The influence of actual and imputed talker gender on fricative perception, revisited (L)

Benjamin Munson[a)]
University of Minnesota, Department of Speech-Language-Hearing Sciences, 115 Shevlin Hall, 164 Pillsbury Drive, SE, Minneapolis, Minnesota 55455

(Received 12 May 2011; revised 29 July 2011; accepted 26 August 2011)

To examine the role of perceived gender on fricative identification, a study was conducted in which listeners identified /s/-/ʃ/ and /s/-/θ/ continua combined with vowels produced by a man and a woman. These were acoustically modified to be consistent with different-sized vocal tracts (VT), and were presented with pictures of men or women. Listeners identified more tokens of /s/ in the /s/-/ʃ/ and more tokens of /θ/ in the /s/-/θ/ continuum when these sounds were combined with men's vowels, with vowels consistent with a 17 cm VT, and with pictures of men. Results support the hypothesis that listeners incorporate information about talker gender during fricative perception.
© *2011 Acoustical Society of America*. [DOI: 10.1121/1.3641410]

PACS number(s): 43.71.Bp, 43.71.Es Pages: 2631–2634

I. INTRODUCTION

Perhaps the most daunting challenge facing listeners is that of normalization. Listeners are capable of providing consistent perceptual responses to speech signals whose acoustic characteristics are highly variable. This question of how listeners transform variable signals to invariant responses has arguably been at the center of the field of speech perception research since its inception.

The acoustic detail of speech varies as a function of social categories. Consider one well-studied social variable, gender. The acoustic characteristics of men and women's speech differ. As argued by Johnson (2006), these are not reducible to simple anatomic differences between sexes, but partly reflect learned, socially and culturally specific gendered ways of speaking. This hypothesis is supported by Stuart-Smith (2007), who examined that magnitude male-female differences in the acoustic characteristics of /s/ productions of speakers of Glaswegian English. Stuart-Smith found that these differed as a function of age and socioeconomic status: smaller differences were found for younger, working-class people than for middle-class people or older working-class ones. The hypothesis is also supported by Fuchs and Toda (2009), who found that male-female differences in /s/ acoustics were not predictable from a number of anatomical measures known to predict /s/ acoustics.

A number of recent studies have shown that listeners are sensitive to socially stratified phonetic variation during speech perception, as reviewed by Thomas (2002). More-recent studies, such as Drager (2011), continue to find this association. Drager showed that New Zealand listeners' perception of low-front vowels differed when they were presented with pictures of older and younger speakers, reflecting knowledge of the fact that younger New Zealand speakers' productions reflect an ongoing sound change in which the TRAP vowel is raised to [ɛ]. Older speakers' productions of TRAP reflect the older [æ] pronunciation. Consistent with this, stimuli intermediate between [ɛ] and [æ] were more likely to be identified as members of the TRAP lexical class when presented with pictures of older speakers, and the DRESS class when paired with pictures of younger speakers.

The focus of this brief communication is on the influence of presumed speaker gender on the perception of fricatives. Johnson (1991) showed that listeners identify more /s/ tokens from an /s/-/ʃ/ continuum when it is combined with men's productions of vowels than when it is combined with women's productions. The peak frequency of /s/ is higher than that for /ʃ/, and the peak frequency of both fricatives is lower in men's productions than in women's. This finding was later replicated by Munson *et al.* (2006). Johnson and Munson *et al.*'s findings appear to reflect listeners' perceptual compensation for sex differences in production. Strand and Johnson (1996, see also Strand, 1999) conducted an audiovisual speech perception experiment in which listeners identified a *sod-shod* continuum while viewing video clips of men and women. Stimuli were constructed by combining a synthetic /s/-/ʃ/ continuum with natural productions of *od* from tokens of *sod* and *shod* produced by four talkers: two men and two women. One man and one woman were identified as prototypically male or female sounding; the other two talkers' voices were judged to be non-prototypically gendered. Listeners' identified more tokens of /s/ when the non-prototypical talkers' continua were paired with men's faces than with women's. This pattern parallels what Johnson (1991) and Munson *et al.* (2006) found, and suggests that listeners' access their tacit knowledge of gendered speech patterns when identifying fricatives. Strand and Johnson argue that this finding is strong evidence that the findings in Johnson (1991) and Munson *et al.* (2006) do not reflect mere vocal-tract normalization, i.e., estimating the talker's vocal-tract length from the acoustic characteristics of their vowels and adjusting the fricative perception accordingly. If that explanation were true, then visual images would not have affected perception.

This manuscript presents the results of an experiment in which listeners identified fricatives paired with pictures of men and women, a manipulation we refer to henceforth as *imputed gender*. This experiment has three purposes. The

a)Author to whom correspondence should be addressed. Electronic mail: Munso005@umn.edu

Downloaded 30 Nov 2011 to 134.84.138.14. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

first is to replicate Strand and Johnson's finding. While studies have shown the effect of imputed gender on the perception of other phonemes (i.e., Johnson *et al*. 1999), and on the influence of social categories on the perception of phonemes in general, no study other than Strand and Johnson has demonstrated an effect of imputed gender on fricative categorization. A replication would strengthen Strand and Johnson's potentially very influential finding.

The second purpose is to examine the influence of imputed gender on fricative identification when certain acoustic characteristics of men and women's voices are controlled carefully. The stimuli Strand and Johnson used varied in the acoustic parameters that are known to be correlated with actual and perceived vocal-tract length: their formant frequencies, and their f0 (i.e., Gonzalez, 2004; van Dommelen and Moxness, 1995). This study examined the perception of men's and women's tokens that had been acoustically modified so that two parameters known to affect perceived vocal-tract length—F3 and f0—were equivalent. That is, we synthesized synthesize stimuli with different *apparent vocal-tract length*s (henceforth *aVTL*) This manipulation allows us to examine (a) whether the strength of the influence of imputed gender on identification is equivalent for different talkers and for different aVTLs, and (b) whether speaker gender influences fricative identification when the acoustic parameters known to affect perceived vocal-tract length are controlled statistically.

The third purpose is to examine the influence of talker sex, aVTL, and imputed gender on the perception of an /s/-/θ/ continuum. This contrast is particularly interesting because the acoustic parameters that differentiate between /s/ and /θ/ are different from those that distinguish /s/ from /ʃ/. The principle acoustic difference between /s/ and /ʃ/ is in the peak frequency (Jongman *et al*. 2000), which itself is a related to the size of the resonant cavities anterior and posterior to the constriction. The size of this cavity potentially differs between men and women, and may explain at least some of the observed sex differences in this acoustics of this sound. Jongman *et al*. found that the peak frequency of the /θ/ is lower than that of /s/, just as that of /ʃ/ is. Based on this, we might predict that the effects of speaker sex, aVTL, and imputed gender on /s/-/θ/ identification would parallel their effects on /s/-/ʃ/ perception: listeners might be biased to expect lower peak-frequency stimuli for men than for women, leading listeners to be biased toward /s/ responses when listening to men and to /θ/ responses when listening to women. However, though /s/ and /θ/ differ in peak frequency, the principal acoustic difference between them in the distribution of energy in the spectrum rather than to the location of the peak frequency. This parameter is less transparently related to vocal-tract size than peak frequency is. Moreover, unlike /s/, the center frequencies of men and women's /θ/ productions do not differ. Hence, we predict that perception of this contrast will be less influenced by sex, aVTL, and imputed gender would the perception of /s/-/ʃ/. There are, however, ample anecdotal evidence of social stereotypes about /s/ variation and gender typicality (Munson, 2010) which may cause actual or imputed gender to affect fricative perception. These stereotypes typically characterize the /s/ productions of less-prototypically male-sounding men as more frontal than those of more prototypically male-sounding men. This stereotype is supported by the finding that listeners rate men to sound less prototypically masculine when they produce /θ/-like /s/ tokens (Munson and Zimmerman, 2006). Based on this, we might expect that listeners would identify stimuli ambiguous between /s/ and /θ/ as /s/ (i.e., as a frontal /s/, rather than as /θ/) when paired with any voice that isn't prototypically male (including a woman's voice) than when paired with a voice that is prototypically male. A finding that sex and actual and imputed female gender leads listeners to identify more /s/-/θ/ stimuli as /s/ would buttress Strand and Johnson's argument that these effects are due more to sociocultural knowledge of gendered speech patterns than to mere vocal-tract normalization.

II. METHODS

A. Participants

Listeners were 20 individuals from the University of Minnesota community, who responded to advertisements posted on campus. They ranged in age from 18 to 40, and included three men and 17 women. All were native, monolingual speakers of American English, and none reported a current or past speech, language, or hearing impairment. They were paid $10 for their participation.

B. Stimuli

Stimuli consisted of eight continua: four *sigh-shy* continua, and four *sigh-thigh* continua. These were created by combining two different fricative continua with a set of eight acoustically modified vowels, based on productions by one man and one woman. Both talkers spoke American English natively, and each produced a fully diphthongal token of the vowel /a͡ɪ/, as judged by two trained phoneticians, excised from productions of the word *sigh*. The initial fricative was edited off of these stimuli. To create different aVTLs, the rime portions of the continua were created by varying the formant frequencies and the fundamental frequencies of the naturally produced /a͡ɪ/ tokens. This was accomplished using the PSOLA algorithm in Praat. PSOLA includes a tool to scale the talker's apparent vocal-tract size, using Wakita's (1977) algorithm for estimating vocal-tract size from acoustic signals. For both the man and the woman's /a͡ɪ/, one token was made that was consistent with a 14.2 cm vocal tract (i.e., an F3 of 2600 Hz, assuming that the speed of sound is 34,000 cm/s). A second set of the man and the woman's /a͡ɪ/ were made to be consistent with a 17 cm vocal tract (i.e., an F3 of 3000 Hz). All stimuli had a peak f0 of 135 Hz. The man and woman's /a͡ɪ/ differed in their perceived voice quality, with the woman's token sounding breathier than the man's. Moreover, the two differed in length: the woman's token was approximately 7% longer than the man's production. The first and second formant frequencies of the man's productions were lower than those of the woman's, even after the formants had been scaled so that the F3 values were equal.

Two seven-step fricative continua were created. The first of these was the middle seven steps of the very same /s/-/ʃ/ continuum used by Strand and Johnson (1996), with the additional manipulation that the intensity of the fricatives

Downloaded 30 Nov 2011 to 134.84.138.14. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

was scaled so that the /ʃ/ endpoint was more intense than the /s/ endpoint. The second was an /s/-/θ/ continuum that was created by first mixing the /s/ endpoint of the /s/-/ʃ/ continuum with a naturally produced /θ/ that had been equated to the amplitude of the /s/ at seven different amplitude ratios. The fricatives' amplitudes were modified because fricative /vowel intensity ratios have been shown to affect fricative categorization (Hedrick and Ohde, 1993). Following this, the overall amplitude of the resulting seven fricatives was altered so that the /s/ endpoint was the most intense and the /θ/ was the least intense. Further details of the fricative-combining algorithm can be found in McGuire (2007). The fricatives were combined with the natural /a͡ɪ/ bases. Though the /a͡ɪ/ was excised from a token of *sigh* and therefore had formant transitions most appropriate for an alveolar fricative, the *shy* and *thigh* endpoints were sufficiently natural to be identified as such in a pilot test 100% of the time.

Three visual stimuli were used. Two of these were pictures taken from the Caltech Frontal Facial Database (http://www.vision.caltech.edu/Image_Datasets/faces/README) of one man and one woman, chosen because their faces were prototypically male and female, respectively. The third image, used for filler trials, was of a checkerboard pattern approximately the same size as the pictures.

C. Procedures

The experiment was administered using the E-Prime experiment management software. Participants were seated in a double-walled sound-treated room wearing Sennheiser HD 280 Pro headphones. On each trial, the text "listen carefully" was presented at the center of a 15" monitor in 36-point courier font, followed by a 1 s presentation of one of the three visual stimuli, followed by the sound file presented at approximately 65 dB IL, followed by a response screen. Responses were elicited using visual analog scaling (VAS). Participants viewed a 425-pixel double-arrow line with a vertical mark at the midpoint, the text "the 's' sound" to the left of the left arrow, and "the 'sh' sound" or "the 'th' sound" to the right of the right arrow. Participants responded by clicking on the line where they thought the token fell relative to the /s/, /ʃ/, or /θ/ endpoints. The choice to use VAS was done, in part, so that we could compare these participants' identification of controlled fricative stimuli with their VAS ratings of children's productions of /s/, /ʃ/, and /θ/, the results of which are presented elsewhere (Munson *et al.* 2010). They were also chosen because of the finding, presented in Urberg-Carlson *et al.* (2008), that VAS ratings track gradient perception of fricatives better than simple binary categorization responses. The *sigh-shy* and *sigh-thigh* trials were presented in separate blocks. Block order was randomized across subjects. The trials with the man's face, the woman's face, and the filler checkerboard stimulus were interspersed with each other randomly. Four ratings of each stimulus were elicited; hence, a total of 672 responses were elicited over the course of the experiment (2 voice sexes × 2 aVTLs × 3 visual stimuli × 2 continua × 7 steps per continuum × 4 repetitions of each stimuli). The entire experiment was conducted in a single session that took approximately 45 min.

D. Analyses

Each individual click was logged in pixels on the *x* axis. These were transformed to proportions of the total line length in pixels, from 0.0 for clicks at the left edge to 1.0 for clicks at the right edge. These proportions were subjected to probit analysis. The hypothetical fractional step on the continuum that would elicit a response at the midpoint of the line was calculated. We refer to this henceforth as the *crossover point*. The slope of the Probit function was also calculated. Slopes were uninformative, and are thus not analyzed here. One subject's ratings of the /s/-/θ/ stimuli were almost exclusively at or near the endpoint of the visual analog scaled labeled with "the 's' sound." Hence, this person's crossover could not be calculated. This person's data for both the /s/-/ʃ/ and /s/-/θ/ trials were excluded from further analysis.

III. RESULTS

Individual subjects' crossover points were subjected to two fully within-subjects three-factor (2 voice sex × 2 aVTL × 2 picture sex) ANOVAs. For each significant effect or factor, a measure of effect size, η^2_{partial}, was calculated. For the crossover points on the /s/-/ʃ/ continua, there were significant main effects of talker sex ($F[1,18] = 13.8$, $p = 0.002$, $\eta^2_{\text{partial}} = 0.44$) and VTL ($F[1,18] = 9.3$, $p = 0.007$, $\eta^2_{\text{partial}} = 0.34$). There was no main effect of imputed gender, but voice sex interacted with imputed gender significantly, $F[1,18] = 4.422$, $p = 0.05$, $\eta^2_{\text{partial}} = 0.20$. As expected, ratings closer to the /s/ end of the visual analog scale were found for the 17 cm aVTL, for the man's /a͡ɪ/, and for the tokens paired with a man's picture. The interaction between voice sex and face sex can be seen by comparing the bar heights in Fig. 1. As this shows, the effect of picture sex on listeners' crossover points was strongest for the stimuli based on the woman's /a͡ɪ/, and in particular on the woman's /a͡ɪ/ with the aVTL.

For the /s/-/θ/ continua, all three main effects affected crossover points significantly: Voice sex: $F[1,17] = 7.2$, $p = 0.02$, $\eta^2_{\text{partial}} = 0.30$; aVTL: $F[1,17] = 29.3$, $p < 0.001$, $\eta^2_{\text{partial}} = 0.63$, imputed gender, $F[1,17] = 6.7$, $p = 0.02$, $\eta^2_{\text{partial}} = 0.28$. Ratings closer to the /θ/ end of the visual analog scale were elicited for men's voices, for 17 cm vocal tracts, and for stimuli paired with men's faces. None of the factors interacted significantly. Although there were no significant interactions, the bar heights in Fig. 2 suggest that the effect was strongest for the 14 cm aVTL, and for stimuli appended to men's voices.

IV. DISCUSSION

The first purpose of this investigation was to determine if Strand and Johnson's finding regarding /s/-/ʃ/ perception. Results showed that indeed it could be: listeners identified fricatives as more /s/-like when they were presented with pictures of men than with pictures of women. The second purpose of this investigation was to examine the relative size of the effects of talker sex, imputed gender, and apparent vocal-tract length on the identification of /s/-/ʃ/. The ANOVA results showed a robust effect of aVTL on identification that was statistically independent of the effects of talker sex and imputed gender. The effect of talker sex was also significant.

Downloaded 30 Nov 2011 to 134.84.138.14. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

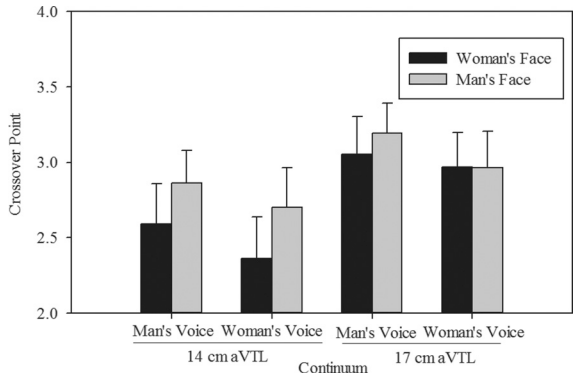

FIG. 1. Location of the crossover points for the /s/-/ʃ/ continuum, separated by talker sex, apparent vocal-tract length, and imputed gender. Higher values indicate identification functions with relatively more/ʃ/responses. Lower values indicate identification functions with relative more /s/ responses.

Though the effect of talker sex interacted with imputed gender, a qualitative inspection of the bar heights of Fig. 1 suggests that the effect of talker sex was robust across different imputed genders, but that the effect of imputed gender was present only for the two continua based on the woman's /a͡ɪ/. The η^2_{partial} values showed that talker sex had a larger influence on identification patterns than did aVTL.

The third purpose of this investigation was to examine the influence of talker sex, aVTL, and imputed gender on /s/-/θ/ perception. More /s/ tokens were identified when this continuum was paired with a woman's voice, a woman's face, and a 14.2 cm vocal tract than when paired with a male voice or face, or with a 17 cm vocal tract. This is not predicted by the acoustic characteristics of men and women's /s/ and /θ/ productions. Though the /s/ productions differ acoustically, the /θ/ productions do not. Hence, the finding arguably supports Strand and Johnson's hypothesis that gender effects on fricative perception reflectknowledge of culturally specific gendered ways of speaking (for /s/-/ʃ/ perception) or stereotypes about gender and speech (for /s/-/θ/ perception). We also predicted that the effects of aVTL of actual and imputed gender would be stronger for the /s/-/ʃ/ continuum than for the /s/-/θ/ continuum. This hypothesis was not supported. Indeed, the η^2_{partial} suggested stronger effects of aVTL and imputed gender on /s/-/θ/ than on /s/-/ʃ/ perception. There are well-established stereotypes that less-prototypically masculine talkers' productions are /θ/-like. There are no clear stereotypes regarding the relationship between /s/-/ʃ/ and gender. Hence, the stronger effect of imputed gender on /s/-/θ/ perception than on /s/-/ʃ/ perception is again evidence that stereotypes about gender are activated during speech perception.

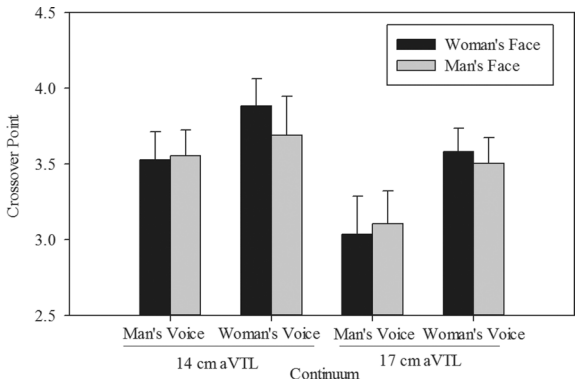

FIG. 2. Location of the crossover points for the /s/-/θ/ continuum, separated by talker sex, apparent vocal-tract length, and imputed gender. Higher values indicate identification functions with relatively more /θ/ responses. Lower values indicate identification functions with relative more /s/ responses.

ACKNOWLEDGMENTS

This research was supported by a McKnight Presidential Fellowship and by NSF grant BCS 0729277 to Benjamin Munson. Generous thanks to Keith Johnson and Elizabeth Strand for allowing me to use the fricative stimuli from Strand and Johnson (1996).

Drager, K. (**2011**). "Speaker age and vowel perception," Lang. Speech **54**, 99–121.

Fuchs, S., and Toda, M. (**2009**). "Do differences in male versus female /s/ reflect biological factors or sociophonetic ones?" *An Interdisciplinary Guide to Turbulent Sounds*, edited by S. Fuchs, M. Toda, and M. Zygis (Mouton de Gruyter, Berlin), pp. 281–302.

González, J. (**2004**). "Formant frequencies and the body size of the speaker: a weak relationship," J. Phonetics **32**, 177–187.

Hedrick, M., and Ohde, R. (**1993**). "Effect of relative amplitude of frication on perception of place of articulation," J. Acoust. Soc. Am. **94**, 2005–2026.

Johnson, K. (**1991**). "Differential effects of speaker and vowel variability on fricative perception," Lang. Speech **34**, 265–279.

Johnson, K. (**2006**). "Resonance in an exemplar model of phonology," J. Phonetics **43**, 485–499.

Johnson, K., Strand, E. and D'Imperio, M. (**1999**). "Auditory-visual integration of talker gender in vowel perception," J. Phonetics **27**, 359–384.

Jongman, A., Wayland, R. and Wong, S. (**2000**). "Acoustic characteristics of English fricatives," J. Acoust. Soc. Am. **108**, 1252–1263.

McGuire, G. (**2007**). "Phonetic Category Learning," Doctoral Dissertation, Department of Linguistics, Ohio State University, Columbus, OH, Downloaded on May 12, **2011** from http://rave.ohiolink.edu/etdc/view?acc_num = osu1190065715.

Munson, B. (**2010**). "Variation, implied pathology, social meaning, and the 'gay lisp': a response to Van Borsel et al. (2009)," J. Commun. Disorders **43**, 1–5.

Munson, B., Edwards, J., Schellinger, S. K., Beckman, M. E., and Meyer, M. K. (**2010**). "Deconstructing phonetic transcription: covert contrast, perceptual bias, and an extraterrestrial view of *Vox Humana*," Clin. Linguist. Phonetics **24**, 245–260.

Munson, B., Jefferson, S.V., and McDonald, E. C. (**2006**). "The influence of perceived sexual orientation on fricative perception," J. Acoust. Soc. Am. **119**, 2427–2437.

Munson, B., and Zimmerman, L. J. (**2006**). "Perceptual Bias and the Myth of the 'Gay Lisp'," Poster Presentation at the Annual Meeting of the American Speech-Language-Hearing Association, Miami, FL.

Strand, E. (**1999**). "Uncovering the role of gender stereotypes in speech perception," J. Lang. Soc. Psychol. **18**, 86–99.

Strand, E., and Johnson, K. (**1996**). "Gradient and visual speaker normalization in the perception of fricatives," in *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference, Bielfelt*, edited by D. Gibbon, October 1996 (Mouton de Gruyter, Berlin), pp. 14–26.

Stuart-Smith, J. (**2007**). "Empirical evidence for gendered speech production: /s/ in Glaswegian," in *Laboratory Phonology 9*, edited by J. Cole and J. Hualde (Mouton de Gruyter, Berlin), pp. 65–86.

Thomas, E. (**2002**). "Sociophonetic applications of speech perception experiments," Am. Speech **77**, 115–147.

Urberg-Carlson, K., Munson, B., and Kaiser, E. (**2008**). "Assessment of Children's Speech Production 2: Testing Gradient Measures of Children's Productions," Poster presented at the American Speech-Language-Hearing Association Convention, Chicago, IL, http://www.tc.umn.edu/~munso005/Urberg-CarlsonEtAl_Final. pdf (Last viewed 26 April 2001).

Van Dommelen, W., and Moxness, B. (**1995**). "Acoustic parameters in speaker height and weight identification: sex-specific behavior," J. Phonetics **38**, 267–287.

Wakita, H. (**1977**). "Normalization of vowels by vocal-tract length and its application to vowel identification," Trans. IEEE Acoust., Speech, Signal Process. Soc. **25**, 183–192.

Downloaded 30 Nov 2011 to 134.84.138.14. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp