EFFECTS OF ACOUSTIC MANIPULATION ON THE PERCEPTUAL STABILITY OF CANADIAN ENGLISH ISOLATED SIBILANTS

San-Hei Kenny Luk, Daniel Pape

McMaster University, Hamilton, Canada
luksanhei@gmail.com, paped@mcmaster.ca

ABSTRACT

Articulatory and motor perception theories hypothesize that listeners recover underlying articulatory information when mapping acoustic speech signals to phonetic categories. To test certain paradigms of this hypothesis for Canadian English sibilants, we acoustically manipulated /s ʃ/ such that the manipulated phonemes were supposedly *articulatorily* cued as the original sibilants but *acoustically* cued as the alternative sibilants (i.e. /s/ as /ʃ/ and /ʃ/ as /s/). Results of a forced-choice perception experiment showed that switching phoneme identification strongly depends on the underlying sibilant category: Listeners identified acoustically /s/-like /ʃ/ completely as the alternative sibilant /s/, however acoustically /ʃ/-like /s/ only at chance level as the alternative sibilant /ʃ/. We conclude that, although acoustic information dominates the identification process for the tested sibilants, additionally articulatory information seems to be recovered, however in dependence of the underlying phonetic category (i.e. in this case restricted to alveolars). Different explanations are proposed for the observed imbalance.

Keywords: fricative perception, sibilant acoustics, articulation, Canadian English, sibilants

1. INTRODUCTION

1.1. Acoustic versus articulatory speech perception theories

One of the crucial tasks in speech perception is the mapping of the acoustic signals onto the underlying phonetic categories in listeners' minds. Different speech perception theories have been proposed to explain how listeners accomplish this mapping. Classically, the two major groups of theories are *acoustic theories* and *articulatory theories* (for recent review, see [17] and [5]). The major difference between acoustic and articulatory theories lies in whether mental representations of phonetic categories are acoustic or articulatory in nature. For acoustic theories, such as e.g. *Acoustic Invariance Theory* [1, 2, 3] and the *Adaptive Variability Theory* [11, 12], acoustic information on its own is sufficient for successful mapping. For articulatory theories, such as *Motor Theory* [13, 14] and *Direct-Realist Theory* [4], there are intermediate processes that recover articulatory information from acoustic signals, thus 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 phonetic category identification. While some researchers believe that recovering articulatory information is unnecessary (e.g., [16]), others support an articulatory account of speech perception (e.g., [5]). Our study aims to provide additional empirical evidence to fuel this debate by examining the perception of manipulated sibilant fricatives.

1.2. Acoustics of sibilants

Previous acoustic studies have identified acoustic cues that can be used to distinguish sibilants produced with different places of articulation (for an overview, see [8]). The two main acoustic measurements, restricted to the fricative noise, are the frequency location of the highest spectral peak(s) and the center of gravity (COG). The highest spectral peak location is a specific range of frequencies at which the highest amplitudes occur, i.e. the frequencies where the sibilants' acoustic energy is strongest. This frequency region is almost exclusively determined by the size of the anterior cavity (or front cavity) that defines the place of articulation for a given sibilant. In contrast, the COG measurement is the mean of the overall distribution of energy spread over *all* frequencies.

If we compare the type of information each of the two measures represents, the location of the highest spectral peak is more related to the articulatory (place of articulation) information since it is primarily determined by the size of anterior cavity and thus by the most important articulatory setting (i.e. place of articulation). COG is also partially determined by the size of the anterior cavity, but, most importantly, also heavily influenced by acoustic energy occurring *outside* the frequency range of the highest spectral peak. As an example, increasing the (laryngeal) source strength strongly

excites energy in frequency regions outside the main spectral peak, mainly in higher frequency regions, and thus significantly influences COG computation. On the other hand, COG measurements give an overall representation of the energy distribution over all frequencies and thus defined by the sibilant's overall spectral shape. If we aim to relate these two acoustic measurements to the perceptual domain, then the highest spectral peak location(s) provides a measure for articulatory information (i.e. front cavity location and thus place of articulation), and COG a measure of the overall acoustic representation of the produced sibilant.

1.3. Perception of sibilants

We are interested in how sibilant identification may be influenced by manipulations of acoustic information available in the spectral distribution of fricative noise. We acoustically altered /s/ and /ʃ/ so that the acoustic shape of the manipulated sibilants becomes very similar to their sibilant alternatives (i.e. the alternative sibilant of the manipulated /s/ resembles a prototypical /ʃ/, and the alternative sibilant of the manipulated /ʃ/ a prototypical /s/), while keeping the underlying articulatory information (based on the front cavity information) constant. This should create an auditory confusion based on available acoustic information, namely that the manipulated /s/ acoustically resembles a prototypical /ʃ/, and the manipulated /ʃ/ a prototypical /s/. Our hypothesis is that if listeners, indeed, only rely on acoustic information, they should completely switch identification (from /s/ to /ʃ/ and /ʃ/ to /s/) because the manipulated sibilants were acoustically highly similar (and thus confusable) to their alternative sibilants. However, if listeners are able to rely on the recovery of articulatory information of the (acoustically unaltered) primary spectral peaks, such a switch would not occur because the underlying original sibilants, but not the acoustically manipulated part, would be responsible for the building of the perceptual construct. Previous research suggests that English listeners mainly use fricative noise as the primary acoustic cue (e.g., [6] [7] [10]). Therefore, and in order to avoid any perceptual influence of preceding and/or following vowel formants, we used isolated fricatives for the perception experiment.

2. METHODS

2.1. Stimuli

The stimuli were recorded by a female native speaker of Canadian English (GTA region, Southern Ontario, Canada). Pseudowords in a VCV structure, [asa] and [aʃa], were recorded, and the most prototypical items (based on native listener perception) were selected for further processing. The pure fricative noises of /s/ and /ʃ/ were extracted from their vocalic contexts and normalized in length. These normalized tokens were then manipulated with the goal to alter acoustic information while keeping underlying articulatory information as identical as possible. The overall spectrum of a manipulated sibilant was manipulated to become extremely similar to the spectrum of their prototypical alternative sibilant (i.e. the sibilant produced at the alternative place of articulation; prototypical /ʃ/ for manipulated /s/, and prototypical /s/ for manipulated /ʃ/), while the spectral peak location representing the anterior cavity in the vocal tract configuration remained identical. Most importantly, the spectra of manipulated /s/ would be identical to the prototypical /ʃ/, and spectra of manipulated /ʃ/ would be identical to a prototypical /s/. Thus, only based on spectral (acoustic) information, a prototypical sibilant and its manipulated alternative sibilant were made to be as identical and thus perceptually confusable as possible, whereas the underlying articulatory information of the two stimuli was still contrary and thus possibly perceptually contrastive.

The frequency range of each sibilant was divided into two sub ranges divided at the frequency of 5 kHz, in our acoustic data (and based on [18] the frequency midpoint between the main spectral peaks of /s/ (around 7 kHz) and /ʃ/ (around 3 kHz)). We defined this 5 kHz point as the division between *relevant* and *irrelevant frequencies* with respect to the articulatory front cavity resonances. *Relevant frequencies* represent the frequency ranges where the prototypical highest spectral peaks appear, and *irrelevant frequencies* represent all other frequency ranges. For the /s/ phoneme, *relevant frequencies* are above 5 kHz and *irrelevant frequencies* below 5 kHz, all seen with respect to the front cavity resonance region. However, these regions are switched for /ʃ/: *relevant frequencies* are here below 5kHz and *irrelevant frequencies* above 5 kHz. Note that the frequency range above 5 kHz also includes frequencies above 10 kHz (up to *Nyquist* frequency).

For each sibilant, we created an acoustic continuum with seven steps by amplifying either *relevant* or *irrelevant* frequencies by different amplitudes. Frequencies to be amplified were filtered with the *Waves Linear Phase Equalizer Filter* using an professional audio engineering high-accuracy shelving filter (V-Slope high-shelf and V-Slope low-shelf). All acoustic stimuli were then RMS amplitude normalized using European Broadcasting Union's (EBU) loudness standards to

ensure that perceived loudness was identical across all manipulated stimuli. Table 1 summarizes the manipulations for each step.

Table 1: Manipulation summary (stimulus steps)

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

Figure 1 compares the spectra of the prototypical sibilants (left panels) with the endpoint of their manipulated spectra (step 7 in table 1). As can be seen, the overall spectral energy distribution of prototype /s/ (lower left panel) and manipulated /ʃ/ stimulus (upper right panel) are very similar. Likewise, the overall spectral energy distribution of prototype /ʃ/ and manipulated /s/ are very similar.

In sum, the overall spectral shapes of the manipulated stimuli closely resemble those of their natural alternative sibilants and thus should be acoustically confusable when presented in a perceptual identification task.

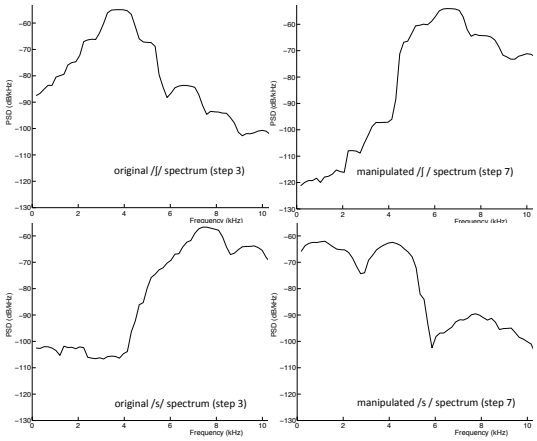

Figure 1: Comparison between prototypical recorded stimuli (/s/ and /ʃ/, left panels) and the acoustically manipulated stimuli with +48dB amplification (step7 in table 1) of the irrelevant frequencies (right panels).

2.2. Participants and experimental procedure

We recruited 32 native speakers of Canadian English. They were undergraduate students who received course credit for participation. All of them reported normal hearing. Due to the special location of McMaster University (Ontario) none of the listeners was monolingual[1], however we excluded all participants that had knowledge of languages with a three-way voiceless sibilant contrast (e.g., Mandarin or Polish) as such linguistic experience may result in perception different from other participants (e.g., Polish vs. English listeners in [19]).

The experiment was run as a (perceptual) phoneme identification task. For each trial, participants were presented with an isolated fricative noise sound: either one of the prototypical sibilant phonemes or any variant of the acoustically manipulated stimuli (see 2.1.). Listeners were then asked to identify the stimulus as either /s/ or /ʃ/ as accurately and fast as possible in a forced-choice identification[2]. There was no time limit for each trial, the ISI was 1.5 s. We excluded responses over 2.5 standard deviations for each listener. There was a practice session with 10 selected stimuli. In total each participant was presented with 112 stimuli (8 repetitions of the complete continuum, each continuum consisting of 14 stimuli (7 /s/ stimuli + 7 /ʃ/ stimuli).

3. RESULTS

Figure 2 shows, over all listeners, the mean probabilities of /s/ responses for all continuum steps described in 2.1, split by underlying stimulus identity /s ʃ/. As expected, the underlying /s/ and /ʃ/ stimuli (the prototypical stimulus, step3) and the manipulated stimuli with amplified *relevant* frequencies (step1, step2) were perceived as the original underlying sibilants (i.e. /s/ stimuli perceived as /s/ and /ʃ/ stimuli as /ʃ/).

Increasing the amplification of *irrelevant* frequencies (steps 4 to 7) led to a shift of listener responses towards the alternative sibilant (i.e. /s/ stimuli as /ʃ/ and /ʃ/ stimuli as /s/), thus introducing phoneme identity shift. However, here we observed an imbalance of the shifting magnitude towards the alternative phoneme based on the underlying phoneme identity: Whereas for the manipulated /ʃ/ stimuli at step 6 and step 7 the presented stimuli were reliably identified as the alternative sibilant /s/ (i.e. 100% as /s/), that behavior could not be observed for the manipulated /s/ stimuli: For this phoneme, listeners did *not* switch perception to the alternative sibilant but rather identified the stimuli at chance level (around 50%). In other words, there was a complete phonetic categorical switch for the underlying /ʃ/ stimuli (from /ʃ/ to /s/), but not for the underlying /s/ stimuli. This lack of phoneme shift is remarkable since the acoustic spectral distribution was extremely similar to a prototypical /ʃ/ sound (see figure 1). In sum, although our manipulation to amplify irrelevant frequencies influenced the listeners' identification of both sibilants to some extent, its effect on the two sibilants was not symmetric and the underlying *articulatory* structure

of the sibilant does seem to play a role here, leading to this perceptual asymmetry.

To statistically verify whether the manipulated sibilants with amplified *irrelevant* frequencies were perceived significantly different compared to the prototype stimuli of their alternative sibilants, we compared the proportions of /s/ and /ʃ/ prototype responses to those for the manipulated stimuli with amplified irrelevant frequencies (step 7) of the *alternative* sibilant. There was no significant difference between the prototypical /s/ and manipulated /ʃ/ stimuli (step7), $t(31) = 0.329$, $p = 0.745$, suggesting that the manipulated /ʃ/ stimuli were perceived similarly as the natural /s/ stimuli. In contrast, the comparison between the prototypical /ʃ/ and manipulated /s/ stimuli showed a significant difference, $t(31) = -6.211$, $p < .001$.

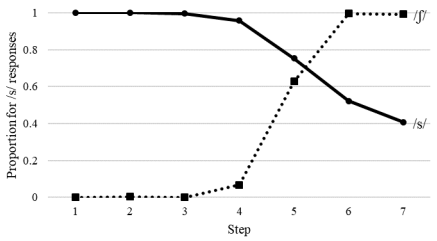

Figure 2: Probabilities of /s/ identifications (y-axis) against presented stimulus continuum (x-axis). The original recorded sibilant (/s/ or /ʃ/) is step 3, amplified *relevant* frequency stimuli are steps 1-2, and increasing amplification of *irrelevant* frequency regions are steps 4-7. See table 1 and text for further details.

Follow-up study: Since we divided the acoustic frequency range at 5 kHz to create one frequency region for acoustic manipulation below 5 kHz and one region above 5 kHz, acoustic information above 10kHz might have influenced identification decisions. In order to exclude this influence of residual high frequencies on listeners' sibilant identification we ran a follow-up pilot study (10 listeners) where we categorically filtered out all acoustic energy above 10kHz. All other parameters were identical to the previously described experiment. Results showed that identification results were identical (and the imbalance even stronger) to the previous experiment, thus showing that residual energy was not the reason for the perceptual imbalance.

4. DISCUSSION

Our results show that the acoustic manipulation of *irrelevant frequencies* influences the identification of voiceless sibilants, but in dependence of the *underlying sibilant category*: whereas identification of underlying postalveolar sibilants completely switches to the alternative sibilant, underlying alveolar sibilants largely resist the perceptual category shift. We now present two explanations for the observed perceptual imbalance.

Acoustic plus articulatory account: According to our original research hypothesis, manipulating acoustic information while maintaining articulatory information should allow the listeners to recover articulatory information (i.e. vocal tract configurations) for identification and disregard acoustic variation. In line with this hypothesis, *an acoustic plus articulatory account* claims that listeners indeed recover articulatory information, but this articulatory information is restricted to certain phonemes and does not expand to other phoneme categories. One reason why articulatory information would only influence /s/ (but not /ʃ/) identification is that different articulatory gestures are active comparing /s/ and /ʃ/: /ʃ/ is often produced with secondary articulation (i.e. lip rounding) in English and other languages, thus making the process of reliably extracting articulatory information more complicated, and thus more difficult than for /s/ without secondary articulatory configuration.

Universals and phoneme frequency: Another reason for the influence of articulatory information being restricted to /s/ identification could be that /s/ is generally favored in identification due to frequency effects. Ladefoged and Maddieson [9] showed that alveolar /s/ is much more common than postalveolar /ʃ/ in the world's languages (/s/: 85% of the examined languages; /ʃ/: 46%). There might be a universal reason why the production and/or perception of alveolar sibilants is generally preferred over postalveolars, and that preference, be it acoustic, articulatory or perceptual, could be the reason for the observed imbalance.

However, without considering articulation and the extraction of articulatory information we do not believe the last explanation is probable. As described, the acoustic spectra of the prototype (step3) and strongly manipulated alternative sibilants (step7) were extremely similar. All purely acoustic explanations would fail to explain how the *underlying* sibilant phoneme category (/s/ in this case) influenced our listeners to identify these acoustically extremely similar stimuli differently (comparing identification of strongly manipulated /s/ and prototype /ʃ/). Therefore, we consider the first explanation the most likely to explain our results: articulatory information anchors identification for alveolar, but not postalveolar sibilants. Follow-up experiments with more speakers and other phoneme categories will provide more data to solidify an explanation.

5. ACKNOWLEDGEMENTS

This work was partly funded by the Canadian NSERC grant RGPIN-2018-06518.

6. REFERENCES

[1] 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.

[2] 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.

[3] 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.

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

[5] Galantucci, B., Fowler, C., Turvey, M. T. 2006. The motor theory of speech perception reviewed" *Psychonomic Bulletin & Review*, *13*(3), 361-373.

[6] Harris, K. S. 1958. Cues for the discrimination of American English fricatives in spoken syllables, *Language and Speech, 1*(1), 1–7.

[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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

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

[14] Liberman, A. M., Mattingly, I. G. 1985. The motor theory of speech perception revised. *Cognition*, *21*, 1–36.

[15] Munson, B., Ryherd K, Kemper S. 2017. Implicit and explicit gender priming in English lingual sibilant fricative perception, Linguistics 55(5), 1073-1107.

[16] Ohala, J. J. 1996. Speech perception is hearing sounds, not tongues. *The Journal of the Acoustical Society of America* 99(3), 1718-25.

[17] Perrier, P. 2005. Control and representations in speech production. *ZAS Papers in Lingustics*, *40*, 109–132.

[18] Stevens, K.N. 1998. *Acoustic Phonetics*, MIT Press: Cambdridge.

[19] 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.

[1] English was the first and dominant language for all speaker. The definition of *native English* was being born or arrived in Canada (non-French part) before the age of five. Most listeners had French as L2 and some listeners had knowledge of Spanish, Italian or Hindi.

[2] In the full perception experiment, the isolated fricative noises (reported in this paper) were mixed with VCV stimuli (C=sibilant), so listeners were able to identify the sex of the speaker based on the presented vowel (V_V) information. It is assumed here that listeners judged the sex of the speaker for the isolated fricative noises identical to the intermittent vowels, thus corresponding to female sibilant productions [15]. The data for the VCV sequences will be reported elsewhere. Our female speaker self-identified as female.