Contributions of Semantic and Facial Information to Perception of Nonsibilant Fricatives

Allard Jongman
University of Kansas, Lawrence

Yue Wang*
Brian H. Kim**
Cornell University, Ithaca, NY

Most studies have been unable to identify reliable acoustic cues for the recognition of the English nonsibilant fricatives /f, v, θ, ð/. The present study was designed to test the extent to which the perception of these fricatives by normal-hearing adults is based on other sources of information, namely, linguistic context and visual information. In Experiment 1, target words beginning with /f/, /θ/, /s/, or /ʃ/ were preceded by either a semantically congruous or incongruous precursor sentence. Results showed an effect of linguistic context on the perception of the distinction between /f/ and /θ/ and on the acoustically more robust distinction between /s/ and /ʃ/. In Experiment 2, participants identified syllables consisting of the fricatives /f, v, θ, ð/ paired with the vowels /i, ɑ, u/. Three conditions were contrasted: Stimuli were presented with (a) both auditory and visual information, (b) auditory information alone, or (c) visual information alone. When errors in terms of voicing were ignored in all 3 conditions, results indicated that perception of these fricatives is as good with visual information alone as with both auditory and visual information combined, and better than for auditory information alone. These findings suggest that accurate perception of nonsibilant fricatives derives from a combination of acoustic, linguistic, and visual information.

KEY WORDS: perception, fricatives, semantic context, facial context

Most research on fricatives has not been able to identify consistent acoustic characteristics that may serve to distinguish the labiodental (/f, v/) and dental (/θ, ð/) fricatives. Neither spectral, temporal, nor amplitude properties of the frication noise have been shown to reliably distinguish /f/ from /θ/ and /v/ from /ð/. Results from perceptual experiments suggest that cues to the perception of labiodental and dental fricatives are located in the transition and vocalic portion of fricative-vowel syllables rather than the noise portion (e.g., Harris, 1958; LaRiviere, Winitz, & Herriman, 1975). However, stable acoustic differences between /f, v/ and /θ, ð/, in terms of vocalic attributes, remain to be documented. This difficulty in determining the defining acoustic properties has been mirrored at the perceptual level. Among fricatives, /f/ and /θ/ and /v/ and /ð/ are most easily confused (e.g., Balise & Diehl, 1994; Jongman & Wang, submitted). Given the reported difficulty in recognition of these fricatives, G. A. Miller and Nicely (1955) hypothesized that the distinction between /f/ and /θ/ and between /v/ and /ð/ may be based on nonacoustic information:

AQ1

AQ2

*Currently at SUNY Buffalo, Buffalo, NY

**Currently at Michigan State University, East Lansing

> The distinctions between /f/ and /θ/ and between /v/ and /ð/ are among the most difficult for listeners to hear and it seems likely that in most natural situations the differentiation depends more on verbal context and on visual observation of the talker's lips than it does on the acoustic difference. (G. A. Miller & Nicely, 1955, p. 347)

Similar observations were made by Massaro (1987, 1998), who argued that the contribution of visual information increases as auditory distinctiveness decreases (see Sumby & Pollack, 1954, for one of the earliest quantitative reports). Research has clearly demonstrated that providing contextual or visual information generally improves speech perception (see Massaro, 1987, for a review). However, it is not clear whether all speech sounds benefit to the same extent from these kinds of information. The present experiments specifically focused on the role of linguistic context and visual information in the perception of nonsibilant fricatives because it has been so difficult to find reliable acoustic and perceptual cues to their distinction.

Experiment 1: Effects of Linguistic Context

AQ3 Rationale

A number of studies have shown that the linguistic content of a carrier sentence can affect the categorization of phonetic information in spoken words. A variety of paradigms have been used, including shadowing, gating, and monitoring (e.g., Cole, 1973; Connine, 1987; Garnes & Bond, 1976; Grosjean, 1980; Marslen-Wilson, 1973; G. A. Miller & Isard, 1963; Pollack & Pickett, 1963). Experiments using these paradigms have included response accuracy, response latency, or both as response measures and have typically shown that a related context leads to faster responses and fewer errors in the identification of a target sound or word. For example, Garnes and Bond (1976) demonstrated that linguistic context could affect perception of place of articulation of word-initial stop consonants. Garnes and Bond created a continuum of target words from *bait* to *date* to *gate* by manipulating F2 transition. Each continuum member was then spliced in as the last word in each of three carrier sentences, each one biased toward one of the three words: *Here's the fishing gear and the* ___, *Check the time and the* ___, and *Paint the fence and the* ___. Listeners were to identify the target words. When the stimulus information was phonetically unambiguous, listeners reported hearing the words correctly. This sometimes resulted in semantically anomalous sentences (e.g., *Paint the fence and the date*). However, when the onset of the target-initial consonant was phonetically ambiguous, listeners would report hearing the word that was semantically congruous with the sentence context. These results indicate that in those cases in which phonetic information is less clear, listeners use contextual information to help them decode the message. J. L. Miller, Green, and Schermer (1984) showed that linguistic context also affects perception of voicing in word-initial stop consonants. In this study, members of a VOT continuum ranging from *bath* to *path* were preceded by either a sentence semantically congruent with *bath* (*She needs hot water for the…*) or *path* (*She likes to jog along the…*). Results showed that although the endpoints were unambiguously identified regardless of the precursor, continuum members with intermediate VOT values were identified on the basis of their congruity with the preceding context. AQ4

The present experiment examined the effects of linguistic context on the perception of the English fricatives /f, θ, s, ʃ/. By preceding fricative-initial target words by different precursor sentences, the extent to which perception of fricatives is affected by sentential linguistic information could be evaluated. Precursor sentences were selected such that they were either semantically congruous or incongruous with a fricative-initial target word. The use of fricatives rather than stops also allowed us to explore an additional variable in these perceptual studies, namely the acoustic–phonetic robustness of the stimulus. Stimulus robustness was manipulated naturally by contrasting nonsibilant fricatives to sibilant fricatives. Because the acoustic properties of nonsibilant fricatives are still poorly understood, the use of naturally produced, nonmanipulated fricatives was preferred over the creation of a synthetic /f–θ/ continuum. Previous research on fricatives has shown that the nonsibilant fricatives are often confused, whereas the sibilant fricatives are rarely confused (Balise & Diehl, 1994; Jongman & Wang, submitted; G. A. Miller & Nicely, 1955). Minimal-pair targets with confusable onsets (e.g., *first–thirst*) and acoustically robust onsets (e.g., *suit–shoot*) were used. By contrasting these types of target-initial fricatives, the influence of context on phonetically confusable and robust targets could be evaluated. This manipulation allowed us to determine if context effects are modulated by degree of acoustic robustness in the target word. For the present experiment, then, it was expected that perception of the nonsibilant fricatives may be more affected by context than that of the sibilant fricatives. Two response measures were used in this study: both accuracy and latency of responses. Response latency can sometimes reveal patterns that are not apparent when only response accuracy is taken into account (Pisoni & Tash, 1974). In general, responses will take longer when additional processing of the stimulus is required or when conflicting sources of information AQ2

lead to increased ambiguity (see Massaro, 1987; Sawusch, 1996, for a discussion of reaction time as a measure of mental processes in speech perception research).

Method

Participants

Twenty listeners (12 women, 8 men) were recruited from the Cornell University student population. All were native speakers of English and stated that they did not have any speech or hearing impairments. Participants were paid for their participation.

Stimuli

Participants heard contexts followed by target words. The target words were 20 minimal pairs: 10 pairs began with either nonsibilant /f/ or /θ/, and 10 pairs began with sibilant /s/ or /ʃ/. This choice of minimal pairs was motivated by previous experiments on fricative perception (Balise & Diehl, 1994; Jongman & Wang, submitted; Tomiak, 1990). These studies showed nonsibilant confusions with nonsibilants and sibilant confusions with sibilants. This research also indicated that the contrast between sibilants can be related to specific acoustic cues whereas the contrast between nonsibilants is more elusive. Only voiceless fricatives were included, because some comparisons (e.g., /v/–/ð/) cannot occur in identical sentential contexts. In the present experiment, then, the nonsibilant fricatives /f, θ/ were compared with the sibilant fricatives /s, ʃ/.

AQ2

To assess the predictability of the target words as a function of the precursors, a written version of each precursor was presented to a group of 9 undergraduates. To mimic the conditions of the speech perception experiment, separate lists were made for words that started with /f/ and /θ/ or /s/ and /ʃ/. For a given list, participants were instructed to write down the first word that came to mind that, depending on the list, started with either /f/ and /θ/ or /s/ and /ʃ/, respectively. Results showed that sentences were highly predictable and comparable in terms of their degree of predictability across sibilant and nonsibilant contexts (correct identification rates were 84%, 89%, 87%, and 80% for words starting with /f/, /θ/, /s/, and /ʃ/, respectively).

Targets and sentences were recorded by a female speaker of General American English in the Cornell Phonetics Laboratory. The recordings were made within a soundproof booth (IAC) with a high-quality microphone (Electro-Voice RE20), microphone pre-amp (Gaines Audio MP-1), and cassette deck (Carver TD1700). The microphone was placed at approximately a 45° angle and 15 cm away from the corner of the speaker's mouth to prevent turbulence due to direct airflow from impinging on the microphone. All recordings were sampled at 22 kHz (16-bit quantization, 11-kHz (antialias) low-pass filter) on a Sun SPARCstation 5. Because the acoustic characteristics of segments have been shown—at least in the temporal domain—to vary as a function of their predictability (Charles-Luce, 1993), all target words were produced in isolation and spliced into each of the two contexts (congruous and incongruous). All context sentences were produced with the same unrelated word in target position. This word was subsequently replaced by the appropriate congruous or incongruous target word. This was done to avoid any coarticulatory effects that could bias the results. Because the target words were not pronounced as part of the utterance, there was no coarticulatory information in the word preceding the target word that potentially could cue the identity of its initial fricative. The mean duration of the target words as a function of initial fricative was as follows: /f/ = 461 ms, /θ/ = 423 ms, /s/ = 485 ms, and /ʃ/ = 499 ms. A 55-ms silence interval was inserted between context sentence and target word. The sentences were matched in terms of number of syllables (six). Examples of target words and congruous and incongruous context sentences are shown in Table 1. A list of all materials can be found in the Appendix. Stimulus sentences (context plus target word) were played binaurally from disk over headphones (Sony MDR-7506, frequency response from 10 to 20000 Hz) at a comfortable listening level, using Bliss software (Mertus, 1989).

Procedure

The experiment was designed such that each target served as its own control. That is, listeners responded

Table 1. Examples of target words and congruous and incongruous context sentences used in Experiment 1. See the Appendix for a complete list of stimuli.

Target word	Congruous condition	Incongruous condition
thirst	The lemonade quenched my ___	The top swimmer came in ___
first	The top swimmer came in ___	The lemonade quenched my ___
shag	The rug she chose was a ___	The old bridge began to ___
sag	The old bridge began to ___	The rug she chose was a ___

to the same target as a function of different precursors. Sibilant and nonsibilant targets were blocked. Order of blocks was counterbalanced across subjects. Listeners were tested in groups of two to four. For a given block, the listeners' task was to indicate whether the final word started with either /f/ or /θ/ or with either /s/ or /ʃ/ by pressing one of two response buttons. The buttons were labeled either *f* and *th* or *s* and *sh*, respectively. Listeners could only make one decision per trial. Stimuli were presented to listeners in random order at 3-s intervals. Listeners heard 40 stimuli for the nonsibilant contrast (10 words × 2 fricatives × 2 contexts) and 40 stimuli for the sibilant contrast (10 words × 2 fricatives × 2 contexts), for a total of 80 stimuli. Feedback was not provided.

Both response accuracy and latency were measured. Response accuracy was simply whether the correct fricative was identified. Because all materials were presented in quiet and ceiling effects could occur, response latency was also collected. Response latency was measured from onset of the target word.

This design allowed for the comparison of responses to the *same* target as a function of the two precursors. If context influences perception of either /f, θ/ or /s, ʃ/, then responses to the target are expected to be faster and/or more accurate when preceded by the congruous context as compared to the incongruous context. The pattern of results across nonsibilant and sibilant contrasts can also be compared.

Results

Correct fricative identification scores are displayed in Figure 1. A two-way (Context × Fricative) analysis of variance (ANOVA) revealed a main effect for context, $F(1, 7) = 40.73$, $p = .000$. Fricative identification was significantly better in the congruous context (97%) than in the incongruous context (89%). There was also a main effect for fricative, $F(3, 7) = 35.97$, $p = .000$. Bonferroni post hoc tests indicated that identification of targets beginning with /s/ (99%) and /ʃ/ (98%) was significantly more accurate than of those starting with /θ/ (92%) and /f/ (83%). In addition, the difference between the two nonsibilant fricatives was also significant, with perception of /θ/ more accurate than /f/. Importantly, the Context × Fricative interaction was significant, $F(3, 7) = 17.33$, $p = .000$. Simple effects tests showed that although identification of nonsibilant-initial words benefited from a semantically congruous context (95% in congruous context vs. 81% in incongruous context), identification of sibilant-initial words was very high and not affected by context (99% in both contexts).

Response latencies to targets identified correctly are shown in Figure 2. A two-way (Context × Fricative) ANOVA revealed a main effect for context, $F(1, 7) = 7.52$, $p = .007$. Fricative identification was significantly faster in the congruous context (778 ms) than in the incongruous context (856 ms). There was also a main effect for fricative, $F(3, 7) = 3.64$, $p < .014$. Bonferroni post hoc tests indicated that identification of targets beginning with /f/ (897 ms) was significantly slower than of those starting with /θ/ (784 ms), /s/ (802 ms), and /ʃ/ (783 ms). There was no significant difference among the latter three fricatives. Finally, there was no significant Context × Fricative interaction, $F(3, 7) = 0.195$, $p = .90$.

An analysis of error latencies was also conducted. These are reaction times of responses for which the

Figure 1. Mean correct identification rates (%) and standard deviations of word-initial fricatives (/f, θ, s, ʃ/) preceded by either a semantically congruous (white bars) or incongruous (shaded bars) precursor (Experiment 1).

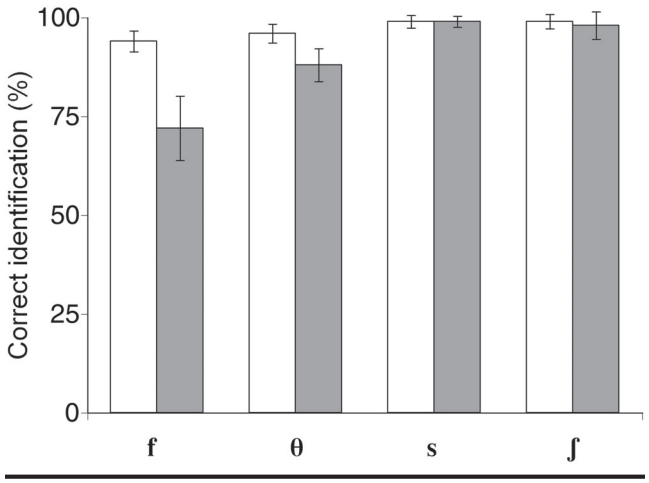
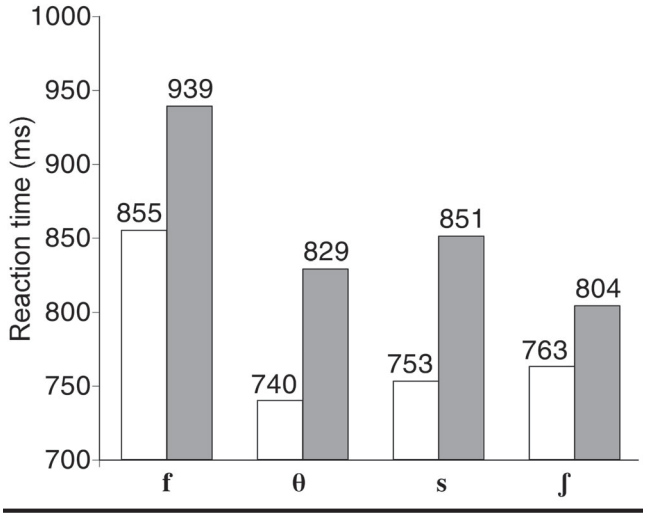


Figure 2. Response latencies for correctly identified word-initial fricatives (/f, θ, s, ʃ/) preceded by either a semantically congruous (white bars) or incongruous (shaded bars) precursor (Experiment 1).



intended fricative was misperceived as the other fricative of a minimal pair (e.g., *suit* perceived as *shoot*). A two-way (Context × Fricative) ANOVA revealed no main effects or interactions.

Discussion

Experiment 1 was designed to explore the effect of linguistic context on the perception of English fricatives. The role of stimulus robustness was investigated by comparing the perception of nonsibilant (acoustically ambiguous) and sibilant (acoustically robust) fricatives. Accuracy data indicated that context only affects perception of nonsibilant fricatives. Perception of /f, θ/ was more accurate when preceded by a semantically congruous precursor. Moreover, in terms of accuracy, perception of /s, ʃ/ was shown not to be affected by linguistic context. The fact that for /s, ʃ/ this context effect was only observed in the reaction time data and not in the accuracy data is most likely due to ceiling-level recognition scores for these fricatives. Because accuracy rates were high initially for the sibilants, the effect of context was not observable. An analysis of the reaction time data was more revealing. The reaction time data showed that context affected perception of both nonsibilant and sibilant fricatives. Perception of all fricatives was faster when preceded by a semantically congruous precursor as compared to an incongruous precursor. Context not only affected the more phonetically ambiguous stimuli /f/ and /θ/ but also was influential for the more robust, phonetically unambiguous stimuli /s/ and /ʃ/. Context appears to influence perception of all fricatives. Response latencies were analyzed for the errors as well. There were very few errors for /s/ and /ʃ/ and more errors for /θ/ and /f/. Despite differences in overall error rates, no effects of context were obtained in the error data.

Experiment 2: Contribution of Auditory and Visual Information

The results from Experiment 1 suggest that semantic context may aid in fricative perception. Another factor that has been implicated in consonant perception is visual information. Research on the role of visual information in consonant perception in persons with a hearing loss suggests that it may serve to differentiate place of articulation. Walden, Prosek, Montgomery, Scherr, and Jones (1977) studied the visual recognition of English initial consonants in consonant–vowel syllables (the vowel was always /ɑ/) in adults with a hearing loss. Walden et al. found that participants with a hearing loss distinguished five categories of consonants, known as *visemes*. Based on a criterion of at least 75% identification within the viseme, Walden et al. showed that five visemes could be distinguished: /fv, θð, pbm, szʃʒ, w/. These results show that /f/ and /v/ were often confused with each other, as were /θ/ and /ð/, but that the labiodental and dental fricative categories were rarely confused with each other. Similar results have been reported by other researchers (Benguerel & Pichora-Fuller, 1982; Binnie, Montgomery, & Jackson, 1974; Fisher, 1968).

In addition, a few studies have addressed similar issues in normal-hearing participants while exploring the McGurk effect (McGurk & MacDonald, 1976). Repp, Manuel, Liberman, and Studdert-Kennedy (1983, as discussed in Massaro, 1998) investigated perception of the syllables /bɑ, vɑ, ðɑ, dɑ/ produced by a female speaker. Repp et al. compared perception of these four syllables in conditions where the auditory and visual information was either consistent or conflicting. Results showed that listeners made no errors on consistent trials. On conflicting trials, auditory information seemed to contribute more than visual information to overall perception. Unfortunately, Repp et al. did not include any unimodal trials in which only auditory or visual information was presented, making it difficult to evaluate directly the relative importance of these two sources of information. Massaro (1998) investigated perception of the same four syllables using synthetic visible speech (Baldi) and natural audible speech. Trials consisted of consistent and conflicting bimodal stimuli, as well as unimodal auditory and visual stimuli. Consistent with Repp et al., performance on conflicting trials suggested a greater contribution of auditory rather than visual information. However, Massaro (1998) found a relatively smaller effect of visible speech than did Repp et al. Massaro (1998) attributed the discrepancy to a difference in quality between natural and synthetic visible speech. Finally, Massaro's (1998) results for the unimodal trials indicated that fricative recognition based on visible information is only slightly better than that based on only audible information.

Previous research with populations with hearing loss has suggested that visual information alone may be able to differentiate /f, v/ from /θ, ð/, while research in normal-hearing populations has suggested a greater contribution of auditory information when visual and auditory information are conflicting. The present study therefore investigates the role of natural visual information in the perception of these four fricatives by normal-hearing participants. Evaluation of the role of visual information in the perception of this class of nonsibilant fricatives is particularly important because they are highly confusable auditorily.

A general comparison of accuracy scores obtained for fricative perception in quiet (Balise & Diehl, 1994;

AQ2 Jongman, 1989; Jongman & Wang, submitted; Tomiak, 1990) with those obtained by Walden et al. (1977) with participants with hearing loss suggests that perception of nonsibilant fricatives in terms of place of articulation is better based on visual information alone than on auditory information alone. However, these auditory and visual studies cannot be compared directly because they differed in the questions they addressed and the methodologies they used. In the present study, the relative contributions of auditory, visual, and combined audiovisual information in nonsibilant fricative perception were determined in normal-hearing participants.

Method

Participants

Thirty listeners (10 per condition) were recruited from the Cornell University student population. Nineteen were women. All were native speakers of English and stated that they had no speech or hearing impairments, and all had normal or corrected-to-normal eye-
AQ5 sight. Participants were paid for their participation.

Stimuli

Stimuli consisted of 12 fricative-vowel syllables in which the fricatives /f, v, θ, ð/ were paired with each of the vowels /i, ɑ, u/. The stimuli were produced by the same speaker used in Experiment 1. The speaker was simultaneously audio- and video-recorded using a digital video camera (Sony Hi-8) in a soundproofed video studio at Cornell University's Noyes Language Learning Center. The speaker was seated in front of a neutral background and was illuminated by daylight-balanced studio lighting. The audio signal was recorded on one of the audio tracks of the Hi-8 video tape, using an external microphone (Electro-Voice RE20) placed at approximately a 45° angle and 15 cm away from the corner of the speaker's mouth. The speaker was video-recorded to capture the head from vertex to mandible within the video frame, thus excluding the microphone. The speaker began and ended each syllable with her lips closed. The speaker produced multiple repetitions of each syllable, from which one token of each syllable was then selected. All 12 stimuli had similar durations (approximately 800 ms).

Procedure

Three test conditions were contrasted: audiovisual, visual, and auditory. In the audiovisual experimental condition, participants watched the speaker's face on a 19-in. color TV monitor (Sony) and simultaneously heard her voice. Participants were seated at a desk in a dimly lit room at a comfortable viewing distance (130 cm) from the TV screen. Each trial started with a 500-ms trial-warning tone, followed by a black screen (1,000 ms), the speaker's neutral face (667 ms), and the target production (800 ms), and ending with a black screen (3,033 ms). The speech stimuli were placed on one audio track of the Hi-8 video tape, and the warning tones were placed on the other audio track. In the audio and audiovisual experiments, auditory information (both tracks) was provided through the loudspeakers of the TV monitor. The audio signal was presented at a comfortable listening level of approximately 70 dB SPL, measured for the peak intensity of the vowel at the approximate location of the listener's head. In all conditions, participants responded by circling 1 of the 13 alternatives: fi, fa, fu, vi, va, vu, thi, tha, thu, dhi, dha, dhu, or "other" provided on answer sheets. The alternatives th and dh were used to indicate /θ, ð/, respectively. Participants were asked to repeat a few words with /θ, ð/ in initial position to ensure that they were aware of the difference between these two sounds and of their correspondence to the response alternatives. In the visual condition, procedures were identical, except that the audio track with the speech stimuli did not accompany the video. Only the track with the warning tones was audible. In the audio condition, the TV monitor was simply disconnected and only the audio targets were presented.

All participants were tested individually. Trials were presented every 6 s. Five repetitions of each stimulus were presented in random order, yielding a total of 60 trials (4 fricatives × 3 vowels × 5 repetitions). Each test was preceded by a brief instruction period and by a series of practice trials to familiarize participants with the type of stimuli and the presentation rate.

Results

Participants identified both the fricative and the following vowel. Use of the "other" response category for fricative identification accounted for only 1.3% of responses. Vowel identification was very good, with no errors in both the audio + video and the audio conditions and with a 1.8% error rate in the video condition. Correct fricative identification scores are shown in Figure 3. A two-way (Modality × Fricative) ANOVA indicated a main effect for modality, $F(2, 11) = 84.59$, $p = .000$. Bonferroni post hoc tests for modality revealed that all three conditions were significantly different from each other: Fricative identification on the basis of both auditory and visual information (97%) was significantly better ($p = .024$) than that based on only auditory information (86%), which was in turn better than that based on visual information alone (48%, $p = .000$). A main effect was also found for fricative, $F(3, 11) = 3.42$, $p = .033$. Post hoc tests for fricative indicated that, across conditions, identification

Figure 3. Mean correct identification rates (%) and standard deviations of syllable-initial fricatives (/f, v, θ, ð/) on the basis of audio and video information combined (white bars), audio information only (light shaded bars), and video information only (dark shaded bars).

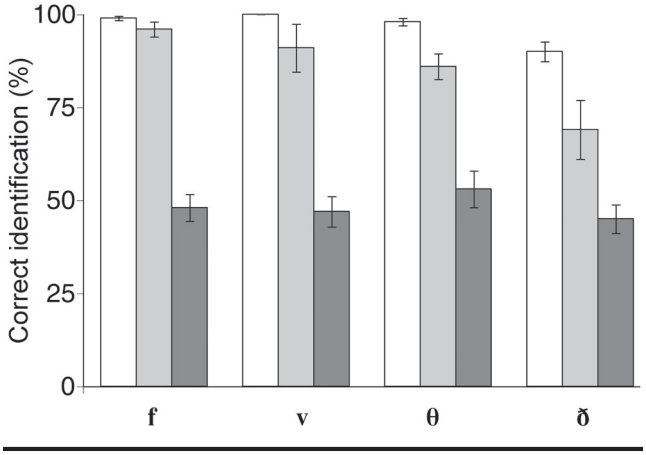
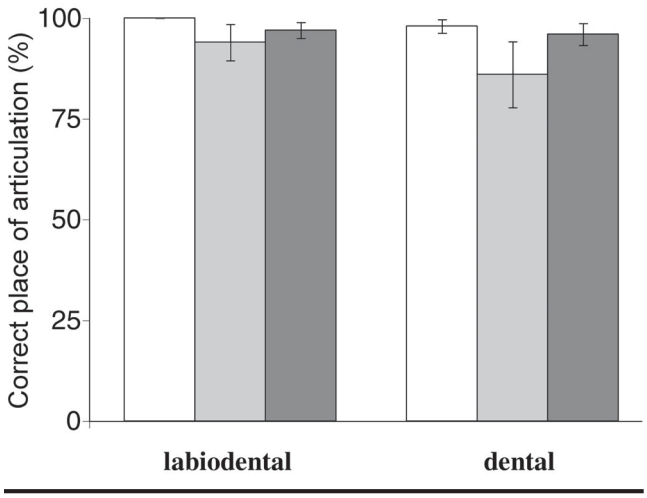


Figure 4. Mean correct identification rates (%) and standard deviations for place of articulation of syllable-initial fricatives on the basis of audio and video information combined (white bars), audio information only (light shaded bars), and video information only (dark shaded bars).



of /f/ (81%) was significantly better than identification of /ð/ (68%) and that there were no other significant differences. There was no significant Modality × Fricative interaction, $F(6, 11) = 0.97$, $p = .464$.

An analysis for Modality × Place of Articulation revealed similar main effects for modality, $F(2, 5) = 81.69$, $p = .000$, and for place, $F(1, 5) = 4.17$, $p = .049$. Across conditions, perception of labiodentals (80%) was significantly better than perception of dentals (73%). However, there was no significant Modality × Place interaction, $F(2, 5) = 2.28$, $p = .119$. Across conditions and fricatives, fricative identification was better in the context of /u/ (83%), than in the /a/ (75%) and /i/ (73%) contexts. This context effect, however, did not quite reach significance, as shown by the effect for vowel in a two-way (Modality × Vowel) ANOVA, $F(2, 8) = 3.30$, $p = .052$.

The poor scores in the visual information only condition may be due to the fact that facial information does not typically contain any cues to voicing (Fisher, 1968). The traditional visemes (/fv, θð, pbm, szʃʒ, w/) also suggest that voicing cannot be distinguished on the basis of visual information. Therefore, a second analysis was performed. Listeners' responses for all three conditions (audio, video, audio + video) were scored for correct identification for place of articulation regardless of voicing. For example, for an /f/ target, both /f/ and /v/ responses were considered correct.

Results for correct identification of place of articulation are shown in Figure 4. A two-way (Modality × Place of Articulation) ANOVA revealed a main effect for modality, $F(2, 5) = 3.79$, $p = .034$. Post hoc tests indicated that accuracy in the audio + video condition was as high as in the video condition ($p = .99$) and these two conditions were both significantly better than the audio condition ($p = .036$). There were no other significant main effects or interactions.

Discussion

Experiment 2 explored the effects of visual information on the perception of nonsibilant fricatives. Perception was quite accurate on the basis of simultaneous auditory and visual information. Perception based on auditory information alone was significantly worse, but was still reasonably accurate. Finally, fricative perception from visual information alone was quite poor. Although visual information did not seem to contain much information for fricative perception, visual information did make a perceptual contribution because fricatives were perceived more accurately with both auditory and visual information together than with auditory information alone.

When analyzing identification in terms of place of articulation, discounting voicing errors, a different picture emerged. Perception of fricatives was again highly accurate with simultaneous auditory and visual information. Interestingly, visual information alone now yielded accuracy rates comparable to the audio + video condition. Finally, perception based on auditory information was significantly worse. The present results for the audio + video condition are comparable to those reported by Repp et al. (1983) and Massaro (1998) for similar conditions. Comparison of the unimodal conditions indicates that perception of the dental place of articulation from visual information only was substantially

better in the present study (96%) than in Massaro (84%). This difference may be due to the use of natural facial information in the present study. Despite this difference, Massaro also showed that fricative perception is less accurate for auditory information alone relative to video information alone.

General Discussion

The present study explored the influence of linguistic context and the contribution of visual information on the perception of English fricatives. The effect of linguistic context on fricative perception was assessed in Experiment 1. Specifically, this experiment was designed to test the hypothesis that the influence of linguistic context is affected by the acoustic or perceptual salience of the stimulus. This study therefore compared perception of an acoustically murky contrast for place of articulation (/f/–/θ/) to an acoustically robust contrast (/s/–/ʃ/). In terms of accuracy, linguistic context was found to have an effect for contrasts that are not well-defined acoustically. That is, perception of /f/- and /θ/-initial target words was more accurate when preceded by a semantically congruous sentence relative to an incongruous sentence. In contrast, no effect of context was observed for /s/- and /ʃ/-initial words. Although the accuracy data suggest that recognition of words starting with /s/ and /ʃ/ was not affected by context, this may have been due to ceiling-level performance. The reaction-time data showed a facilitatory effect of semantically congruous context for target words starting with all four fricatives, indicating that semantic context affected recognition of even the acoustically robust distinction.

In a second experiment, an additional contextual variable, facial information, was examined. This information is often ignored in research on speech perception in normal-hearing participants. The influence of facial information has been well documented for consonant perception in populations with hearing loss (see Campbell, Dodd, & Burnham, 1998, for a recent review). Research in this area suggests that perception of the contrast between /f, v/ and /θ, ð/ should be possible on the basis of visual information alone. The contribution of visual (facial) information to perception of the nonsibilant fricatives /f, v, θ, ð/ was evaluated in Experiment 2. When the data were analyzed for correct identification for place of articulation, the results suggested that perception of nonsibilant fricatives is equally as good for visual information alone as for both auditory and visual information combined.

Although stimulus materials were provided by a single speaker, we are confident that the present results will generalize to other speakers of English. The speaker recorded in this study was 1 of 20 (10 women, 10 men) included in detailed acoustic (Jongman, Wayland, & Wong, 2000) and perceptual (Jongman & Wang, submitted) studies of English fricatives. Results from both studies indicated that this speaker was highly representative of this sample, suggesting that similar results would be obtained for other speakers. Research on the extent to which variability in facial features may affect speech reading suggests that similar results would be obtained with other speakers for the video materials used in Experiment 2 as well. Kricos and Lesner (1982) showed that although visemes do vary across speakers, the stimuli used in Experiment 2, labiodental /f, v/ and dental /θ, ð/, always constituted separate visemes. AQ2

With respect to G. A. Miller and Nicely's (1955) original claim, the present study showed that fricative perception is affected by both semantic and visual context. Much research on the role of context in the perception of speech has revolved around the heated debate about whether speech perception is a purely bottom–up process or whether top-down information can influence its outcome. Autonomous theories such as "shortlist" or "merge" (Norris, 1994; Norris, McQueen, & Cutler, 2000) suggest that prelexical phonological processing proceeds independently of lexical processing, whereas interactionist models like "cohort" or "TRACE" (Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986) propose that lexical context can influence phoneme perception. Although there is broad agreement that there is competition among candidates and that an integration of phonological, lexical, syntactic, and semantic information must occur, considerable controversy surrounds the specific details of that integration. One general model of perception that may be most likely to accommodate the present results is Massaro's (1987, 1998; Massaro & Cohen, 1991) fuzzy logical model of perception (FLMP). FLMP is a powerful model designed to account for perceptual results across a range of domains and tasks. In FLMP, different sources of perceptual information are evaluated independently. These sources are then integrated for perceptual decisions. FLMP has been shown to be capable of modeling the integration of a wide range of sources of information (Massaro, 1998). In the present study, Experiment 1 required the integration of lower level acoustic–phonetic information and higher level semantic information in phonetic decision making, while Experiment 2 required integration of auditory and visual information. Few models besides FLMP are comprehensive enough to address these issues given the range of these data. However, as Norris et al. (2000) pointed out, FLMP is a generic model of perceptual decision making, not a fully developed model of lexical access. Consequently, specific details of the model's implementation (e.g., the notion of lexical competition) are not fully available, making it difficult to determine the extent to which FLMP could successfully account for the current results.

The present data do suggest a number of insights into the nature of the context effects themselves. First, the present data support the notion that a number of sources of contextual influence may be operative. In Experiment 1, the ambiguity of the phonemic contrast itself did not affect the influence of the sentential context: The more robust phonetic contrast between the sibilant fricatives showed an effect of context similar to that of the more ambiguous nonsibilant contrast. This finding suggests that the listener applies the contextual information regardless of the robustness of the phoneme in question. Second, the results from Experiment 2 encourage a redefinition of what is typically included in contextual effects. In Experiment 2, facial information, as conveyed by visual evidence, provided a significant contribution to identifying place of articulation in fricatives. In sum, the current findings suggest that accurate perception of nonsibilant fricatives derives from a combination of acoustic, linguistic, and visual information. Although visual information has long been recognized as an important source of information for persons with hearing loss (see De Filippo & Sims, 1988, for a review of speech reading), the present results indicate that visual information also provides important information for identification of fricatives by normal-hearing adults. In this view, visual information is not merely a helpful cue to understanding speech in noisy situations—it is an integral part of speech perception.

AQ6

Acknowledgments

Portions of this research were conducted while the first author was at Cornell University. This research was supported (in part) by Research Grant 1 R29 DC 02537-01A1 from the National Institute on Deafness and Other Communication Disorders. The authors thank Scott Gargash, Eric Evans, Julie Allmayer, Diana Schenck, Michelle Spence, Michael Spivey, and Mike Tolomeo for assistance. Portions of this research were presented at the 137th meeting of the Acoustical Society of America, Berlin, March 1999.

References

Balise, R. R., & Diehl, R. L. (1994). Some distributional facts about fricatives and a perceptual explanation. *Phonetica*, *51*, 99–110.

Benguerel, A-P., & Pichora-Fuller, M. K. (1982). Coarticulation effects in lip reading. *Journal of Speech and Hearing Research, 25*, 600–607.

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research, 17*, 619–630.

Campbell, R., Dodd, B., & Burnham, D. (1998.). *Hearing by eye: II. Advances in the psychology of speechreading and auditory-visual speech.* East Sussex, U.K.: Psychology Press.

Charles-Luce, J. (1993). The effects of semantic context on voicing neutralization. *Phonetica, 50*, 28–44.

Cole, R. A. (1973). Listening for mispronunciations: A measure of what we hear during speech. *Perception & Psychophysics, 13*, 153–156.

Connine, C. M. (1987). Constraints on interactive processes in auditory word recognition: The role of sentence context. *Journal of Memory and Language, 16*, 527–538.

De Filippo, C. L., & Sims, D. G. (1988). New reflections on speechreading [edited monograph]. *The Volta Review, 90,* 3–313.

Erber, N. (1972). Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech and Hearing Research, 15*, 413–422.

AQ7

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research, 11*, 796–804.

Garnes, S., & Bond, Z. S. (1976). The relationship between semantic expectation and acoustic information. *Phonologica*, *3*, 285–293.

Grosjean, F. (1980). Spoken word recognition and the gating paradigm. *Perception & Psychophysics, 28*, 267–283.

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

Jongman, A., & Wang, Y. (XXXX). *Perceptual characteristics of English fricatives*. Manuscript submitted for publication.

AQ8

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

Kricos, P. B., & Lesner, S. A. (1982). Differences in visual intelligibility across talkers. *The Volta Review, 84,* 219–225.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, *244*, 522–523.

Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10,* 29–63.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Massaro, D. W., & Cohen, M. M. (1991). Integration versus interactive activation: The joint influence of stimulus and context in perception. *Cognitive Psychology, 23,* 558–614.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.

Mertus, J. (1989). *Bliss manual*. Providence, RI: Brown University.

Miller, G. A., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, *2*, 217–228.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27*, 338–352.

Miller, J. L., Green, K., & Schermer, T. (1984). A distinction between the effects of sentential speaking rate and semantic congruity on word identification. *Perception & Psychophysics*, *36*, 329–337.

Norris, D. (1994). SHORTLIST: A connectionist model of continuous speech recognition. *Cognition, 52,* 189–234.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral & Brain Sciences, 23*, 299–370.

Pisoni, D., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, *15*, 285–290.

Pollack, I., & Pickett, J. M. (1963). The intelligibility of excerpts from conversation. *Language and Speech*, *6*, 165–171.

AQ9 **Repp, B. H., Manuel, S. Y., Liberman, A. M., & Studdert-Kennedy, M.** (1983). Exploring the McGurk effect [Abstract]. *Bulletin of the Psychonomic Society,* 366–367.

Sawusch, J. R. (1996). Instrumentation and methodology for the study of speech perception. In N. J. Lass (Ed.), *Principles of experimental phonetics* (pp. XXX–XXX). St. Louis, MO: Mosby. AQ10

Sumby, W. H., & Pollack, I. (1954). Visual contributions to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212–215.

Tomiak, G. (1990). *An acoustic and perceptual analysis of the spectral moments invariant with voiceless fricative obstruents*. Unpublished doctoral dissertation, SUNY Buffalo.

Walden, B., Prosek, R., Montgomery, A., Scherr, C., & Jones, C. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research, 20*, 130–145.

Wang, M. D., & Bilger, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *Journal of the Acoustical Society of America*, *54*, 1248–1266. AQ11

Received March 1, 2002

Accepted May 5, 2003

DOI: 10.1044/1092-4388(2003/XXX)

Contact author: Allard Jongman, PhD, Linguistics Department, 412 Blake Hall, University of Kansas, Lawrence, KS 66045. E-mail: jongman@ku.edu

Appendix (p. 1 of 2). Full listing of target words and semantically congruous and incongruous contexts used in Experiment 1.

1. thirst/first
   The lemonade quenched my thirst
   The top swimmer came in first
2. thought/fought
   Amy sat still lost in thought
   In the ring the boxers fought
3. thread/Fred
   You sew with needle and thread
   Mr. Flintstone's name is Fred
4. three/free
   A triplet is made of three
   All of the slaves were set free
5. thrill/frill
   Roller coasters are a thrill
   The gown had a lacy frill
6. throws/froze
   Strikes are what a pitcher throws
   In December the lake froze
7. think/fink
   Humans are able to think
   A scoundrel is called a fink
8. Thor/four
   The god of thunder is Thor
   Seven equals three plus four
9. thin/fin
   A beanstalk is tall and thin
   The big goldfish moved its fin
10. threat/fret
    The man's words were a veiled threat
    When I'm sad mom says don't fret

Appendix (p. 2 of 2). Full listing of target words and semantically congruous and incongruous contexts used in Experiment 1.

1. shag/sag
   The rug she chose was a shag
   The old bridge began to sag
2. shed/said
   Long-haired cats are known to shed
   I misheard what Tom had said
3. shave/save
   The barber gave Ed a shave
   Ten bucks were all she could save
4. sheet/seat
   The bed has a new blue sheet
   The bus had one empty seat
5. shell/sell
   On the beach I found a shell
   Stolen goods are hard to sell
6. show/sew
   I watched the new TV show
   You need a needle to sew
7. shingle/single
   The roof lost another shingle
   If unmarried one is single
8. shock/sock
   His death came as quite a shock
   I took off my shoe and sock
9. shoot/suit
   The cop warned him not to shoot
   He wore a blue pinstripe suit
10. shore/sore
    Seagulls flock down by the shore
    Overworked muscles feel sore

Author Queries

AQ1: Harris (1958) and LaRiviere, Winitz, and Herriman (1975) do not have corresponding references in the ref. list. Please add them.

AQ2: For this "submitted" article, please instead provide the year of the draft you are referring to (or "in press" if that is now the case).

AQ3: Should "Rationale" be a subheading under "Experiment 1…" or should it be part of that heading?

AQ4: Please define VOT on first use.

AQ5: Please indicate how much the participants were paid.

AQ6: For clarity and parallel structure, I have slightly reworded the portion of this sentence after the colon. Are my changes okay?

AQ7: Erber (1972) is not cited in text. Please add a citation or delete this ref.

AQ8: If Jongman and Wang is now in press, please provide the journal name. If it is not, please provide the year of the draft you are referring to.

AQ9: For Repp et al. (1983), the citations I found on the Web give the page number as 358 instead of 366–367 and also indicate that it is an abstract. Are my changes correct?

AQ10: Please provide chapter page numbers for Sawusch (1996).

AQ11: Wang and Bilger (1973) is not cited. Please add a citation or delete this ref.