X. SPEECH COMMUNICATION

Prof. M. Halle	J. Emerson	K. Nakata
Prof. K. N. Stevens	J. M. Heinz	G. Rosen
Dr. T. T. Sandel	D. T. Hess	C. E. Persons
G. W. Hughes	P. Lieberman	M. G. Saslow
R. Capraro	C. I. Malme	M. G. Schachtman

A. REACTION TIME TO CONSONANT-VOWEL SYLLABLES IN ENSEMBLES OF VARIOUS SIZES*

Over-learned acoustic stimulus-response pairs were used in a study of human reaction time. The size of the stimulus ensemble was the principal independent variable. Two experienced subjects were used as listeners. The stimuli were natural speech sounds, disc-recorded by a single speaker (M.G.S.), and included the consonant-vowel syllables [bɑ], [dɑ], [gɑ], [kɑ], [pɑ], [tɑ], [bi], [di], [gi], [ki], [pi], [ti]. Each stimulus was preceded by a short, faint, warning tone, one second before stimulus presentation. Presentations were approximately 30 to 45 sec apart, in groups of approximately 30 stimuli at a time. The instructions for each test listed the stimuli in the test ensemble, stated that the stimuli were equally probable, and asked the subject to "repeat each presented stimulus accurately but quickly." Reaction time was defined as the interval between stimulus onset and response onset, as determined by voice-operated relays that gated a counter.

The responses were essentially 100 per cent correct for ensembles of sizes 2, 4, 8, and 12. A straight line fitted to the data is practically flat, as shown in Fig. X-1; reaction time increases very slowly as a function of ensemble size. In comparison, Albert's (1) and Stevens' (2) data for responses to pure tones, and Hyman's data (3) for verbal responses to visually presented nonsense syllables would be represented on the scale of Fig. X-1 by a line rising at a rate of 0.2 sec for each doubling of ensemble size.

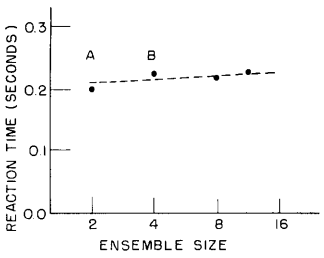


Fig. X-1. Reaction time as a function of ensemble size. The data represent average response times for two subjects, 120 judgments per point per subject. Because of partial unbalance of the design and dissimilar ease of enunciation of certain responses, point A is probably a low estimate and point B is possibly a high estimate.

The low rate of increase in reaction time and the lack of errors for increasing ensemble size are not like the results of previous related experiments (1, 2, 3, 4, 5). The difference may be explained by the fact that the subjects had been practicing the stimuli and responses of this experiment since early childhood.

---

* This work was supported in part by National Science Foundation.

The author acknowledges with appreciation the assistance and practice of the subjects, Mr. H. M. Chapman and Miss Marla M. Moody.

M. G. Saslow

References

1. A. Albert, S.B. Thesis, Department of Electrical Engineering, M.I.T., June 1956.
2. K. N. Stevens, Experiments in pitch discrimination and reaction time, Report No. 1, Instituut Voor Perceptie Onderzoek, Eindhoven, Netherlands, 9 Oct. 1957.
3. R. Hyman, Stimulus information as a determinant of reaction time, J. Exptl. Psychol. 45, 188-196 (1951).
4. G. A. Miller, The magical number seven, plus or minus two; some limits on our capacity for processing information, J. Psychol. Rev. 63, 81-97 (1956).
5. W. E. Hick, Proc. Symposium on Information Theory, London, September 1950, Trans. IRE, PGIT-1, pp. 130-133 (Feb. 1953), refers to work by Hick in 1950 and by Merkel in 1885.

B. SYNTHESIS OF NASAL CONSONANTS BY TERMINAL ANALOG SYNTHESIZER

In a previous report (1) we presented the results of a listening test in which the stimuli were synthetic consonant-vowel syllables each of which consisted of a nasal consonant plus the vowel [ɑ]. Both consonant and vowel segments were generated by a resonance-analog synthesizer, and the second formant transition of the vowel was given a range of values. Formant bandwidths and frequencies and temporal characteristics of the stimuli were described in detail. The principal findings of that study were: (a) Very low (approximately 200 cps) starting frequency and broad bandwidth for the first formant are adequate cues for distinguishing nasal consonants from other consonants. (b) Nasal consonants can be distinguished from each other on the basis of the starting frequency of the second formant (F_2 locus), and well-defined frequency ranges for the F_2 locus are associated with each of the nasal consonants, at least in the vowel context [ɑ]. (c) Duration of the initial (nasal) segment and transition time had a small but significant effect on identification of the nasal consonants.

The initial study has now been extended to include the vowel contexts [i, I, ɛ, ɔ, u]. Listening tests were performed with stimuli consisting of each of these vowels preceded by the same synthetic nasal consonants. Data from the new tests show that the frequency ranges for the F_2 loci appropriate to each of the nasal consonants /m, n, ŋ/ are somewhat dependent upon the adjacent vowels. The results, summarized in Fig. X-2, show the range of the F_2 loci that yield more than 50 per cent identification for each of the nasal consonants in each vowel context. The data of Fig. X-2 summarize average responses for several values of consonant duration and transition time. The best F_2 locus for each consonant apparently decreases in frequency as the frequency of the

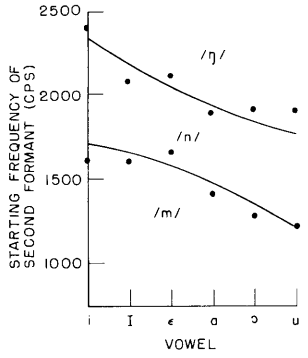


Fig. X-2. Range of second-formant transitions for stimuli that elicit responses of /m/, /n/, and /ŋ/ more than 50 per cent of the time. Each stimulus consists of a synthetic nasal consonant plus one of the vowels indicated on the abscissa.

second formant of the adjacent vowel decreases. This finding is not entirely consistent with the hypothesis that a particular nasal consonant is associated with a fixed articulatory configuration and hence with a fixed F_2 locus. The effect is not marked, however, and we can still specify F_2 loci for /m/, /n/, and /ŋ/ at frequencies of approximately 900, 1700, and 2300 cps, respectively, as demonstrated previously (1, 2).

We have performed another study that was designed to determine the identifiability of isolated synthetic nasal consonants. The stimuli consisted of the outputs of the resonance analog synthesizer, with the following characteristics:

1. Frequency of first resonance, 200 cps.
2. Bandwidth of first resonance, 300 cps.
3. Frequency of second resonance, 900, 1100, 1300, 1500, 1700, 1900, 2100, 2300 cps.
4. Bandwidth of second resonance, 30-100 cps.
5. Frequency of third resonance, 2500 cps.

The stimuli were presented in random order to a group of subjects, who were asked

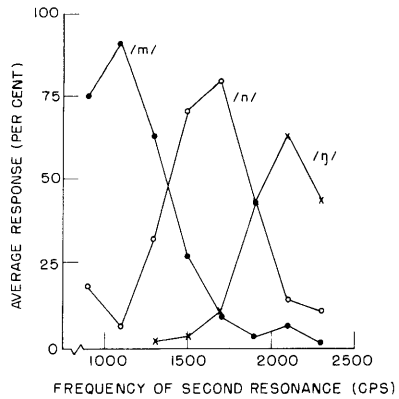


Fig. X-3. Responses to synthetic nasal consonants presented in isolation. The ordinate indicates the frequency of the second resonance of the nasal consonant.

to identify each stimulus as one of /m, n, ŋ/. The results indicate that the identification of the synthetic nasal consonants is dependent to some extent on the second-resonance

frequency. Maximum response percentages for /m/, /n/, and /ŋ/ were 64, 52, and 41 at frequencies of 1100, 1700, and 2300 cps for the second resonances. We repeated the test with a broader bandwidth (200 cps) for the second resonance; the results are shown in Fig. X-3. These response curves are more clearly defined than those for narrow-bandwidth second resonances.

It is of some interest to compare the data of Fig. X-3 with similar data of Malécot (3), who used nasal consonants spoken in isolation as stimuli. In a forced-judgment test with the natural stimuli he found response percentages of 96, 56, and 12, for /m/, /n/, and /ŋ/, respectively. Apparently, the synthetic stimuli, particularly [n] and [ŋ], can be more readily identified than the natural stimuli! In the synthetic sounds the cues are probably accentuated, whereas they are obscured by "noise" in the spoken versions.

K. Nakata, K. N. Stevens

References

1. K. Nakata, Synthesis of nasal consonants by terminal analog synthesizer, Quarterly Progress Report, Research Laboratory of Electronics, M.I.T., April 15, 1958, p. 104.
2. A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. J. Gerstman, The role of consonant-vowel transitions in the perception of the stop and nasal consonants, Psychol. Monographs 68.8, 1-13 (1954).
3. A. Malécot, Acoustic cues for nasal consonants, Language 32, 274-284 (1956).

C. MODEL STUDIES OF THE PRODUCTION OF FRICATIVE CONSONANTS*

Fricative consonants are produced when the vocal tract is excited by acoustic noise generated at a constriction by turbulent flow. For these sounds, the vocal tract is characterized by a tongue constriction that separates the front and back cavities. This narrow constricted passage is usually small enough so that the back cavity is essentially decoupled from the front cavity. The radiated sound is, therefore, mainly characteristic of the front cavity, of the constriction itself, and of the source.

To investigate some aspects of the mechanism of the production of fricative consonants, an idealized mechanical model of the human vocal organs was constructed. This model, shown in Fig. X-4, consists of a tube that is comparable in length with the human

* This work was supported in part by Contract AF19(604)-2061 with Air Force Cambridge Research Center and in part by Contract N5ori-07861 with the Navy (Office of Naval Research).

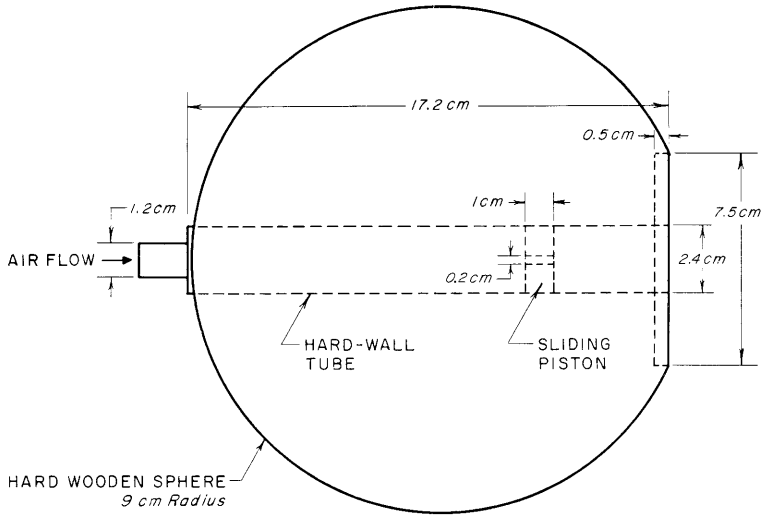


Fig. X-4. Mechanical model for fricative consonant production.

vocal tract and is imbedded in a spherical baffle that represents the human head. A constriction is formed in the "tract" by inserting a piston with a hole in its center into the tube. Air flow is supplied to the model in such a way that turbulent flow is created near the exit of the constriction. The turbulent noise excites the cavities of the model and sound is radiated from its mouth.

For this investigation, a constriction of 1 cm in length and 0.2 cm in diameter was used. These dimensions give negligible coupling to the back cavity. The constriction was used in two positions, at the mouth and 4 cm from the mouth of the "tract." These positions give extremes of coupling between the source and the resonator. In the front position the source is largely decoupled from the resonator; in the back position it is completely within the resonator. Four values of air-flow velocity, ranging from 95 to 195 cc/sec, were used.

Measurements of the radiated sound were made by means of a pressure microphone that was placed 38 cm from the model. Data were taken at each of nine positions around the model for each combination of the variables. The over-all sound pressure level and the spectrum level were measured with a continuously variable filter of 4-cps bandwidth.

The over-all sound pressure measurements for the constriction at the mouth were numerically integrated over all positions to obtain the total acoustic-power output as a function of air flow. The acoustic power was found to be proportional to the fifth power of the flow velocity. Typical data for the constriction at the mouth are shown in Fig. X-5. Since the source is very loosely coupled, the radiated sound below 10 kc is mainly characteristic of the acoustic source itself and is fairly flat, with a broad maximum between 4 and 6 kc. Near the frequency that corresponds to the one-half wavelength resonance

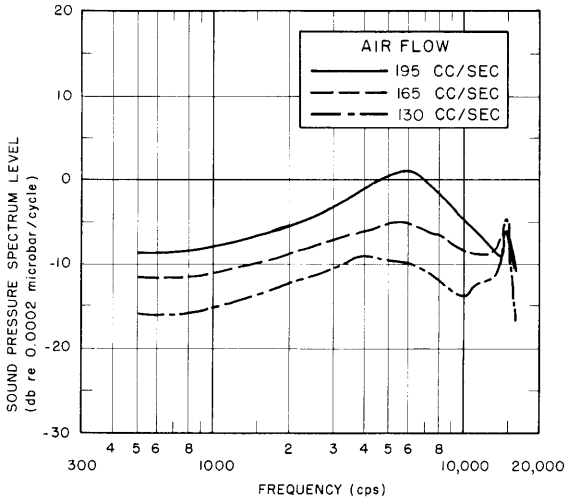


Fig. X-5. Typical pressure spectral data for constriction at the mouth of the model.

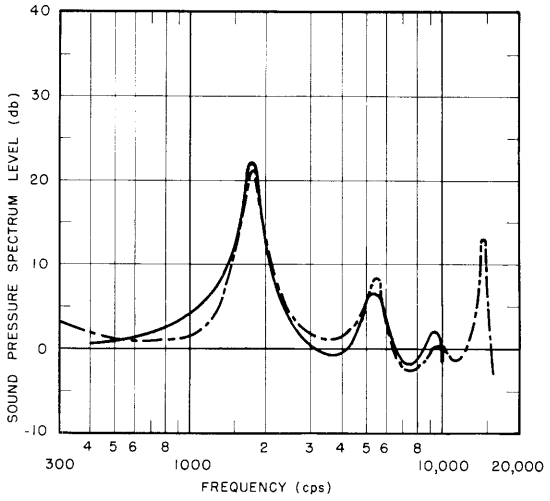
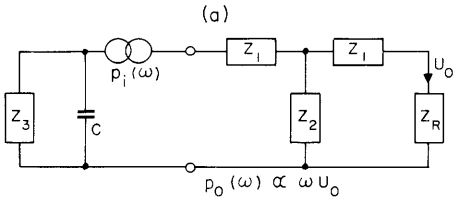
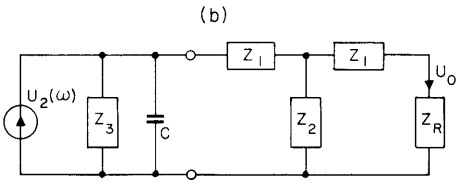


Fig. X-6. Comparison of calculated system function and spectrum measured with constriction at 4 cm from the mouth of the model.

——— Calculated.
— - — Measured.





$$Z_1 = \frac{\rho_0 c}{\pi a^2} \text{ TANH } \left[j \frac{\omega \ell}{2c} \right]$$

$$Z_2 = \frac{\rho_0 c}{\pi a^2} \text{ CSCH } \left[j \frac{\omega}{c} \ell \right]$$

Z_3 = INPUT IMPEDANCE OF CONSTRICTION

Z_R = ACOUSTIC RADIATION IMPEDANCE

Fig. X-7. (a) Thévenin's and (b) Norton's equivalent circuits for plane wave propagation in a lossless tube of radius a and length ℓ. $\rho_0(\omega)$ is the sound pressure measured on the axis in the far field; $\rho_i(\omega)$ is the equivalent pressure source; and $U_2(\omega)$ is the equivalent volume velocity source representation of the actual acoustic source.

of the constriction, the radiated sound depends mainly on the resonator, and a sharp peak can be seen in the spectra at 14.5 kc. The spectral data were numerically integrated to find the total power-density spectra. Below 10 kc they are very similar to the spectra shown in Fig. X-5.

A typical measured spectrum for the constriction that is 4 cm from the mouth is shown by the dotted curve in Fig. X-6. This curve shows peaks that result from the first three normal resonances of the front cavity, as well as a peak that is caused by the constriction resonance.

If it is assumed that only plane waves propagate in the tube, the electrical analog circuits shown in Fig. X-7 can be used to represent the acoustic behavior of the tube. The upper circuit consists of Thévenin's equivalent source that excites a transmission line that has the same length as the front cavity and is terminated in the radiation impedance. The second circuit shows Norton's equivalent source. The internal impedance of the "source" consists of Z_3, the input impedance of the constriction, in parallel with a compliance C that represents the volume between the source and the constriction. Some doubt still exists about the exact nature of the internal source impedance because the actual acoustic source is distributed over a finite region with different parts of the radiated spectrum arising from different parts of the region. Figure X-6 gives a comparison between the transfer function U_o/U_2, calculated from the electrical analog, and one of the measured spectra. In this calculation the internal impedance of Fig. X-7b was assumed to be high enough to be negligible.

In a model with a larger constriction size the internal impedance Z_3 could no longer be neglected, and it would modify the spectrum of the output. At frequencies for which the impedance, as we look back from the source in Fig. X-7a, is infinite, there would be antiresonances or zeros in the output. Such antiresonances can be observed in the spectra of spoken fricative consonants (1, 2, 3).

These investigations with the mechanical analog and the studies of fricative spectra have brought out several things about the production of fricative consonants. The spectral maxima are primarily the natural resonances of the cavity in front of the tongue constriction, while antiresonances or zeros are determined mainly by the constriction. The spectrum of the acoustic source is fairly flat within the range that is of interest, and the total power of the source, for a given geometry, is roughly proportional to the fifth power of the flow velocity.

J. M. Heinz

References

1. G. W. Hughes and M. Halle, Spectral properties of fricative consonants, J. Acoust. Soc. Am. 28, 303-310 (1956).

2. J. M. Heinz, A terminal analog of fricative consonant articulation, Quarterly Report, Acoustics Laboratory, M.I.T., July-September 1957, pp. 1-3.

3. G. Fant, Acoustic theory of speech production, Report No. 10, Royal Institute of Technology, Division of Telegraphy-Telephony, Stockholm, Sweden, 1958.

D. SYNTHESIS AND PERCEPTION OF FRICATIVE CONSONANTS*

Studies with the mechanical analog, discussed in Section X-C, and measurements of the spectra of spoken fricative consonants have shown that fricative sounds can be approximated by excitation of a simple passive electric circuit by white noise. The transfer characteristic of the circuit has poles and zeros that reflect the resonances of the vocal-tract constriction and of the cavities anterior to the constriction. To yield a reasonably good approximation to the fricative spectra up to 8-10 kcps, the circuit must generally have two or more poles and one or more zeros. Some of the principal features of the spectra can be simulated, however, if a simpler circuit with only one pole and one zero is used.

In the development of speech synthesizers for practical application it is desirable to utilize the simplest possible circuits for generation of speech sounds and at the same time maintain a reasonable degree of naturalness and intelligibility. Consequently, we have been examining the degree to which a simple pole-zero circuit (with modifications that will be discussed) can synthesize acceptable versions of unvoiced fricatives in consonant-vowel syllables. We hope also that this study will help us to understand the cues that are important for the perception of fricative consonants.

The particular circuit used to generate the stimuli is shown in Fig. X-8a, and a typical output spectrum for white-noise excitation of the circuit is shown by the solid curve, labeled "-20 db," in Fig. X-8b. In this example the resonant frequency is 8000 cps, but the stimuli that were tested had resonant frequencies of 2500, 3500, 5000, 6500, and 8000 cps. The zero in the transfer function is always approximately one octave below the pole. The bandwidths of the resonances ranged from 400 cps at 2500 cps to 1500 cps at 8000 cps, and corresponded roughly to the observable bandwidths in the spectra of the spoken fricatives. A preliminary listening test indicated that a variation of bandwidths over a 2:1 range around the selected values did not greatly affect the identification of the stimuli.

Although spectra of the type shown by the solid line in Fig. X-8b are quite similar to spectra of spoken /ʃ/ and /s/ sounds, there are some /f/ and /θ/ spectra that are characterized by broad low-frequency noise in addition to the high-frequency peak. Apparently the source of this low-frequency noise is turbulence at the lips, which generates noise that is relatively uncoupled to the vocal-tract cavities. In order to examine the effect of this feature in the perception of fricatives, we added low-frequency noise electrically to some of the basic stimuli that had resonant frequencies of 6500 and 8000 cps.

---

* This work was supported in part by Contract AF19(604)-2061 with Air Force Cambridge Research Center.

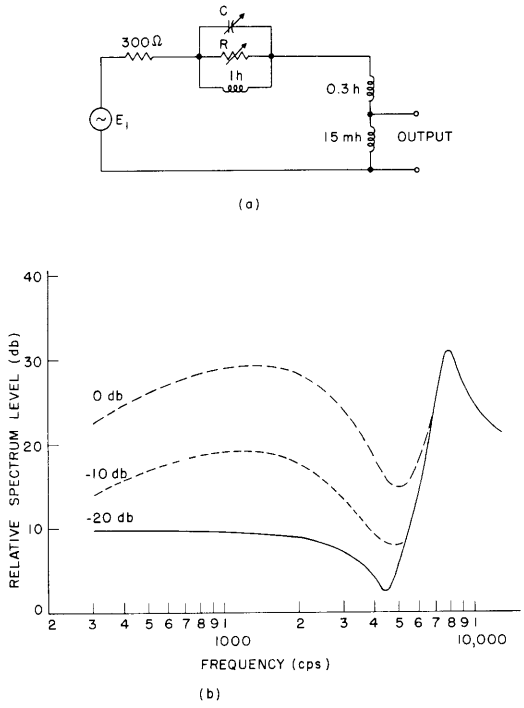


Fig. X-8.

(a) Circuit for generating synthetic fricatives in the listening test. The transfer function has one pole and one zero, and these critical frequencies can be varied over a wide range of values, subject to the restriction that the frequency of the zero is approximately one-half of the frequency of the pole. (b) The solid line shows a typical output spectrum with white-noise excitation and a pole frequency of 8000 cps. The dashed curves show how the basic output spectrum was modified by the addition of low-frequency noise. The parameter is the intensity of the low-frequency part relative to that of the high-frequency part (in db).

The values of low-frequency noise are shown in Fig. X-8b, in which the levels of the low-frequency noise are labeled 0 db, -10 db, and -20 db. Thus a total of nine noise spectra were tested, four of which were characterized by additional low-frequency noise.

Each of the nine noise spectra was synthesized in combination with the vowel [ɑ] to form a syllable. The vowel was generated by a resonance synthesizer (1). Timing of fricative onset and decay, vowel onset and decay, and formant transitions was accomplished by the Rosen timer and control apparatus (2). The over-all level of the fricative relative to the vowel was given three values: -5, -15, and -25 db. The initial 25 msec of the vowel was characterized by a rising first-formant transition and three values of second-formant transition, in which the starting frequencies or loci were 900, 1700, and 2400 cps. Stimuli with no vowel transitions were also used. The 108 stimuli (9 spectra, 3 levels, 4 transition values) were presented in random order to a group of listeners

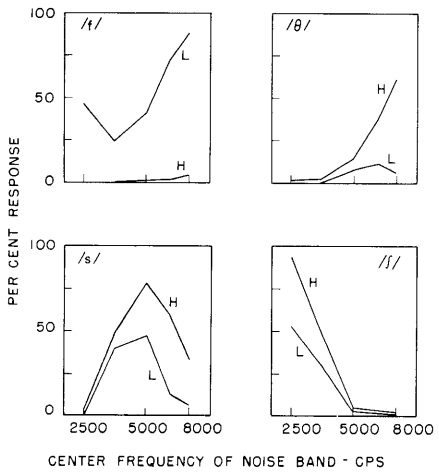


Fig. X-9. Results of fricative identification test. Each part of the figure shows responses to one of the fricatives. Curve "L", data for a rising vowel formant transition; curve "H", average data for two values of falling vowel formant transition.

who were asked to identify the consonant as one of /f, θ, s, ʃ /.

In general, the addition of low-frequency noise did not greatly modify the pattern of responses. The best /f/ and /θ/ responses were obtained with a relative noise level of -15 or -25 db and a resonance of 8000 cps (with or without low-frequency noise). For the best /f/ responses, a rising second-formant transition was required, whereas a falling second-formant transition gave best /θ/ responses. Levels of -5 db or -15 db with falling formant transitions yielded best /s/ and / ʃ / responses. Maximum /s/ and / ʃ / responses were obtained for resonant frequencies of 5000 and 2500 cps, respectively.

A partial summary of the results is given in Fig. X-9, in which the percentage of response is plotted as a function of the center frequency of the noise band. Each response is summarized by two curves: L represents a second-formant transition rising from 900 cps, and H represents the average response for the two values of the falling second-formant transition. The curves represent average responses for all three levels of fricative noise. The data show clearly that formant transitions play a greater role in the perception of /f/ and /θ/ than in the identification of /s/ and / ʃ /. These results are in general agreement with those of Harris, who generated synthetic fricatives by passing white noise through highpass filters at various cutoff frequencies (3). The results also demonstrate that a simple circuit of the type shown in Fig. X-8, excited by white noise, is capable of generating acceptable versions of unvoiced fricative consonants.

K. N. Stevens, K. Nakata

References

1. K. N. Stevens, Synthesis of speech by electrical analog devices, J. Audio Eng. Soc. 4, 2-8 (1956).

2. G. Rosen, Dynamic analog speech synthesizer, Quarterly Progress Report, Research Laboratory of Electronics, M.I.T., April 15, 1958, p. 102.

3. K. S. Harris, Cues for the discrimination of American English fricatives in spoken syllables, Language and Speech 1, 1 (1958).

E. A WIDE-RANGE ELECTROSTATIC LOUD-SPEAKER*

Present methods of construction and operation of electrostatic loud-speakers limit the achievable diaphragm excursion, and hence they are useful only at the higher audio frequencies. An investigation of the problems in the design and construction of an electrostatic loud-speaker that will cover the entire audio spectrum was completed by the author as a thesis for the S.M. degree.

The basic design features are:

(a) light, circular, peripherally supported diaphragm, 20 inches in diameter.

(b) high-resistivity coating on the diaphragm surface to give constant-charge operation without the use of an external series resistor.

(c) 16-kv bias voltage applied through a corona ring around the edge of the diaphragm.

(d) high-output-voltage audio amplifier to provide a driving signal of 4.5 kv rms.

(e) large diaphragm excursion to allow adequate low-frequency reproduction.

(f) electrical segmentation of the diaphragm to give a broad directivity pattern at all frequencies.

The pressure response of the experimental loud-speaker is shown in Fig. X-10. It was measured in an anechoic chamber at a distance of 6 ft, with the microphone located 15° off the axis of the loud-speaker. The upper curve was obtained by using an electric network to partially compensate for the front-to-back cancellation of the (4-ft-square) loud-speaker baffle. The lower curve was obtained without compensation.

Figure X-11 is a circuit diagram of the high-output-voltage audio amplifier that was used to drive the electrostatic loud-speaker. The amplifier is capable of delivering an output voltage of 4.5 kv rms to the loud-speaker and incorporates 45 db of negative feedback for reducing the output impedance to approximately 6000 ohms. The frequency response is flat within 4 db from 20 cps to 20,000 cps. The maximum power output is 4.5 watts.

The loud-speaker diaphragm is made of 1/4-mil Mylar plastic film and has an electrically conductive high-resistivity coating on both sides. The film is held between two fixed electrodes composed of vertical wire segments that are electrically connected so that successively smaller sections of the diaphragm are driven at higher frequencies. This prevents the high frequencies from emerging in a sharp beam.

Tests of transient response, obtained by pulsing the loud-speaker with tone bursts from a modulator, showed that the electrostatic loud-speaker is free from the ringing response usually obtained from cone speakers. Another advantage that was found is the

* This work was supported in part by Contract N5ori-07861 with the Navy (Office of Naval Research).

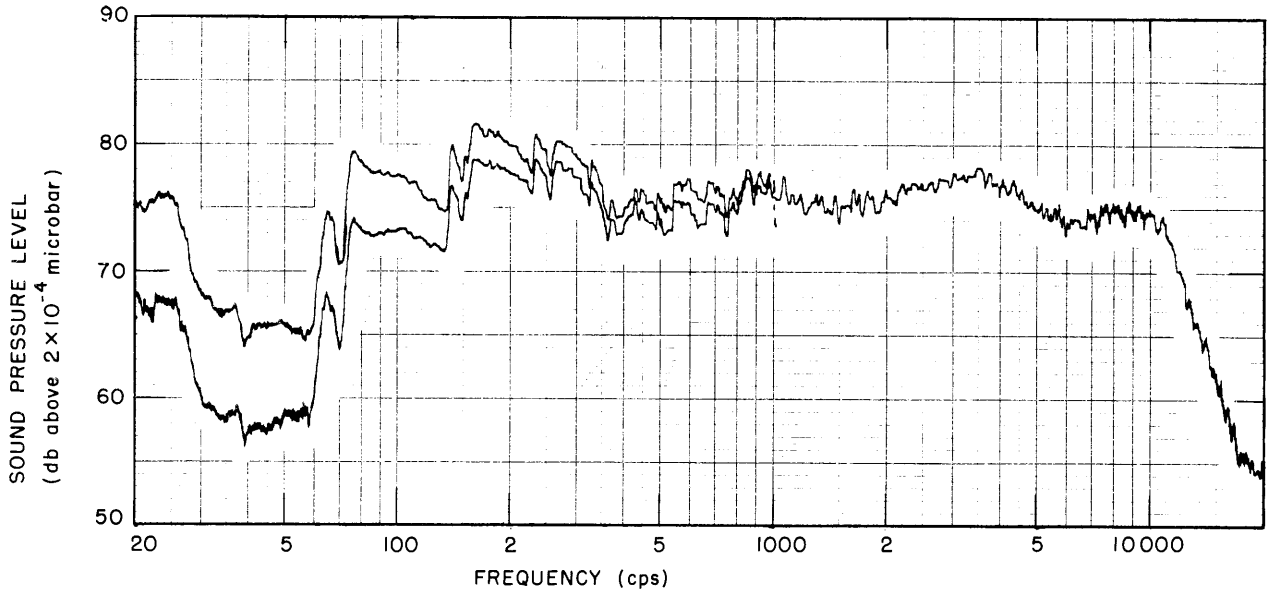


Fig. X-10. Response of the electrostatic loud-speaker measured in an anechoic chamber.

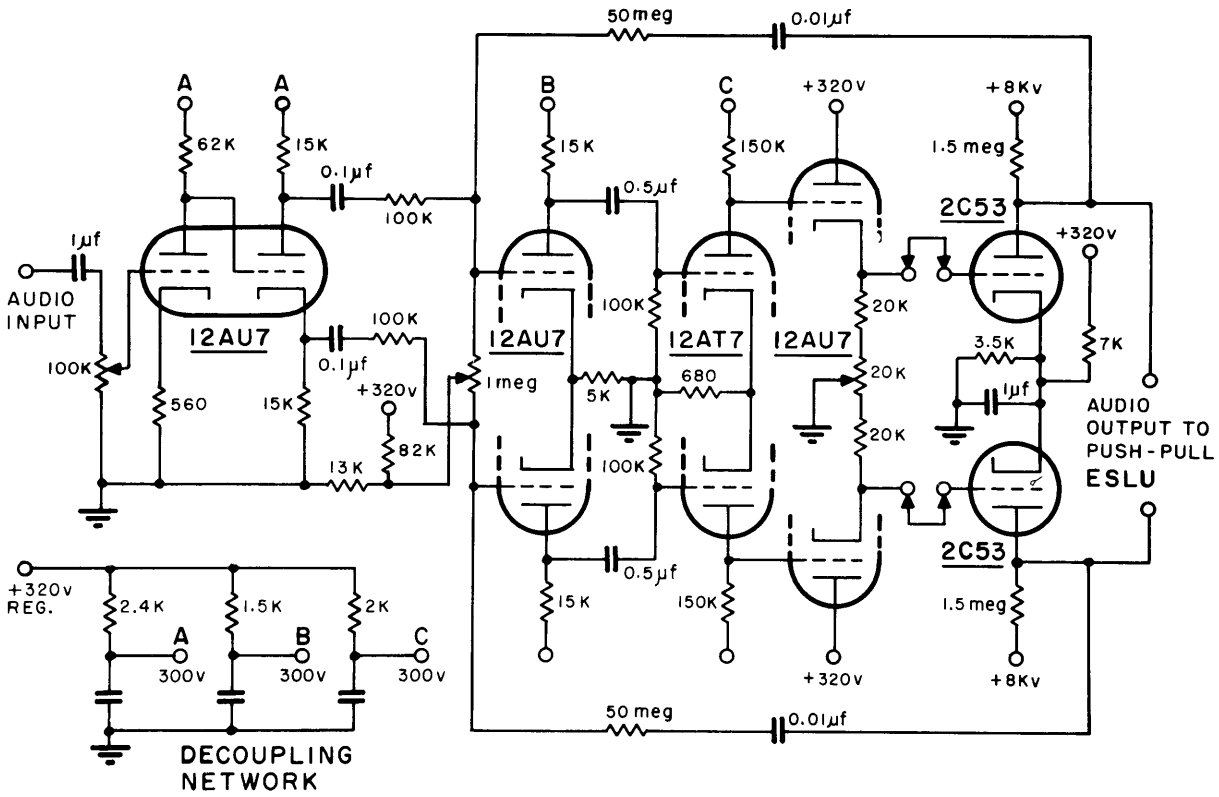


Fig. X-11. High-output-voltage audio amplifier.

inherently low value of harmonic and intermodulation distortion. The advantages of the electrostatic loud-speaker over the conventional magnetic speaker result from the distributed force drive system that is used in conjunction with the light diaphragm.

There are, of course, many other factors that must be investigated before the best design for a wide-range loud-speaker is achieved. The present prototype is a long way from the living room, but it does give definite indication that the electrostatic loud-speaker can serve as a wide-range sound source for high-quality reproducing systems.

C. I. Malme

F. STUDIES OF THE PERCEPTION OF SPEECH-LIKE SOUNDS*

Recently we have begun to examine the feasibility of utilizing non-speech sounds that appear in speech-like sequences as a means for studying the perception of speech. Such an approach is considered desirable for two reasons. First, in the traditional psychophysical approach to hearing, experiments have utilized relatively simple stimuli, and little, if any, generalization from the results of the psychophysical experiments to studies of the perception of speech has been possible. Second, and perhaps more important, most studies of the perception of speech have involved the use of stimuli that were, in fact, speech – phonemes, syllables, words, sentences or phrases. While, at first glance, such signals seem to be the obvious choice for use in these studies, they present the experimenter with difficult problems. Speech sounds are tremendously over-learned and, because of their familiarity to listeners, cannot be specified with any precision from a perceptual viewpoint. Furthermore, it is difficult to provide precise physical descriptions of natural speech stimuli.

In our initial studies we chose signals that appear in sequences of two components. The sequences were characterized by various amplitudes, durations, and frequency spectra of the two components relative to each other. Typical signals are shown in Fig. X-12. These signals are only two from an ensemble of stimuli of equal length, in which the components vary in intensity over a 30-db range, and in duration from 20 to 280 msec. The frequency spectra of these signals is the same for both components; in our initial experiments it was a bandpass noise of 600-2400 cps.

We asked listeners to make two kinds of judgment about signals of this type. In effect, we have tapped two kinds of response processes that are related to the same signal sequences. Because the signals varied both in intensity and duration, we could ask the

* This work was supported in part by National Science Foundation.

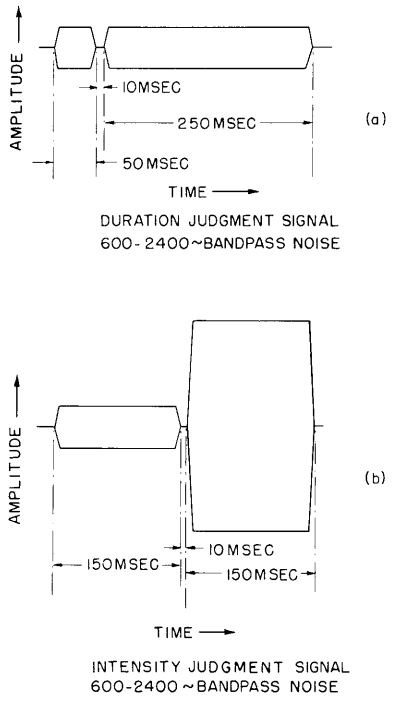


Fig. X-12. Typical signal envelopes. (a) The two components of the signal are of equal amplitude but of different duration. (b) The two components of the signal are of equal duration but of different amplitude.

listeners to judge whether the first half of the signal was either louder or longer than the second part.

Preliminary results show the following trends:

1. It appears that accurate loudness judgments are extremely difficult to make when the duration of the first component is 100 msec or less.

2. To some extent, a trading relationship exists between duration and intensity. That is to say, a weak long sound may be heard as louder than a relatively more intense sound of shorter duration. This relationship does not seem to be of the Intensity × time = constant type. Furthermore, it seems to be unsymmetrical on either side of the equality point.

3. The judgment of duration does not appear to be affected if the spectral compositions of the two components differ markedly.

This research is being continued.

T. T. Sandel, K. N. Stevens

G. VOWEL INTONATION CONTOURS*

Sentences that are otherwise similar may differ only in their intonation; John! and John? are two different sentences (1). According to some theories, intonation is produced by variations in the fundamental frequency; other theories propose more complex criteria. This report covers some results of perceptual tests in which intonation was

* This work was supported in part by National Science Foundation.

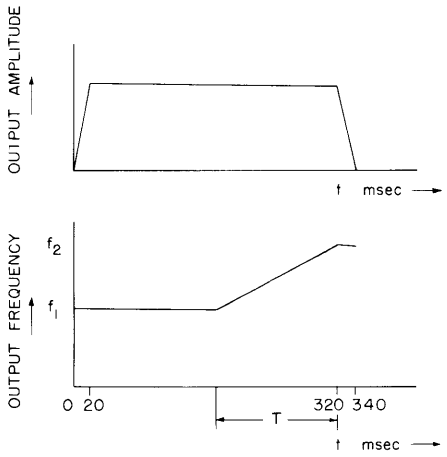


Fig. X-13. Buzz generator output.

simulated by variations in the fundamental frequency only.

Synthetic four-formant vowels were recorded with 32 different intonation contours by varying the fundamental frequency as a function of time. A fixed Pole Voice Synthesizer (POVO) was excited by a buzz generator that produced the waveforms described in Fig. X-13. The values of f_1, f_2, and T were selected on the basis of an earlier experiment in which the glottal excitation frequency was measured as a function of time in various sentence intonation patterns spoken by native male speakers of American English. All of the contours determined by the values of f_1, f_2, and T in Table X-1 were recorded

Table X-1. Parameters f_1, f_2, and T for Intonation Contours 1-32.

$f_1 \rightarrow f_2$ (cps)		T (msec)			
		75	150	225	300
50	100	1	2	3	4
50	200	5	6	7	8
100	200	9	10	11	12
100	150	13	14	15	16
200	50	17	18	19	20
200	100	21	22	23	24
100	50	25	26	27	28
100	66	29	30	31	32

Contours

Formant Frequencies (cps) of Synthesized Vowels

	/a/	/u/	/i/
F_1	715	284	264
F_2	1138	876	2415
F_3	2348	2243	2814
F_4	3242	3242	3242

for the vowels /a/, /u/, and /i/.

Three quantities, X, Y, and Z, were derived in terms of the synthesized intonation contours. X is defined as the direction of the fundamental frequency shift during time T; it is a binary quantity and either rises or falls. Y is defined as the rate of change of the fundamental frequency during time T:

$$Y = \begin{cases} \dfrac{f_2}{f_1 T}, & \dfrac{f_2}{f_1} > 1 \\ \dfrac{f_1}{f_2 T}, & \dfrac{f_1}{f_2} > 1 \end{cases}$$

$Z = |f_2 - f_1|$ is defined as the magnitude of the fundamental frequency shift.

The perceptual effects of these variations were investigated, first, in a test in which listeners were required to discriminate between acoustically different intonation patterns. The intonation contours of a single vowel were presented in groups of four. Three contours in each group were identical; the fourth differed only with respect to one of the derived properties, X, Y, or Z. All four contours started at the same frequency, f_1. The listeners were asked to identify the unique stimulus of each group. These stimuli were randomly distributed with equal probability in each of the four possible positions of each group. Approximately 110 groups were presented for each vowel in three separate half-hour listening tests.

The accuracy with which discriminations of the acoustically different contours were made did not differ substantially for the different vowels. The results are as follows.

For 8 subjects who made 2600 judgments, 97 per cent of the judgments made on the basis of property X were correct; 68 per cent of the judgments made on the basis of quantity Y were correct; and 88 per cent of the judgments made on the basis of property Z were correct. The listener with the highest score identified 99 per cent on the basis of X, 85 per cent for Y, and 95 per cent for Z. The listener with the lowest score identified 94 per cent on the basis of X, 49 per cent for Y, and 68 per cent for Z.

The reliability of discrimination was essentially unaffected by the degree of dissimilarity among the stimuli. This is obvious from the data for judgments made on the direction of the fundamental frequency shift. A detailed examination of the judgments made on the other two parameters revealed the same uniformity. For example, shifts of one-half octave were identified with the same facility as shifts of 2 octaves in the test for parameter 2. This seems to indicate that the values of parameters X, Y, and Z in this test were well above the discrimination levels (dl) for a discrimination test.

The results of this test indicate that discriminations can be reliably made on the basis of the direction of the frequency shift, or the magnitude of the shift, or both. The

rate of change of frequency does not seem to be as strong a cue as the other two factors. The shifts that were employed seem to be well above the discrimination level for discrimination in terms of these quantities. The results, however, cannot be used to demonstrate that listeners can identify particular contours. Discrimination merely sets an upper bound on the number of identifiable stimuli. For speech perception, however, the identification of the stimulus is the crucial problem. A second test was performed to investigate this problem.

The utterance /aha/ was synthesized with 11 different intonation contours with Rosen's Dynamic Analog Speech Synthesizer (2). These contours were selected from an ensemble of 32 combinations produced by varying the binary parameter X and the two quaternary parameters Y and Z. The basis of selection was that the contours should be within the normal experience of speakers of English. The synthesized intonation contours therefore correspond approximately to contours actually noted by Pike (3). Test tapes containing 4, 6, 8, and 11 different contours were presented to three subjects in separate listening tests. The listeners were taught to identify the different /aha/'s by using any notes or mnemonic devices that they could devise. One subject had been trained in the Smith-Trager notation and used it, while the other two subjects, who were untrained, used contour methods spontaneously, drawing pictures of how they thought the fundamental frequency went up or down. Each subject made 400 judgments. The intonation contours of the stimuli are presented in Fig. X-14.

The percentage of correct responses is plotted against the number of different stimuli in Fig. X-15a. The correct identifications are analyzed again in Fig. X-15b with respect to the number of bits of information gained, following the methods outlined by Pollack (4). The maximum information transfer of the identification test is approximately 2.6 bits. This is somewhat greater than that reported for a unidimensional ensemble of pure tones.

The confusions between pairs of contours are presented in Fig. X-16 with respect to the derived acoustic parameters in which these pairs differ. The greatest confusions generally occur when a pair of contours differ only with respect to one acoustic parameter. The next greatest source of confusion is between pairs of contours that differ only with respect to two parameters. There is little or no confusion when pairs of contours differ with respect to all three parameters. For example, 16 per cent of the judgments show confusions between contours 4 and 5, which differ only with respect to the magnitude of their frequency shifts; 17 per cent of the judgments show confusions between contours 3 and 9, which differ only with respect to the rate of change of their fundamental frequencies; and contours 2 and 11, which differ with respect to all three acoustic parameters, were not confused. The percentage of error in Fig. X-16 has been adjusted to take account of the increased error caused by the greater number of stimuli presented in the tests that involved more than four contours.

In conclusion, it is obvious that intonation judgments can be correlated with variations

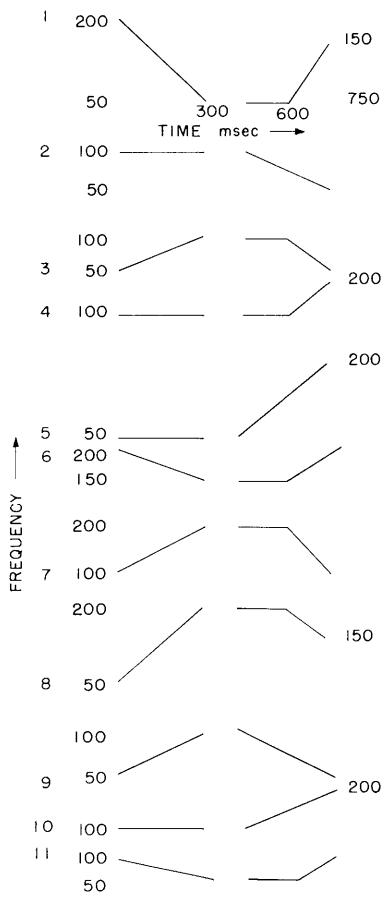


Fig. X-14. Fundamental frequency of utterances versus time.

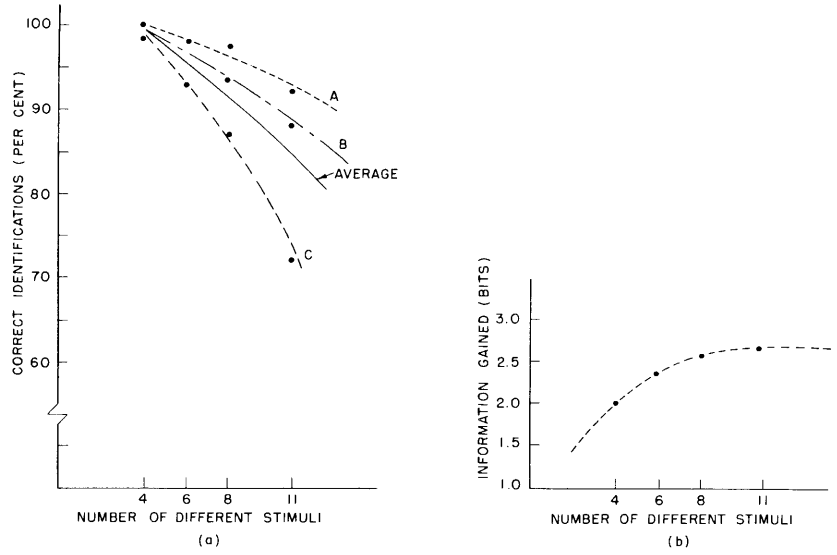


Fig. X-15. (a) Percentage of correct identifications. Curve A, trained listener; curves B and C, untrained listeners. (b) Information gained (measured in binary bits) for average identification versus number of different stimuli.

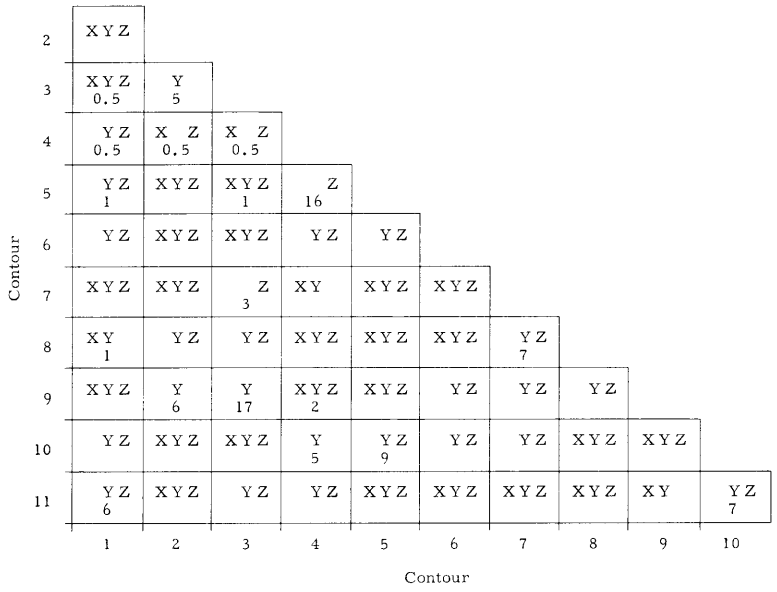


Fig. X-16. Confusions between contours (per cent) with respect to the derived acoustic quantities in which the contours differed.

in the fundamental frequency. It also seems probable that the direction, magnitude, and rate of change of the fundamental frequency can be used as cues for the identification of different intonation contours.

P. Lieberman

References

1. N. Chomsky, Syntactic Structures (Mouton and Company, 'S-Gravenhage, 1957).
2. G. Rosen, Dynamic analog speech synthesizer, Quarterly Progress Report, Research Laboratory of Electronics, M.I.T., April 15, 1958, p. 102.
3. K. L. Pike, The Intonation of American English (University of Michigan Press, Ann Arbor, Michigan, 1945).
4. I. Pollack, Papers on elementary auditory stimuli, J. Acoust. Soc. Am. 27, 1008 (1955); 28, 906, 153A, 412 (1956).