Spectral Properties of Fricative Consonants

George W. Hughes, and Morris Halle

Citation: The Journal of the Acoustical Society of America **28**, 303 (1956); doi: 10.1121/1.1908271

View online: https://doi.org/10.1121/1.1908271

View Table of Contents: http://asa.scitation.org/toc/jas/28/2

Published by the Acoustical Society of America

single presentation of each word and mean scores for tests in which each word was presented four successive times are significant.

In view of the findings of the foregoing investigations and the evidence of the present study, it can be concluded with considerable confidence that repetition *per se*, that is, repetition which involves a mere duplication of the initial presentation of the test word, raises the articulation score, but only to a slight extent. The major portion of the improvement in articulation score which results from repeating the initial presentation of a test word occurs with the second presentation; third and fourth presentations have a negligible effect.

THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA VOLUME 28, NUMBER 2 MARCH, 1956

Spectral Properties of Fricative Consonants*

George W. Hughes and Morris Halle
Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge 39, Massachusetts

(Received August 5, 1955)

Energy density spectra of gated segments of fricative consonants were measured. The spectral data were used as a basis for developing objective identification criteria which yielded fair results when tested. As a further check gated segments of fricatives were presented for identification to a group of listeners and their responses evaluated in terms of the objective identification criteria.

STANDARD English has the following fricative consonants: |*f*| as in "leaf," |*s*| as in "lease," |ʃ| as in "leash," |*v*| as in "leave," |*z*| as in "Lee's," |ʒ| as in "rouge," |θ| as in "teeth," and |ð| as in "seethe." These consonants can normally be distinguished by English speakers in identical phonetic contexts, regardless of whether these contexts are meaningful utterances of English or are nonsense syllables. It follows, therefore, that the cues on which this differentiation is based can only reside in the acoustical stimulus. The purpose of our investigation was to establish what cues are contained in the spectra of the fricative portions taken in isolation.[1] We have investigated in detail all but |θ| and |ð|, since we believe that a solution of this problem will come only after the mechanism involved in their production is more fully understood. A few sample spectra of these two fricatives are given in Fig. 1.

PROCEDURE

A master tape was prepared by recording a number of English speakers, both male and female, reading a list of isolated words. The list was so designed as to place all fricatives in contexts before and after the major classes of vowels. Tape loops containing one word each were recorded from the master tape. Care was taken to assure a high signal-to-noise ratio and a wide frequency response.

The fricative portion of a syllable or word can be easily located and isolated by observing the oscilloscope trace, examples of which are shown in Fig. 2. For our purposes we had to be able to select a segment of any length at any specified time in the sample. In order to do this we recorded on the loops just before the sample a 10 kc tone, which was used to trigger an electronically controlled gate whose position and duration could be separately and accurately adjusted. The gate control

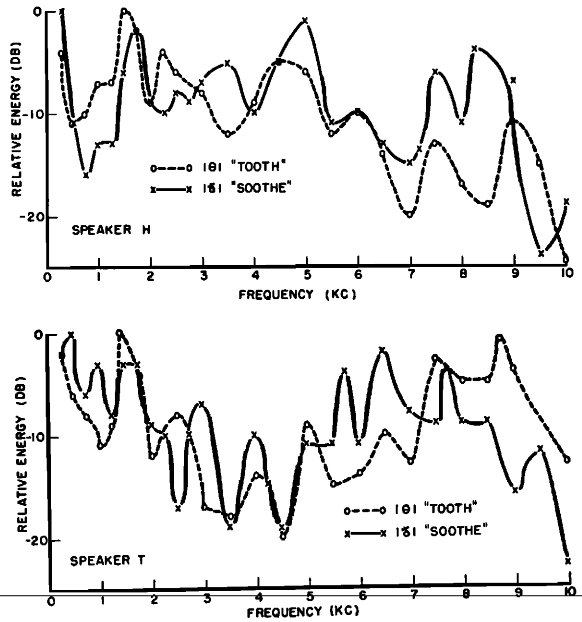

FIG. 1. Energy density spectra of the fricatives |θ| and |ð|.

* This work was supported in part by the Signal Corps, the Office of Scientific Research (Air Research and Development Command), the Office of Naval Research, and the National Science Foundation.

[1] Available evidence indicates that the primary cues are contained in the fricative portion; see K. S. Harris "Cues for the identification of the fricatives of American English," J. Acoust. Soc. Am. **26**, 952 (1954), where it is shown that except for the differentiation between |*f*| and |θ|, the transitions of the formants in the adjacent vowels contribute little towards the identification of the fricatives. For a summary of the literature on fricatives see T. Tarnoczy "Die akustische Struktur der stimmlosen Engelaute," Acta linguistica, (Budapest) **4**, 313–349 (1954).

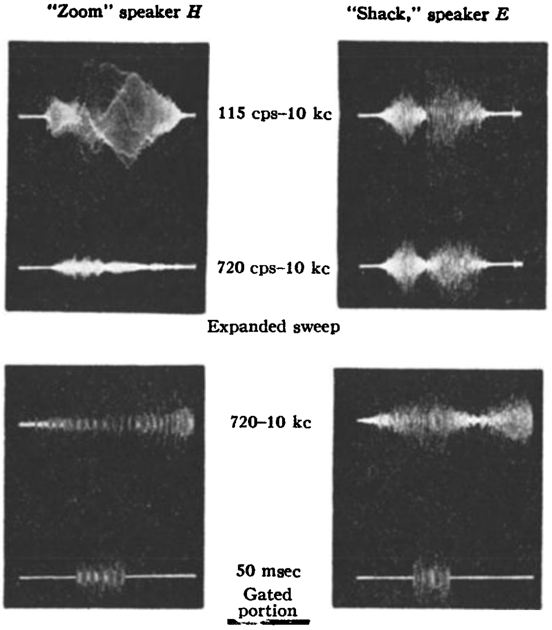

FIG. 2. Oscillograms of two words showing a typical placement of the gate. Note that the unvoiced |ʃ| in "shack" is almost entirely unaffected by the elimination of frequencies below 720 cps.

circuits were such that the gate position settings could be noted down and the same segment reproduced to within ±2 to 3 msec for any number of measurements. This method was adequate, for when measurements of some sounds were repeated several months after the initial study, the two sets of data showed no significant discrepancies. The bottom part of Fig. 2 shows how the gate was adjusted in a typical case. All spectral data on fricatives were taken with a gate length of 50 msec.

Energy density spectra of the gated fricative segments were measured by means of a fixed band-width filter whose center frequency was continuously variable over the range from 300 to 10 000 cps. A Hewlett Packard wave analyzer modified to have a band width of approximately 150 cps was used.

The output of the wave analyzer was amplified, squared or full wave rectified, integrated, and the resulting dc voltage fed to a holding circuit and meter. The meter readings were made to be the same for all settings of the filter frequency by adjusting a precision calibrated attenuator in the amplifier. The relative energy values in db were taken from the readings of this attenuator. A block diagram of the measurement system is shown in Fig. 3. Under this system which insured a relatively constant input voltage into the squarer or rectifier, fricative segments gave almost identical results using either an accurate squaring circuit or a full-wave rectifier. The results reported here were obtained employing the rectifier.

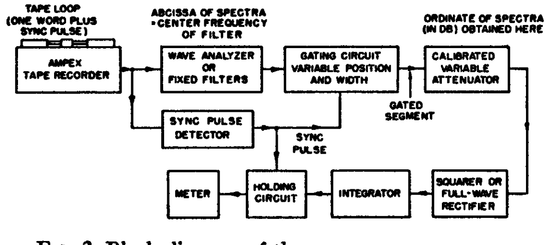
TAPE LOOP (ONE WORD PLUS SYNC PULSE)
AMPEX TAPE RECORDER
ABCISSA OF SPECTRA =CENTER FREQUENCY OF FILTER
WAVE ANALYZER OR FIXED FILTERS
GATING CIRCUIT VARIABLE POSITION AND WIDTH
ORDINATE OF SPECTRA (IN DB) OBTAINED HERE
CALIBRATED VARIABLE ATTENUATOR
GATED SEGMENT
SYNC PULSE DETECTOR
SYNC PULSE
METER
HOLDING CIRCUIT
INTEGRATOR
SQUARER OR FULL-WAVE RECTIFIER

FIG. 3. Block diagram of the measurement system.

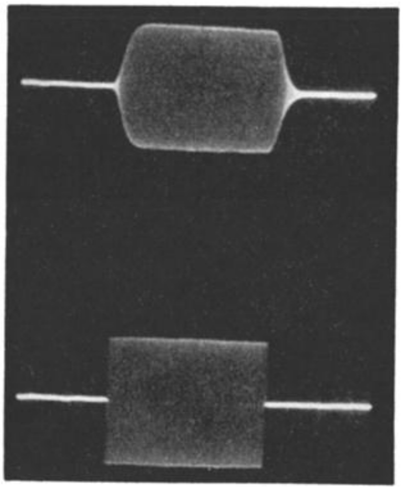

FIG. 4. Input (bottom) and output (top) wave forms of a 50 msec burst of a 2000 cps sine wave passed through a Hewlett Packard Wave Analyzer set at 2000 cps. The rise time of the filter is approximately 5 msec.

To evaluate the transfer characteristics of the system, the following three checks were made. The Hewlett Packard analyzer was set to 2000 cps, and 50 msec of a recorded 2000 cps sine wave was gated into it. Figure 4 shows an oscilloscope trace of the input and output wave forms. The rise time of the filter is approximately 5 msec.

Energy density spectra of a 20 msec and a 50 msec portion of a 2000 cps sine wave were measured using the same procedure as that in the case of the speech sounds. The resultant spectra, which indicate the effects of gating in only a burst of sound as well as the general dynamic range of the measurements, are shown in Fig. 5.

Finally, to evaluate the over-all input to output frequency characteristics, the output of a white noise

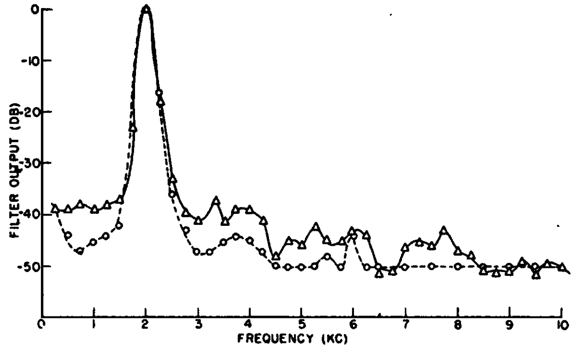

FIG. 5. Energy density spectra of a 20 msec (solid line) and a 50 msec (dotted line) burst of a 2000 cps sine wave.

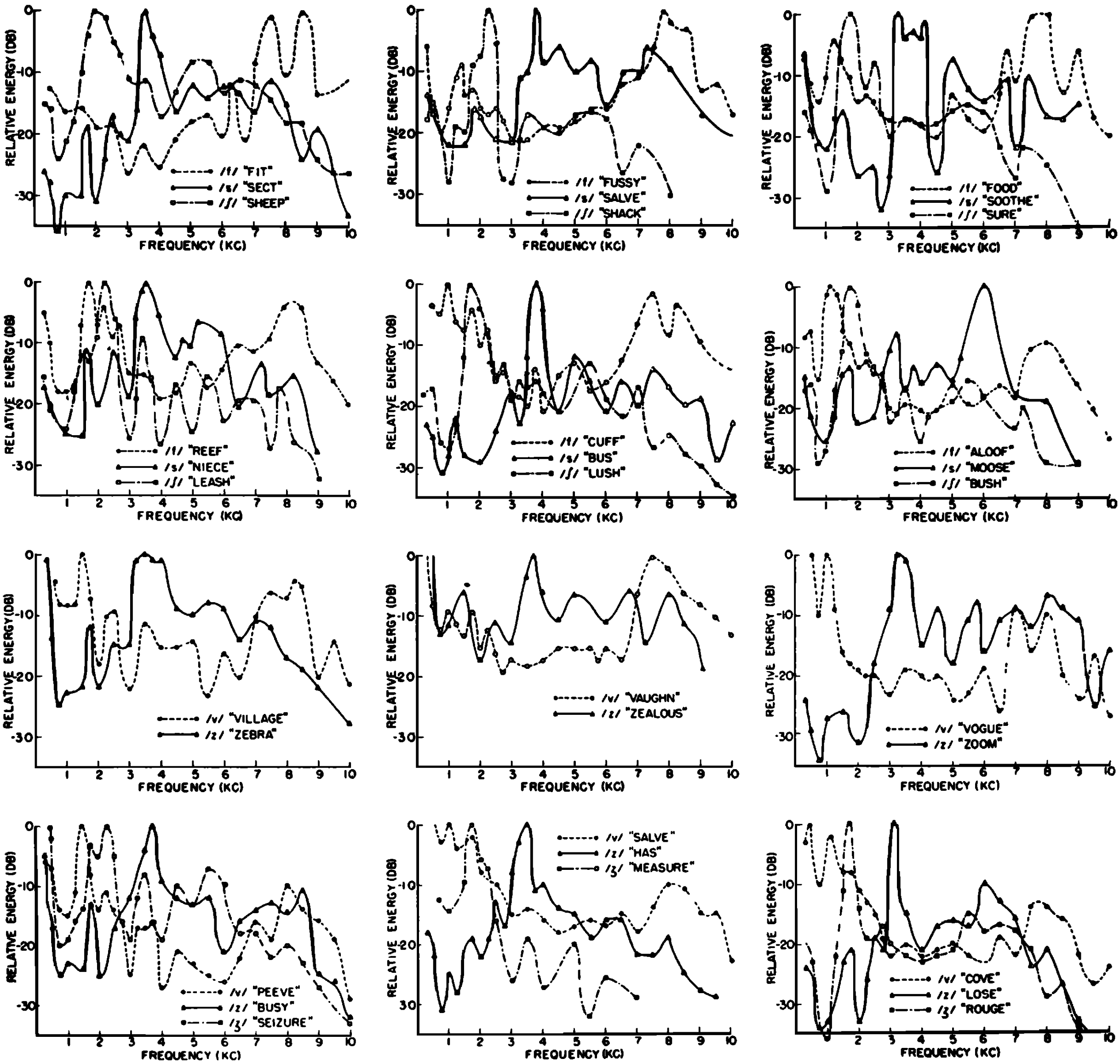

FIG. 6. Energy density spectra of fricative consonants as spoken by speaker E(male).

generator was recorded on a loop and spectra obtained using a 50 msec gate width and several gate positions. Each spectrum was flat to within ±2 db from 300 to 10 000 cps and the ensemble showed no systematic spectral "peaks" or "valleys."

RESULTS

The spectra presented in Figs. 6 to 8 were prepared in the above manner. Although the sounds were recorded as spoken and differed in intensity, each plot of the spectrum was normalized so that its highest peak is represented as 0 db. The spectra are arranged in the following manner: each of the figures is devoted to spectra obtained from a single speaker. (Speakers H and E are male; speaker T is female.) The top two rows in each figure contain spectra of voiceless fricatives, the bottom two rows, spectra of voiced fricatives. Odd numbered rows contain spectra of fricatives in initial position; even numbered rows, spectra of fricatives in noninitial position. In the left-hand column the fricative is adjacent to a front vowel; in the middle column, to a central vowel; in the right-hand column, to a back vowel.

The discrepancies among the spectra of a given fricative as spoken by different speakers in different contexts are so great as to make the procedure of plotting these spectra on one set of axes a not very illuminating one. On the other hand, the differences among the three classes of fricatives (labial, dental, and palatal) are quite consistent, particularly for sounds spoken by a

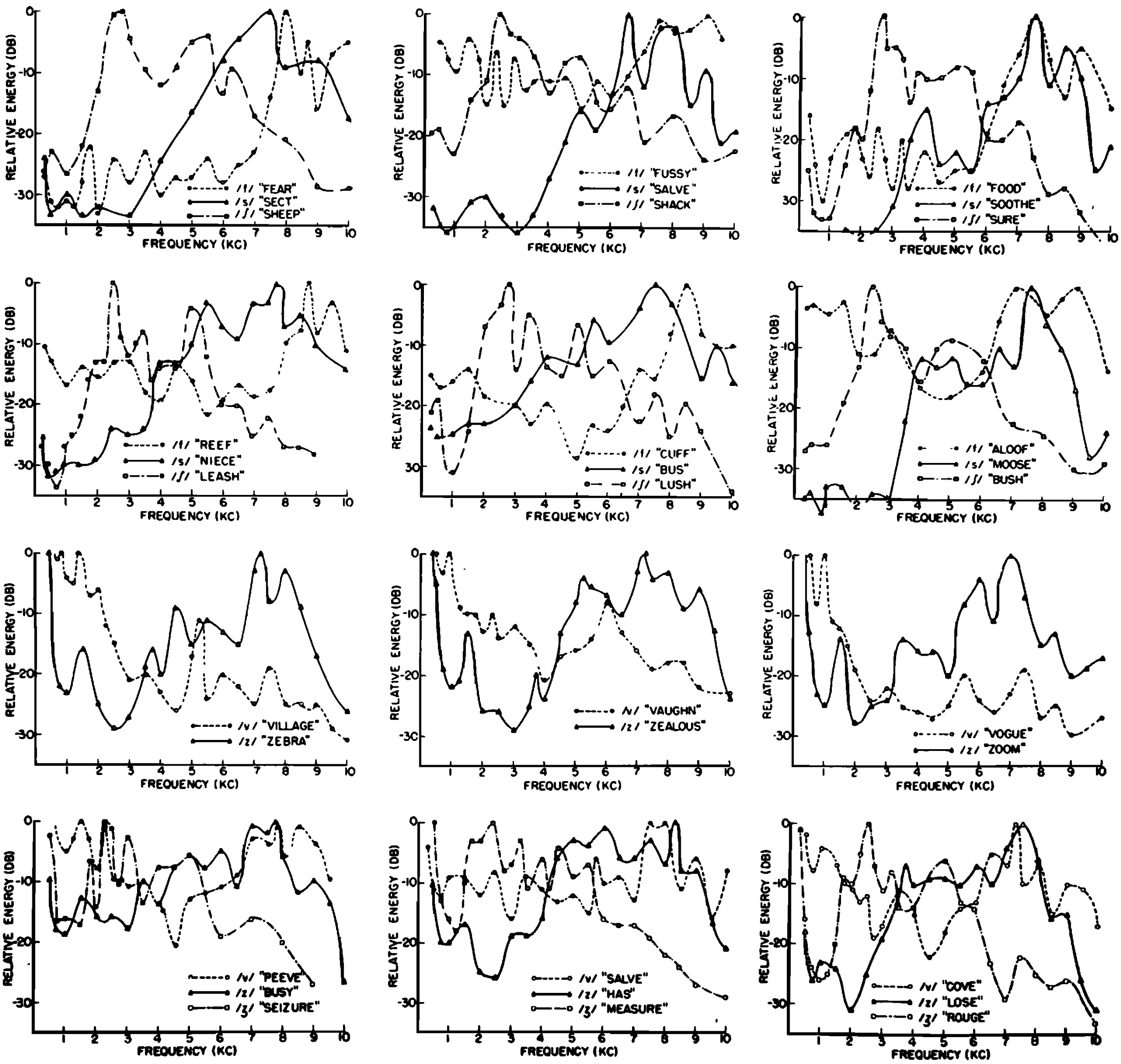

FIG. 7. Energy density spectra of fricative consonants as spoken by speaker H(male).

single speaker. To exhibit this consistency in our spectra we adopted the above method of presentation.

In spectra of the fricatives $|v|$, $|z|$, and $|ʒ|$ a very strong component in the region below 700 cps is often found. In spectra of $|f|$, $|s|$, and $|ʃ|$ this region is never prominent. This strong component is due to the vibration of the vocal cords that often, though not always, occurs during the production of the former sounds, which are commonly known as "voiced." (See spectra of speaker H (Fig. 7), who is particularly consistent in producing the voicing component, with the spectra of "zoom" (speaker E, Fig. 6) and "has" (speaker T, Fig. 8), which do not possess this component.) It appears, therefore, that the distinction between "voiced" and "unvoiced" fricatives is not necessarily made on the basis of a low-frequency component in the spectrum.[2] In the region above 1000 cps the spectra of "voiced" fricatives do not differ appreciably from those of the "unvoiced."

We now turn our attention to the frequency position of the most prominent energy density maximum (peak) in each spectrum. As a first approximation it may be illuminating to view the peaks as resonances of the effective portion of the vocal tract, i.e., of the portion between the point of maximum constriction (point of articulation) and the lips. In such a case we expect an inverse relationship to hold between the length of the effective portion of the vocal tract and the frequency

[2] See pertinent remarks on this point in R. Jakobson *et al.*, *Preliminaries to Speech Analysis*, MIT, Acoustics Lab. TR 13, pp. 26 and 38, and P. Denes "Effects of duration on the perception of voicing," J. Acoust. Soc. Am. 27, 761–764 (1955).

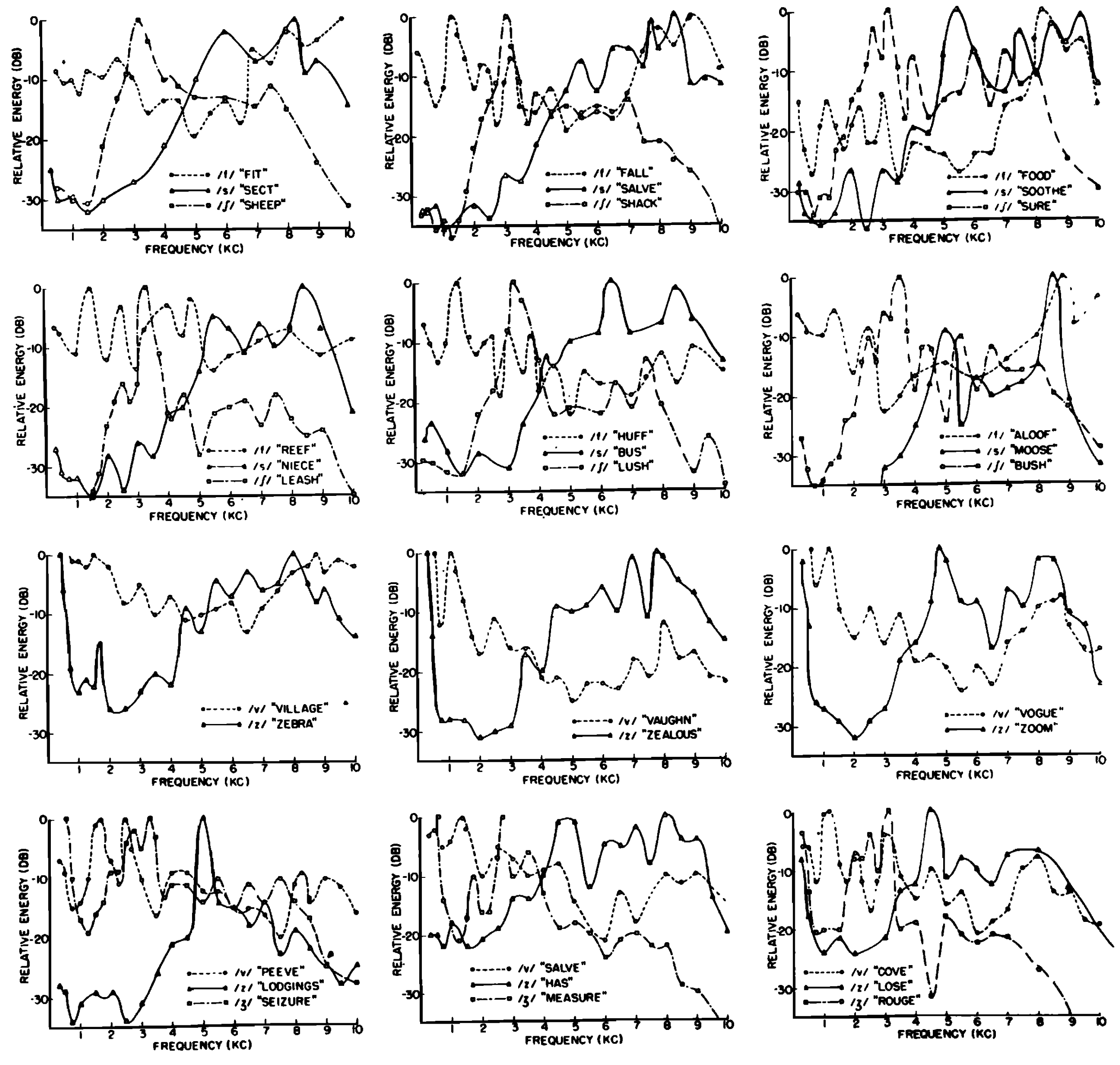

FIG. 8. Energy density spectra of fricative consonants as spoken by speaker T (female).

of the peak. Indeed for any single speaker the spectra of $|s|$ and $|z|$ have peaks at consistently higher frequency than those of $|ʃ|$ and $|ʒ|$. The situation is somewhat more complicated with respect to $|f|$ and $|v|$ where the point of articulation lies at the teeth, so that the effective length of the vocal tract is very small. Consequently in some spectra of $|f|$ and $|v|$ no high-frequency peak can be observed below 10 kc. (See the spectrum of "fussy," in Fig. 6 where a very prominent peak is to be seen at about 8000 cps with that of "huff" in Fig. 8 where no peak is to be seen in the high frequencies.) The low-frequency peaks in spectra of $|f|$ and $|v|$ are due to factors which have been neglected in this approximation.

Individual differences in effective cavity length supply also a partial explanation for the overlaps between the peak frequencies of $|s|$ and $|z|$ of speaker E and those of $|ʃ|$ and $|ʒ|$ of the other speakers; see Fig. 6 "zoom," "lose," "soothe," "zebra" with Fig. 8 "bush," "leash," "seizure." For any one speaker, however, the order among the peak frequencies of different classes of fricatives is consistently maintained.

CRITERIA FOR A MECHANICAL IDENTIFICATION PROCEDURE

In spite of the great divergence among the spectra of different speakers certain common properties emerge. In the following we shall describe a set of criteria which leads towards a separation of the fricative into three classes associated with the three distinct points of

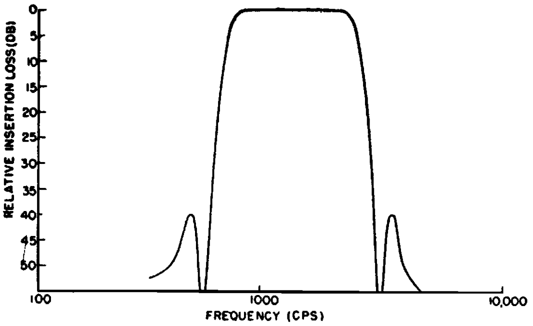

FIG. 9. Measured characteristics of a band pass filter composed of two separate cascaded *M*-derived filters, 720 cps high pass and 2150 cps low pass.

articulation: |*f*| and |*v*| |*s*| and |*z*|, |ʃ| and |ʒ|.[3]

It is clear that detailed spectra such as those in Figs. 6, 7, 8 contain more information than we need. We, therefore, devised a set of measurements which would give us information about gross properties of the spectrum that could be utilized in a mechanical procedure. For these measurements we used a set of fixed filters which lent themselves to rapid data taking. The required pass bands were obtained by cascading *M*-derived high and low pass filters with skirts dropping off about 100 db per octave. A typical filter response curve is shown in Fig. 9. Figure 10 shows an oscillogram of the response of one of the bands (800 to 890 cps—this narrow band was chosen to illustrate the worst transient response) when excited by a 50 msec gated portion of a 825 cps sine wave.

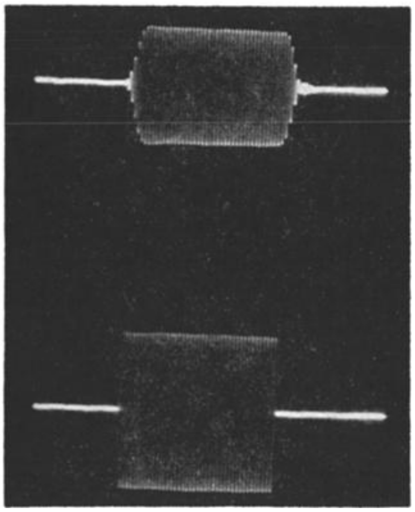

FIG. 10. Input and output wave forms of a 50 msec burst of a 825 cps sine wave passed through two cascaded *M*-derived filters, 800 cps high pass and 890 cps low pass. Note that the rise time is about 6 msec.

The following three measurements were made[4]:

Measurement (1): The energy in db in the band from 4200 cps to 10 kc was subtracted from the energy in db in the band from 720 to 10 kc.

Measurement (2): The energy in db in the band from 720 to 2150 cps was subtracted from the energy in db in the band from 720 to 6500 cps.

Measurement (3): The peak in the region from 1500 cps to 4 kc was located. The energy in db in a band from 720 to 1370 cps was subtracted from the energy in a 500 cps band centered at the peak frequency.

Measurement (1) was formulated on the basis of the observation that |ʃ| almost never had a strong concentration of energy in the range above 4 kc, while |*s*|—with the exception of speaker *E* already discussed—and |*f*| had their peaks above this frequency. If the difference was small, i.e., less than 2 db, the sound was either |*f*| or |*s*|; if it was larger, the sound was |ʃ| or |*f*|.

Measurement (2) is an evaluation of the contribution of the low frequencies in sounds having their peaks above 4 kc. In order to eliminate the effects of the very high-frequency peaks not unusual in |*f*| this measurement was limited to 6500 cps. If the energy in the low-frequency band was substantially equal to that in the 720–6500 cps band (differing by less than 5 db), the sound was identified as |*f*|; if the energy in the 720–6500 cps band exceeded that in the low band by more than 10 db, the sound was identified as |*s*|.

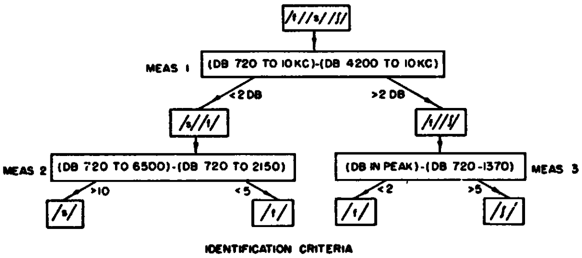
/f//s//ʃ/
MEAS 1
(DB 720 TO 10KC)-(DB 4200 TO 10KC)
<2 DB
>2 DB
/s//f/
/f//ʃ/
MEAS 2
(DB 720 TO 6500)-(DB 720 TO 2150)
(DB IN PEAK)-(DB 720-1370)
MEAS 3
>10
<5
<2
>5
/s/
/f/
/f/
/ʃ/
IDENTIFICATION CRITERIA

FIG. 11. Schematic diagram of the identification procedure.

Measurement (3) evaluated the predominance of the peak in the central region. Large positive values (more than 5 db) indicated predominance of the peak, hence the sound was identified as |ʃ|; negative and small positive values (less than +2 db) indicated absence of such predominance, hence the sound was identified as |*f*|.

A summary of this identification procedure is shown in Fig. 11.

In this manner we measured 190 fricative consonants, voiced as well as unvoiced, taken from English words spoken by two female and three male speakers (the three speakers whose spectra are shown in Figs. 6–8 and two others). Measurement (1) yields critical

[3] Since in the following measurements we eliminate frequencies below 700 cps, we also eliminate information regarding the presence or absence of the voicing component; we shall, therefore, refer to each pair of "voiced" and "unvoiced" fricatives by its "unvoiced" member: |*f*| will thus stand for both |*f*| and |*v*|; and |*s*|, for both |*s*| and |*z*|; |ʃ| for both |ʃ| and |ʒ|.

[4] No special significance is to be attached to the odd frequency values in the following measurements; they are due to design considerations of our fixed filters.

TABLE I. Responses of the listeners to the individual sounds as uttered by the different speakers.

		Sounds intended by											
		Speaker E			Speaker T			Speaker R			Speaker H		
		s	f	ʃ	s	f	ʃ	s	f	ʃ	s	f	ʃ
Sounds judged by listeners to be	s	38	12	16.5	83.5	32.5	22	56	16	16.0	79.5	19	11
	f	6	80	7.5	12	60.5	7	6.5	70	20.5	13	78	6.5
	ʃ	56	8	76	4.5	7	71	37.5	14	63.5	7.5	3	82.5
Total number of judgments		280	180	80	220	160	100	200	160	140	220	140	120

judgments only for |*s*| and |ʃ|. Out of 125 |*s*| and |ʃ|, the correct division was made in 107 cases (86%); 16 of the 18 "mistakes" were with |*s*| uttered by speaker *E*, whose |ʃ| and |*s*| spectra are shifted down by about an octave as compared to those of other speakers. The relative position of peaks in the spectra of |*s*| and |ʃ| was, however, kept invariant by speaker *E*, although in absolute numbers there was an overlap between his |*s*| and |ʃ| of other speakers.

Measurement (2) gave clear separation in 82 out of 88 (93%) of cases. The six dubious cases were all |*f*| which fell in the region between 5 and 10 db; i.e., there was no overlap.

Measurement (3) gave the correct separation in 76 out of 84 (90%) cases. There was no discernible pattern in the mistakes.

PERCEPTUAL TESTS

In order to check the above criteria we devised the following perceptual experiment which tested the identifiability by human subjects of gated portions of fricatives and attempted to establish correlations between the gross spectral properties that we proposed as our mechanical identifying criteria and the responses of our subjects. In other words, suppose human listeners were fed the same information as our machines, could they identify the sounds correctly?

Gated portions, each 50 msec long, of 46 |*s*|, 32 |*f*|, and 22 |ʃ|, taken from isolated words spoken by one female and three male speakers, were recorded twice in random order on a test tape. The test tape, containing 200 samples in all, was played through a sound system having a reasonably flat response up to 10 kc and presented to a group of 10 listeners at very high signal to noise ratio. The listeners were instructed to identify every sample as one of the three fricatives |*f*|, |*s*|, |ʃ|.

The question of learning could not be investigated in detail. Since, however, the listeners had to judge each stimulus twice, the changes in the number of "correct" responses (where the listeners agreed with the speakers) between the first and the second presentation gives some indication of the degree of learning. The first time the hundred stimuli were presented there were 65% correct responses; the second time the same stimuli were presented there were 71%. This shows that there was only a small degree of learning and, therefore, in what follows the results of both runs are taken together.

Since the listeners were asked to choose one out of three possible answers, $33\frac{1}{3}$% correct answers could be expected purely on the basis of random guessing. The actual percentage of correct responses is more than twice that number (i.e., 68%).

In Table I is shown the way in which the listeners

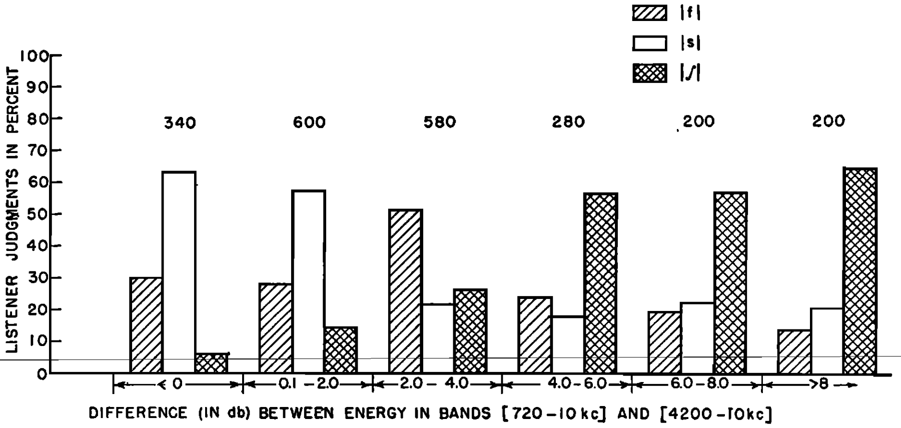

FIG. 12. Responses of the listeners as a function of ranges of increasing values of measurement 1 (the difference in db between energy in bands [720 cps −10 kc] and [4200 cps−10 kc]). The numbers above the bars are the total numbers of listener judgments made on sounds in each of the different ranges of values.

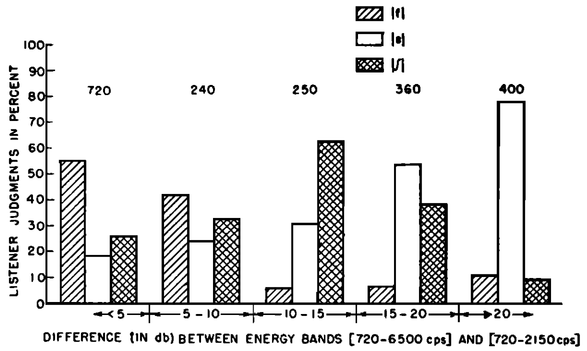

FIG. 13. Responses of the listeners as a function of ranges of increasing values of measurement 2 [the difference in db between the energy in the bands (720–6500 cps) and (720–2150 cps)]. The numbers above the bars are the total numbers of listener judgments made on sounds in each of the different ranges of values.

judged the sounds produced by each of the speakers. We note that in most instances a very high percentage of the responses agreed with the speakers. The cases where there was no such agreement are very significant in that they are precisely those where our identification criteria gave results at variance with the speaker's professed intention: It is seen that both the listeners and our criteria (as mentioned above) tended to identify the |*s*|'s of speaker *E* as |*ʃ*|'s. The high percentage of |*s*| judgments for |*f*| of speaker *T* is due to their very strong high-frequency component. We might add that in the latter case our criteria gave us a significantly better result than would be expected from the listening tests.

The correlation between the three critical measurements which we propose to use for identification and the listener responses was investigated. As shown in Fig. 12 an increase in measurement (1) is accompanied by a significant increase in |*ʃ*| judgments and a decrease in |*s*| judgments. |*f*| judgments increase up to the value 4 and then begin to decrease—it is significant that just at this point |*ʃ*| judgments show a sudden increase.

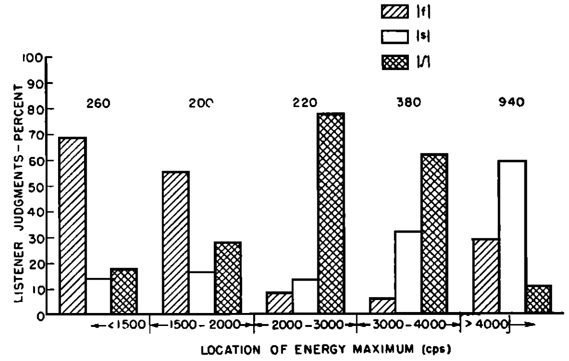

FIG. 14. Responses of the listeners as a function of ranges of increasing values of the location in cps of the energy maximum. The numbers above the bars are the total numbers of listener judgments made on sounds in each of the different ranges of values.

Figure 13 shows the responses as functions of measurement (2). It reflects the dependence of |*s*| and |*f*| on the extremes of the speech spectrum. |*ʃ*| judgments reach a maximum for intermediate values of measurement (2). The same is brought out even more strikingly in Fig. 14 where the responses are shown as functions of the frequency position of the maximum of the spectrum. We call particular attention to the sharp drop in |*f*| identifications for peaks located above 2000 cps; to the increase in |*s*| judgments for peaks above 4 kc; and to the fact that the highest percentage of |*ʃ*| judgments was found for intermediate frequencies, i.e., between 2–4 kc.

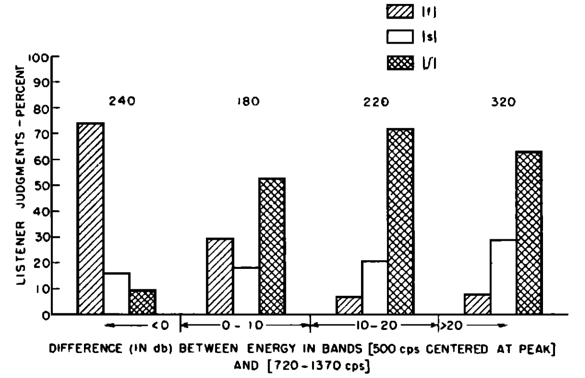

FIG. 15. Responses of the listeners as a function of ranges of increasing values of measurement 3 [the difference in db between the energy in a 500 cps band centered at the spectral peak and the band (720–1370 cps)]. The numbers above the bars are the total numbers of listener judgments made on sounds in each of the different ranges of values.

Figure 15 shows the judgments as functions of measurement (3), which expresses the prominence of the peak above the region between 720–1370 cps. Since sounds having a peak above 4 kc are uniquely identified by measurements (1) and (2), they were not subjected to measurement (3), a fact which is reflected in the absence of high percentages of |*s*| judgments. The data show clearly that a predominance of the peak over the low region leads to a high percentage of |*ʃ*| judgments.

ACKNOWLEDGMENTS

We are happy to acknowledge our indebtedness to Dr. A. Cohen, a Rockefeller Foundation fellow from the University of Amsterdam, for carrying out the perceptual experiments discussed above.