kansas working papers in linguistics

volume 2 1977

Laurel Watkins and Ginny Gathercole

Editors

Cover design by Jeanette Gunn

Funded by Student Activity Fees
University of Kansas
Lawrence, Kansas
1977

The editors are pleased to present this second collection of papers from the Linguistics Department at the University of Kansas. In preparing this issue, we have been aided in many ways by members of the faculty and by our department secretary, Ruth Hillers. We wish to express our appreciation for their kind assistance. We are also grateful to Jeanette Gunn for her work on the cover page.

CONTENTS

A STUDY OF SPEAKER SEX IDENTIFICATION

Ronald P. Schaefer

Introduction

1. The identification of speaker sex from voice samples has been of theoretical and practical interest in the recent past. From a theoretical perspective, it has been of interest because of the desire to know the specific acoustic cues which make sex identification possible. From a more practical perspective, speaker identification has been of interest because of its potential role in law enforcement and voice therapy. Almost all experimentation to date has been conducted with native speakers of English and with isolated productions of vowels or consonants. The experiment described herein was conducted to determine if these two experimental components have an effect on speaker sex identification. Before proceeding to the experiment, it may be beneficial to take note of some of the major research regarding speaker sex identification.

A pioneering experiment with regard to the general area of speaker identification was Pollack, Pickett and Sumby (1954). Using the same set of male speakers and listeners as experimental subjects, they discovered a 30% increment in correct identification of the speaker with each .10 second increment in the sample of continuous speech presented to the listeners. They also discovered that mere repetition of a particular speech sample did not increase the number of correct identifications. Pollack et al. (1954) therefore concluded that the speaker's entire speech repertoire afforded cues for identification and that as more information

Lightner's (1965) *morpheme features* which incorporates harmony features into a generative account of the phonology of Mongolian.

Bibliography

Adiwidjaja, R.I. *Adegan Basa Sunda*, J.B.Wolters, Djakarta: 1951.

Anderson, S.R. "On Nasalization in Sundanese," *Linguistic Inquiry* 3, 253-268.

Chomsky,N and M.Halle (1968) *The Sound Pattern of English*, Harper and Row, New York.

Howard,I. (1971) "On Some Problems of Formalization in Generative Phonology", unpublished paper, M.I.T.

Langendoen,D.T. (1968) *The London School of Linguistics*, M.I.T. Press, Cambridge, Massachusets.

Lightner, Theodore M. (1965) "On the Description of Vowel and Consonant Harmony", *Word* 21. pp. 244-250.

Pavlenko, A.P. *Sundanskii Iazyk*, Moscow 1965.

Robins, R.H. (1953a) "The Phonology of the Nasalized Verbal Forms in Sundanese." In *Bulletin of the School of Oriental and African Studies* 15, 138-145.

__________ (1953b) "Formal divisions in Sundanese". *Transactions of the Philological Society*, 133-142.

__________ (1957) "Vowel Nasality in Sundanese," In *Studies in Linguistic Analysis*, 87-103.

__________ (1959) "Nominal and Verbal Derivation in Sundanese". In *Lingua* 8, 337-369.

about this repertoire was made available, the more accurate was the speaker identification.

Speaker sex identification received detailed attention in a number of experiments which demonstrated that listeners could with greater or lesser accuracy identify speaker sex based on productions of continuous speech samples and isolated vowel samples. Schwartz and Rine (1968) employed whispered productions of the isolated vowels [a] and [i] to determine if the absence of vowel funadmental frequency, due to the whispered condition, affected speaker sex identification. Their results were overwhelming; they found only 4 incorrect identification responses out of a total of 160 responses. After examination of the similarly patterned spectra of the male and female produced vowels, Schwartz and Rine (1969) concluded that the upward displacement of frequency in the spectra of the female-produced vowel relative to the male-produced vowel accounted for the correct identifications.

Weinberg and Bennett (1972) also wished to determine the importance of fundamental frequency to speaker sex identification. Instead of employing the whispered vowel experimental condition as in Schwartz and Rine (1968), Weinberg and Bennett (1972) employed a procedure in which they used 30 seconds of continuous speech from 66 normal-developing 5- and 6-year-old children. For each child's speech sample they computed the mean fundamental frequency and found that there was no difference in fundamental frequency between the male and female samples. However, adults were able to identify the sex of the children in the absence of a sex distinctive fundamental frequency 75% of the time. Weinberg and

Bennett's (1972) findings are thus in agreement with Schwartz and Rine (1968). These findings suggest that the dimensional characteristics of the resonance cavity, which determine the spectra frequency, are the primary factors in sex identification.

In contrast to the findings of Schwartz and Rine (1968) and of Weinberg and Bennett (1972), Lass, Hughes, Bowyer, Waters and Bourne (1976) found that fundamental frequency does play a more important role in speaker sex identification than had been previously thought. Specifically, Lass et al. (1976) found that the normal vowel condition received the highest percentage of correct listener judgments, 96%, that the filtered condition received 91% correct judgments and the whispered condition 75%. The resonance characteristics of vowels may provide cues for speaker sex identification, but the sex distinctive fundamental frequency may provide a more important cue.

There does appear to be agreement that fundamental frequency does play a role in the identification of speaker sex. However, there is a line of research in which speaker sex can be identified even when fundamental frequency is not present. This second line of research relies on the particular acoustic properties of some voiceless fricative consonants to indicate speaker sex.

Schwartz (1968) employed the four voiceless fricatives [f θ s ʃ] produced in isolation as his primary stimuli. He found that [s and ʃ] identified speaker sex more often than chance would allow, but that [f] and [θ] did not. What cues were listeners using to identify the sex of the speakers of [s] and [ʃ]? Schwartz (1968, p. 1179) provided an explanation consonant with the earlier explanation for sex identification based on vowel samples:

> Furthermore, spectrographic analysis of [s] and [ʃ] indicate that the acoustical features that underlie this ability are general--i.e., the female spectra tend not only to be higher in frequency than the male but parallel as well. This higher-frequency displacement is likely a manifestation of the smaller dimensions of the vocal tract in females.

In addition, Schwartz (1968) pointed out that the spectra of the fricatives[f] and [θ] were rather flat and broadband, suggesting therefore that the higher frequency was not transmitted with these vowels.

In general agreement with Schwartz (1968), Ingemann (1968) also concluded that speaker sex could be identified by means of some voiceless fricatives. Ingemann (1968) employed the entire continuum of 9 voiceless fricatives from the back fricative [h] to the front bilabial [ɸ], each produced in isolation. Moreover, she provides an hypothesis into which fit Schwartz's (1968) general conclusions.

In contrast to the general findings of Schwartz (1968), Ingemann's (1968, p. 1144) were more specific and she accordingly hypothesized:

> As the portion of the vocal tract in front of the constriction diminishes, so does the identification of the speaker's sex.

Thus the sex of the speaker of the back fricative [h] was judged correctly 91% of the time and the sex of the speaker of [ɸ] 55%. There is some discrepancy between the findings of Ingemann (1968) and Schwartz (1968) in that the sex of a speaker of [ʃ] was judged correctly 77% of the time for Ingemann and 90% for Schwartz. Despite this discrepancy, listeners were still able to identify speaker sex based on the acoustic information provided by the voiceless fricative [ʃ].

One can conclude from the brief descriptions above that the identification of speaker sex is a general and consistent phenomenon. Native speakers of English have been able to identify speaker sex based on productions of some isolated voiceless fricatives. However, can native speakers of English also identify the speaker sex of a non-native speaker of English? In turn, can non-native speakers of English identify the sex of native speakers of English? These two questions have not been answered to date. Moreover, can listeners identify speaker sex based on sex-identifying voiceless fricatives produced in context as well as the same fricatives produced in isolation? The following experiment was designed to test these three questions.

METHOD

Speakers

2.1 A total of 20 speakers participated in the experiment. 10 were native speakers of Japanese, 5 male Japanese speakers and 5 female Japanese speakers. The other 10 were native speakers of American English, 5 male American English speakers and 5 female American English speakers. All of the speakers were students at the University of Kansas. The Japanese speakers were between the ages of 19 and 26 and the American English speakers between the ages of 20 and 33.

Auditory Stimuli

2.2 The word she produced by each of the 20 speakers served as the basis for the three conditions of the experiment. From each speaker's sample of she, three samples which were to serve for sex identification purposes were made. One sample consisted of the entire word she, the second consisted of the vowel [i] and the third consisted of the voiceless fricative that initiated the word she. She was recorded from each of the 20 speakers in a sound treated room using a UHER Model 4000 tape recorder and a condenser microphone. Each speaker's intensity was monitored by means of a VU meter on the tape recorder. In addition, the distance between the microphone and the mouth cavity was held constant through manipulation of a head-stand.

She was chosen because the segment [ʃ] exhibits a point of constriction that would allow resonance information to be transmitted from speaker to listener. She is also composed of two segments that could be produced without difficulty by both the Japanese and the American English speakers.

Construction of Master Tape

2.3 One master tape was constructed which was composed of the three experimental samples:complete syllable, excised fricative and excised vowel. These three samples were obtained by gating the portion of the word she desired for each particular sample and retaining the desired portion on another tape. The gating mechanism consisted of a switching mechanism which allowed a predetermined portion of the acoustic signal to pass through to the output, while the gated or excised portion was either grounded or monitored on a separate output.

To gate the fricative from the word she, a tape loop of each she sample was made and placed on a tape recorder whose output was connected to the gating mechanism. A pulse was placed on the second channel of this tape. This pulse activated a variable timing mechanism which would allow a portion of the acoustic signal to pass through to the dubbing recorder. When the time limit set on the variable time mechanism was reached, a switch was activated which shunted the remaining signal to a ground or a monitor output.

The tape loop with a pulse on the second channel was played repeatedly while time adjustments were made and until the desired excised or gated portion was obtained. This gating procedure was monitored visually (on a storage oscilloscope) as well as orally to ensure that the desired sample was obtained and that the formant transition information was not included as a part of the experimental sample. Only steady state portions of the fricative and vowel were used. Once the desired sample had been gated, it was dubbed onto a master tape. This procedure was repeated for each of the fricative and vowel samples in the study.

The gating procedure also allowed for the maintenance of a constant duration for the excised fricative and the excised vowel. The duration of the excised fricative was maintained at approximately 150m/sec. The duration of the excised vowel was maintained at approximately 200m/sec. The duration of the excised fricative was the minimum duration found after submitting all complete syllables to wide band spectrographic analysis on a Model 6061B Sound Spectrograph.

Within each experimental condition (fricative, vowel, complete syllable) of the master tape, the 20 stimuli were presented in random order. In order to ensure reliability, the same 20 stimuli were presented in a second random order. There were thus a total of 40 stimuli in each experimental condition.

In order to allow adequate time for listener judgments, there was a 5-second silent interval between all of the stimulus items in each experimental condition. The stimuli were presented in sets of 5 with

25 seconds in between sets. In this fashion the listeners could keep track of the presentation order since they were instructed that after every 5 stimuli there would be a 25 second pause. This pause was chosen rather than numbering with a speaker's voice in order to control for any potential effects of the number-speaking voice on the judgment of the subsequent experimental sample.

Listening Sessions

2.4 20 total listeners participated in the experiment, 10 of the listeners were native speakers of Japanese, 5 male Japanese speakers and 5 female Japanese speakers. The other 10 listeners were native speakers of American English, 5 male American English speakers and 5 female American English speakers. All listeners were students at the University of Kansas and none reported hearing impairments of any kind.

The master tape was presented binaurally by means of a Roberts Model 770XX SS tape recorder and individual headphone sets. All listening sessions were held in a sound treated room at the University of Kansas.

Each listener was given a numbered response sheet on which he circled the letter M or F, corresponding to male and female respectively. To prepare listeners for the stimuli of each experimental condition, the first five stimuli of that condition were played. After the first five stimuli were played, the entire set of 40 stimuli in the experimental condition were played without stopping.

RESULTS AND DISCUSSION

The primary purpose of this study was to determine whether listeners from different language backgrounds could identify speaker sex from selected speech samples. The results show that in each of the three conditions listeners were able to identify speaker sex at a better than chance level. The specific results and their implications are presented in the following paragraphs.

Listener Judgments

3.1 It was found that of the 800 (40 stimuli x 20 listeners) speaker sex identification judgments, 793 (99.1%) were correct in the complete syllable condition. It was also found that of the 800 speaker sex identification judgments, 797 (99.6%) were correct in the excised vowel condition. Finally, it was found that of the 800 speaker sex identification judgments in the excised fricative conditions, 630 (78.7%) were correct.

For each of the three experimental conditions, a breakdown of the data according to the sex of the listener provides the following results. Male listeners correctly identified the samples in the three experimental conditions at the following percentage rates: complete syllable 99.5%, excised vowel 100%, excised fricative 75%. Female listeners correctly identified the samples in the three experimental conditions at the following percentage rates: complete syllable 98.8%, excised vowel 99.3%, and the excised fricative 82%.

For each of the three experimental conditions, a breakdown of the data according to the sex and language background of the listeners provides the following results. Male Japanese listeners correctly identified the samples in the three experimental conditions at the following percentage rates: complete syllable 100%, excised vowel 100% and excised fricative 85%. Female Japanese listeners correctly identified the samples in the three experimental conditions at the following percentage rates: complete syllable 95%, excised vowel 95% and excised fricative 82%. Male American English listeners correctly identified the samples in the experimental conditions at the following percentage rates: complete syllable 99%, excised vowel 100% and excised fricative 66%. Finally, the female American English listeners correctly identified the samples in the three experimental conditions at the following percentage rates: complete syllable 98%, excised vowel 99% and excised fricative 82%. As the above data indicate, all listeners, regardless of category (male or female, Japanese or English) did exceptionally well in identifying the sex of a speaker from complete syllable, excised vowel and excised fricative samples.

The primary acoustic cue accounting for the consistency in identifying the excised vowel samples, and the complete syllable samples, as strongly suggested by Lass et al. (1976), appears to be a sex distinctive fundamental frequency. In order to verify the significance of fundamental frequency for speaker sex identification, each speaker's production of the word she was submitted to frequency analysis by a Honeywell Visicorder Model 1508A. After the determination of fundamental frequency for the vowel [i] of each speaker, it was found that vowels with a fundamental frequency of 120 Hz or below were identified consistently as male, and vowels with a fundamental frequency of 220 Hz or above were identified as female. These frequency levels are in agreement with Lass et al. (1976) who found a mean fundamental frequency for male-produced vowels of 111.43 Hz and a mean fundamental frequency for female-produced vowels of 224.08 Hz.

Listener Judgments in the Fricative Condition

3.2 The primary focus of this study was the excised fricative condition. It was found that of the 800 speaker sex identification judgments in the excised fricative condition, 630 (78.7%) were correct. This percentage correct figure is more in line with Ingemann's (1968) finding that [ʃ] received 77% correct identification than Schwartz's (1968) finding that [ʃ] received 90% correct identification. The percentage of correct identifications in the excised fricative condition is also somewhat lower than that found in the excised vowel and complete syllable conditions, but the findings obtained from each of the conditions are statistically significant. Using a table of binomial probabilities all judgments exceeded chance probability (given a .5 a priori chance probability) at the .05 level of confidence.

The total percentage of correct identifications in the excised fricative condition is somewhat over-simplified, however. Figure 1 illustrates the percentage of correct speaker sex identifications for the 4 categories of listeners. Figure 1 indicates that male American English listeners had the greatest difficulty identifying speaker sex.

Table 2, which illustrates the number of correct sex identification judgments made by each of the four categories of listeners for each of the four categories of speakers, clearly shows that the male American English listeners did not have equal difficulty identifying the sex of all four speaker categories. As Table 2 shows, male American English listeners had consistent difficulty identifying male American English speakers and male Japanese speakers. Moreover, since the number of errors ranged between 10 and 17 for male American English listeners, the difficulty seemed to be relatively uniform for all members of this category. Thus male American English listeners, not male listeners in general, had the most difficulty identifying male voices in the excised fricative condition. There was no evidence of a similar difficulty in either the excised vowel or the complete syllable condition.

That male American English listeners had difficulty identifying speaker sex is not reported in previous experimentation on sex identification by means of voiceless fricatives. Schwartz (1968) made no mention of a sex determined sex identification difficulty. Ingemann (1968), who used 5 male and 5 female listeners, states quite explicitly that neither male nor female listeners perceived sex differences better than the other and that neither the male nor female voice was easier to identify. However, Ingemann (1968) implies that her experimentation was conducted in Scotland and therefore was conducted with British-born listeners. If this is the case, then either a cultural factor may have constrained the identification abilities of male American English listeners or a sampling factor in the particular group of male American English listeners in this experiment may have influenced the findings.

Acoustic Cues for Sex Judgments

3.3 Despite the particular sex identification difficulty of male American English listeners, who still correctly identified speaker sex at a better than chance level, it is a general finding of this study that speaker sex can be identified by means of the excised fricative [ʃ]. In accordance with Schwartz (1968), it might be expected that the spectra of a female-produced [ʃ] exhibit a higher frequency displacement relative to a male-produced [ʃ]. Therefore spectra data were obtained for a selected number of [ʃ] samples used in the experiment. The spectra data were obtained by submitting a central point in the duration of the selected fricative sample to narrow band section analysis on a 6061 B Sound spectrograph.

As one examines A and B in both Figure 1 and Figure 2, which are tracings obtained from the spectra of a selected number of [ʃ] samples, one observes a tendency for the female spectra in Figure 2-A and 2-B to exhibit peaks of energy prominence above the 4000 Hz level that the male spectra, 1-A and 1-B, do not exhibit. However, as the peak of energy prominence in 1-A (male) suggests, a simple displacement of energy toward the higher frequencies may not be sufficient to identify a female voice.

Consider A and B in Figure 1 and Figure 2 again. These samples were selected since they received the greatest percentage of correct identification of male and female samples in the excised fricative condition: 1-A 90%, 1-B 92%, 2-A 100%, 2-B 92%. After examining closely the male spectra in Figure 1, one finds that in both samples the peaks of energy prominence occur between 500 and 4000 Hz. In the female spectra of Figure 2, in contrast, it appears that the peaks of energy prominence occur between 2000 and 6000 Hz with the major peaks occurring just above the 4000 Hz level. It is important to note that although 1-A exhibits its major peak of energy prominence between 2000 and 6000 Hz, as in the female spectra, it also exhibits peaks of energy prominence between 500 and 2000 Hz that the female spectra do not exhibit.

There are two conclusions one might draw from a comparison of these two sets of spectra. First one might conclude that a peak of energy prominence in the low energy region of 500-2000 Hz may be an important cue in identifying a male voice which also exhibits peaks of energy prominence in the high energy regions, particularly above the 4000 Hz level. Second, one might conclude that a major peak of energy prominence above the 4000 Hz level serves as an important cue in identifying the female voice.

Given these tentative conclusions, consider C in Figure 1 and Figure 2. C in both figures was obtained from [ʃ] samples that were consistently misidentified by listeners in the experiment. 2 -C was obtained from a female sample that received 52% correct identification.

In both 1-C and 2-C, notice the major peak of energy prominence. 1-C, which should have been identified as male, exhibits its major peak of prominence near the 4000 Hz level and no peaks of energy prominence between the 500 and 2000 Hz region. Since low energy cues may be important in identifying male voices with high energy cues, the fact that 1-C was not consistently identified as a male voice is understandable. 2-C also appears to be lacking an important cue that would have allowed it to be identified as a female voice. In particular, the major peak of prominence in 2-C occurs between 2000 and 4000 Hz, not between 4000 and 6000 Hz, as in the other samples identified as female.

One additional [ʃ] sample illustrates that the cues for sex identification are more complex than those tentatively suggested here. The spectra 2-D, which was obtained from a female sample, received only 40% correct identification. Given the tentative conclusions regarding the acoustic cues of [ʃ] which allow sex identification, one notices in 2-D that peaks of energy prominence are present in the high energy region of 4000 and 6000 Hz. But, the amplitude of these peaks of energy prominence is considerably reduced relative to other amplitude levels in Figure 1 and Figure 2.

Such a reduced amplitude level in a sample which was consistently misidentified suggests that amplitude may interact in a particular manner with the placement of frequency to determine sex identification. It may be profitable to pursue further the relationship of amplitude levels of the voiceless fricative [ʃ] to speaker sex identification. With particular reference to 2-D it would be important to determine the specific amplitude level which would allow correct speaker sex identification to occur.

Conclusions

3.4 In conclusion, this study has found that the excised fricative [ʃ] allows speaker sex identification at a better than chance level. In addition this study has found that speaker sex identification by means of the voiceless fricative [ʃ] is possible for native and non-native speakers of English. Further, as expected, listeners in all categories ~~made~~ relatively few errors on speaker sex judgments obtained from vowel or complete syllable samples. Speaking fundamental frequency can account for most, if not all, of the correct judgments of these syllable and vowel samples. Taken together with the data from the fricative condition, these data suggest that the listener has available to him/her a variety of acoustic cues on which to make speaker sex judgments. Primary among these cues seems to be speaking fundamental frequency and resonance characteristics. Undoubtedly, as the literature suggests, these cues interact and provide the listeners with redundant information in cases such as she where both types of cues are operative. In any case, the data from the present study do support the claim that listeners can extract resonance cues from the fricative [ʃ] and use this information to allow them to make judgments regarding the sex of a speaker.

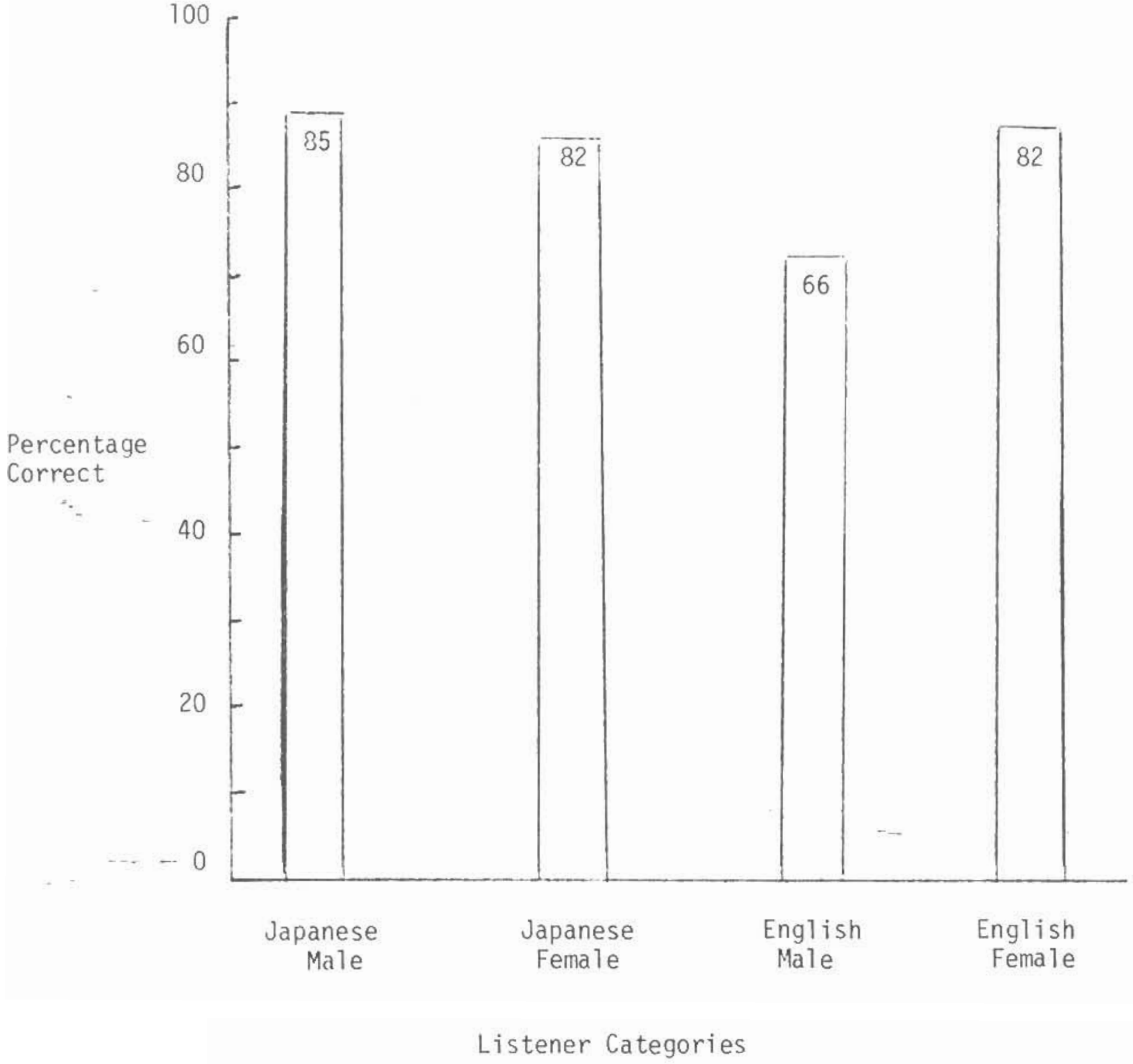

Table 1. Percentage of correct speaker sex identification judgments in the excised fricative condition for each of the four listener categories.

	Listener Category			
Speaker Category	Japanese Male	Japanese Female	English Male	English Female
Japanese Male	42	39	28	36
Japanese Female	40	43	35	44
English Male	43	45	29	45
English Female	44	38	40	39

Table 2. The number of correct sex identification judgments in the excised fricative condition made by each of the listener categories for each of the speaker categories. 50 possible judgments in each category.

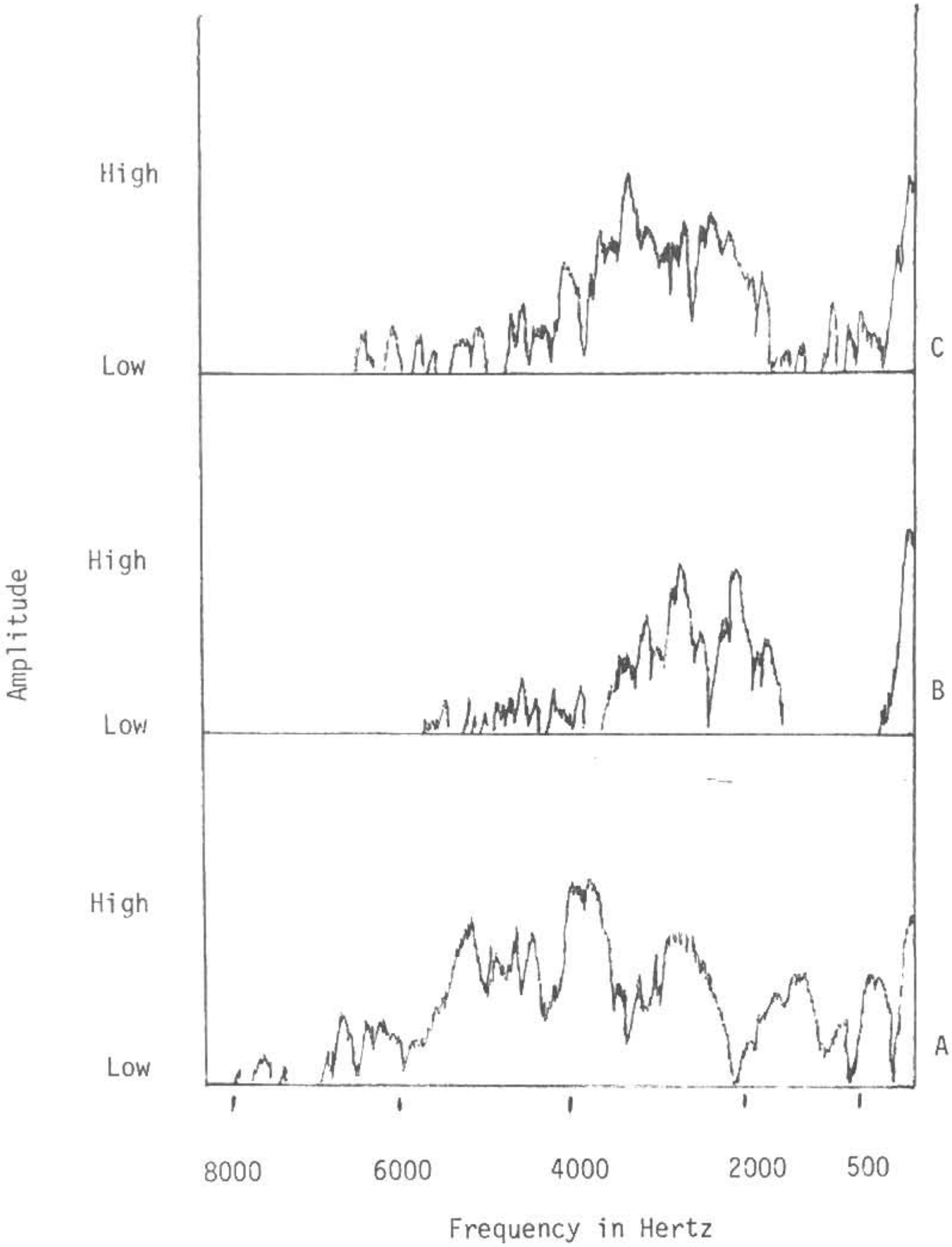

Figure 1. Tracings from the spectra of selected samples of [ʃ] produced by male speakers.

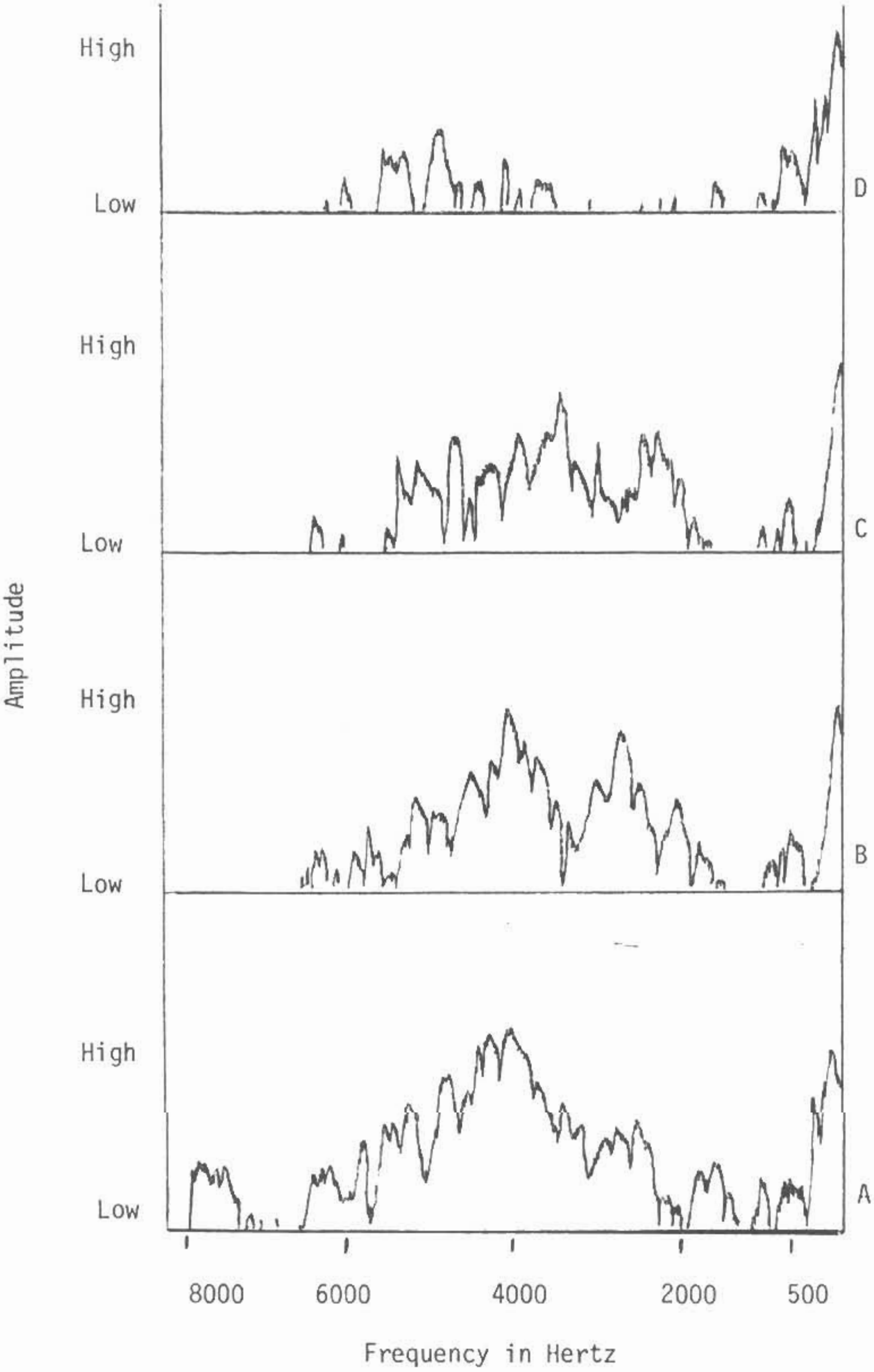

Figure 2. Tracings from the spectra of selected samples of [ʃ] produced by female speakers.

REFERENCES

Coleman, R. 1971. Male and female voice quality and its relationship to vowel formant frequencies. Journal of Speech and Hearing Research, 14, 565-577.

Fant, Gunnar. 1960. Acoustic Theory of Speech Production. The Hague: Mouton.

Ingemann, F. 1968. Identification of the speaker's sex from voiceless fricatives. Journal of the Acoustical Society of America. 44, 1142-1144.

Lass, Norman, J., Hughes, K.R., Bowyer, M.D., Waters, L.T., and Bourne, V.T. 1976. Speaker sex identification from voiced, whispered, and filtered isolated vowels. Journal of the Acoustical Society of America. 59, 675-678.

Pollack, I., Pickett, J., and Sumby, N. 1954. On the identification of speakers by voice. Journal of the Acoustical Society of America. 26, 403-406.

Schwartz, M. 1968. Identification of speaker sex from isolated voiceless fricatives. Journal of the Acoustical Society of America. 43, 1178-1179.

Schwartz, M., and Rine, H. 1968. Identification of speaker sex from isolated whispered vowels. Journal of the Acoustical Society of America. 44, 1736-1737.

Stevens, K., and House, A. 1961. An acoustical theory of vowel production and some of its implications. Journal of Speech and Hearing Research. 4, 303-320.

Weinberg, B. and Bennett, S. 1972. Speaker sex recognition of 5- and 6-year-old children's voices. Journal of the Acoustical Society of America. 50, 1210-1213.