.

Harris, S. (1957). Cues for the Discrimination of American English, (1947), 1–7.

Heinz, J. M., & Stevens, K. N. (1961). On the Properties of Voiceless Fricative Consonants. *The Journal of the Acoustical Society of America*, *33*(5), 589–596.

Hillenbrand, J. M., Gayvert, R. T., & Clark, M. J. (2015). Phonetics exercises using the Alvin Experiment-Control software. *Journal of Speech, Language, and Hearing Research*.

Jongman, A. (1989). Duration of frication noise required for identification of English fricatives. *The Journal of the Acoustical Society of America*, *85*(4), 1718–1725.

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

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Oxford & Cambridge, MA: Blackwell Publishers.

LaRiviere, C., Winitz, H., & Herriman, E. (1975). The Distribution of Perceptual Cues in English Prevocalic Fricatives. *Journal of Speech Language and Hearing Research*, *18*(4), 613.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

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

Lindblom, B. (1988). Phonetic invariance and the adaptive nature of speech. In B. A. G. Elsendoom & H. Bouma (Eds.), *Working Models of Human Perception* (pp. 139–173). London, UK: Academic Press.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modeling* (pp. 403–439). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Magen, H. S. (1997). The extent of vowel to vowel coarticulation in English. *Journal of Phonetics*, *25*(2), 187–205.

McGuire, G. (2007). English listeners' perception of Polish alveopalatal and retroflex voiceless sibilants: A pilot study. *UC Berkeley Phonology Lab Annual Report*.

Nittrouer, S. (1992). Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries. *Journal of Phonetics*, *20*, 351–382.

Nittrouer, S. (2002). Learning to perceive speech: How fricative perception changes, and how it stays the same. *The Journal of the Acoustical Society of America*, *112*(2), 711–719.

Nittrouer, S., & Miller, M. E. (1997). Predicting developmental shifts in perceptual weighting schemes. *The Journal of the Acoustical Society of America*, *101*(4), 2253–

2266.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204–238.

Nowak, P. M. (2006). The role of vowel transitions and frication noise in the perception of Polish sibilants. *Journal of Phonetics, 34*(2), 139–152.

O'Shaughnessy, D. (1996). Critique: Speech perception: Acoustic or articulatory? *Journal of the Acoustical Society of America, 99*(July 1995), 1726.

Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *The Journal of the Acoustical Society of America*.

Perrier, P. (2005). Control and representations in speech production. *ZAS Papers in Lingustics, 40*, 109–132.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.

Rogalsky, C., Love, T., Driscoll, D., Anderson, S. W., & Hickok, G. (2011). Are mirror neurons the basis of speech perception? Evidence from five cases with damage to the purported human mirror system. *Neurocase*, *17*(2), 178–187.

Sjerps, M. J., & McQueen, J. M. (2010). The Bounds on Flexibility in Speech Perception. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(1), 195–211.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, *64*(5), 1358–1368.

Whalen, D. H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception & Psychophysics, 35*(1), 49–64.

Whalen, D. H. (1991). Perception of the English /s/-/integral of/ distinction relies on fricative noises and transitions, not on brief spectral slices. *The Journal of the Acoustical Society of America*, *90*(October), 1776–1785.

Zygis, M., & Padgett, J. (2010). A perceptual study of Polish fricatives, and its implications for historical sound change. *Journal of Phonetics*, *38*(2), 207–226.

Zygis, M., Pape, D., & Jesus, L. M. T. (2012). (Non-)retroflex Slavic affricates and their motivation: Evidence from Czech and Polish. *Journal of the International Phonetic Association*, *42*(3), 281–329.