opening (an antinode). The second formant has one node and antinode more along the tract, the node at the side of the mouth, the antinode at the side of the larynx. These can shift when the cavities are tuned in two different ways for the same vowel, there being many degrees of freedom by the adjustment of lower jaw, tongue, palate and constriction of the pharynx wall, and only a few critical formants for each vowel.‡

The extra node of the second formant will be located somewhere near the soft palate, e.g. at 1000 cps a quarter-wavelength in free air of 37°C is 8.8 cm, with the result that the soft palate lies in a region of relative high pressures and vibrates considerably, the exact location of the node depending on the singing pedagogical treatment.

So we conclude that the resonances in the head are caused by the second and not by the first formant of the cavities.

ACKNOWLEDGMENT

The author wishes to express his appreciation to Dr. Kenneth N. Stevens of the Massachusetts Institute of Technology for his generous editorial assistance.

‡ All formants have a node at the larynx, the throat wall thus being the proper place for a wall microphone.

THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA VOLUME 27, NUMBER 2 MARCH, 1955

An Analysis of Perceptual Confusions Among Some English Consonants

GEORGE A. MILLER AND PATRICIA E. NICELY
Lincoln Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts

(Received December 1, 1954)

Sixteen English consonants were spoken over voice communication systems with frequency distortion and with random masking noise. The listeners were forced to guess at every sound and a count was made of all the different errors that resulted when one sound was confused with another. With noise or low-pass filtering the confusions fall into consistent patterns, but with high-pass filtering the errors are scattered quite randomly. An articulatory analysis of these 16 consonants provides a system of five articulatory features or "dimensions" that serve to characterize and distinguish the different phonemes: voicing, nasality, affrication, duration, and place of articulation. The data indicate that voicing and nasality are little affected and that place is severely affected by low-pass and noisy systems. The indications are that the perception of any one of these five features is relatively independent of the perception of the others, so that it is as if five separate, simple channels were involved rather than a single complex channel.

THE over-all effects of noise and of frequency distortion upon the average intelligibility of human speech are by now rather well understood. One limitation of the existing studies, however, is that results are given almost exclusively in terms of the articulation score, the percentage of the spoken words that the listener hears correctly. By implication, therefore, all of the listener's errors are treated as equivalent and no knowledge of the perceptual confusions is available. The fact is, however, that mistakes are often far from random. A closer look at the problem suggests that we might learn something about speech perception and might even improve communication if we knew what kinds of errors occur and how to avoid the most frequent ones. Such was the reasoning that led to the present study.

Perhaps the major reason that confusion data are not already available is the cost of collecting them. Every phoneme must have a chance to be confused with every other phoneme and that large number of potential confusions must be tested repeatedly until statistically reliable estimates of all the probabilities are obtained. Such data are obtained from testing programs far more extensive than would be required to evaluate some specific system.

In order to reduce the magnitude of the problem to more manageable size, we decided to study a smaller set of phonemes and to explore the potential value of such data within that smaller universe. Since the consonants are notoriously confusable and are quite important for intelligibility, we decided to begin with a comparison of 16 consonants: |*p*|, |*t*|, |*k*|, |*f*|, |*θ*|, |*s*|, |ʃ|, |*b*|, |*d*|, |*g*|, |*v*|, |ð|, |*z*|, |ʒ|, |*m*|, and |*n*|. These 16 make up almost three quarters of the consonants we utter in normal speech and about 40 percent of all phonemes, vowels included. It was our suspicion that when errors begin to occur in articulation tests, the culprits would usually be found among this set of 16 phonemes. A further reason for being interested in consonants is that the information-bearing aspects of these sounds are less well understood than is the case for vowels; we hoped to pick up some clues as to what the important features of these phonemes might be.

The major portion of the work to be reported here was done with the aforementioned 16 consonants. However, a number of other, even smaller, experiments were conducted with subsets of those 16. In general, the results of the smaller studies agree with and support the conclusions of the larger study. These results will be introduced into the discussion where appropriate, but the major emphasis will be placed on the 16-consonant data.

EXPERIMENTAL PROCEDURES

Five female subjects served as talkers and listening crew; when one talked, the other four listened. Since the tests lasted several months, some of the original crew members departed and were replaced; care was taken to train new members adequately before their data were used. The subjects were, with one Canadian exception, citizens of the United States. None had defects of speech or hearing and all were able to pronounce the 16 nonsense syllables without any noticeable dialect. Since rhythm, intonation, and vowel differences were not involved, we have assumed that regional differences in speech habits were not a significant source of variability in the data.

The 16 consonants were spoken initially before the vowel |*a*| (f*a*ther). The list of 200 nonsense syllables spoken by the talker was prepared in advance so that the probability of each syllable was 1 in 16 and so that their order was quite random within the list and from one list to the next. The syllables were spoken at an average rate of one every 2.1 seconds and the listeners were forced to respond—to guess, if necessary—for every syllable. When the speech was near the threshold of hearing, the listeners were kept in synchrony with the talker by a tone that was turned on at fixed intervals. Otherwise, a 2.1-second pause was inserted after every block of five syllables. With four listeners, there were 800 syllable-response events per talker for which confusions could be studied. Pooling the five talkers gives us 4000 observations at each condition tested.

At the completion of each test of 200 syllables, the talker went from the control room back to the test room and the crew proceeded to tabulate their responses. Each listener had a table showing what syllable was spoken and what syllable she had written in response; each cell of the table represented one of the $16\times 16=256$ possible syllable-response pairs, and the number entered in that cell was the frequency with which that syllable-response pair occurred. We shall refer to such tables as "confusion matrices."

A headrest on the talker's chair insured that the distance to the WE-633A microphone was constant at 15 inches. The speech 15 inches from the talker's lips was about 60 db re 0.0002 dyne/cm^2. The speech voltage was amplified, then filtered (if frequency distortion was to be used), then mixed with noise, then amplified again and presented to the listeners by PDR-8 earphones. In all tests the noise voltage was fixed at -32 db below one volt across the earphones and the signal-to-noise ratio was varied by changing the gain in the speech channel. A separate amplifier was used to drive a monitoring VU-meter with the output of the microphone. The gain to the VU-meter was fixed so that the talker could maintain her speech level at a constant value. The talkers did succeed rather well in keeping a constant level; several hundred sample readings of peak deflections gave an average of $+0.18$ VU with a standard deviation of 1.04. However, it should be noted that with this system, the signal-to-noise ratios are set by the peak deflection of the VU needle and that peak occurs during the vowel. The consonants, which are consistently weaker than the vowel, were actually presented at much less favorable signal-to-noise ratios than such a vowel-to-noise ratio would seem to indicate. It was, therefore, especially important to keep the same speech level for all tests since otherwise the vowel-to-consonant ratio might have changed significantly and the data would not be comparable.

The frequency response of the system was essentially that of the earphones, which are reasonably uniform between 200 and 6500 cps. A low-pass filter at 7000 cps in the random noise generator insured that noise voltages could be converted directly to sound pressure levels according to the earphone calibration. A Krohn-Hite 310-A variable band-pass filter was used to introduce frequency distortion into the speech channel; the skirts dropped off at a rate of 24 db per octave and the cutoff frequency was taken as the frequency 3 db below the peak in the pass band.

RESULTS

The results of these tests are confusion matrices. Since these matrices represent a considerable investment and since other workers may wish to apply summary statistics differing from those which we have chosen, the complete confusion matrices are presented in Tables I–XVII. Data for all listeners and all talkers have been pooled so that 4000 observations are summarized in each matrix; on the average, each syllable was judged 250 times under every test condition.

Tables I–VI summarize the data obtained when the speech-to-noise ratio was -18, -12, -6, 0, $+6$, and $+12$ db and the band width was 200–6500 cps. Tables VII–XII summarize the data when the high-pass cutoff was fixed at 200 cps and the low-pass cutoff was 300, 400, 600, 1200, 2500, and 5000 cps with a speech-to-noise ratio corresponding to $+12$ db for unfiltered speech. Tables XII–XVII summarize the data when the low-pass cutoff was fixed at 5000 cps and the high-pass cutoff was 200, 1000, 2000, 2500, 3000, and 4500 cps with a speech-to-noise ratio that would have been $+12$ db if the speech had not been filtered.

In these tables the syllables that were spoken are indicated by the consonants listed vertically in the first column on the left. The syllables that were written by the listener are indicated horizontally across the top of the table. The number in each cell is the frequency that each stimulus-response pair was observed. The number of correct responses can be obtained by totalling the frequencies along the main diagonal. Row sums would give the frequencies that each syllable was written by the listeners.

A GENERALIZATION OF THE ARTICULATION SCORE

The standard articulation score is obtained from Tables I–XVII by summing the frequencies along the main diagonal and dividing the total by *n*, the number of observations. Although this score is useful, it tells us nothing about the distribution of errors among the off-diagonal cells. If we wanted to reconstruct an adequate picture of the confusion matrix, we would need other scores to supplement the usual articulation score.

In order to generalize the articulation score, we can combine stimuli (and their corresponding responses) into groups in such a way that confusions within groups are more likely than confusions between groups. Combining stimuli creates a smaller confusion matrix that shows the confusions between groups, and the sum along the diagonal gives a new articulation score for this new, smaller matrix. The new score will be greater than the original score, since all the responses that were originally correct remain so and in addition all the confusions within each group are now considered to be "correct" in the new score. If the original score, A, is supplemented with such an additional score, A', we would reconstruct the data matrix by spreading the fraction A along the main diagonal. Then $A'-A$ would go off the diagonal but within groups, and $1-A'$ would be distributed off the diagonal between groups. This general strategy can be repeated quite simply if the several groupings used form a monotonic increasing sequence of sets: $A \leq A' \leq A''$, etc.

A simple example will illustrate this technique. A test was conducted at $S/N = -12$ db over a 200–6500-cps channel using six stop consonants in front of the vowel $|a|$. The confusion matrix for 2000 observations

TABLE I. Confusion matrix for $S/N = -18$ db and frequency response of 200–6500 cps.

	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	14	27	22	23	25	22	14	15	16	7	17	11	12	11	16	12
t	16	26	21	15	15	18	14	7	10	6	17	9	13	11	9	13
k	20	22	24	15	14	29	12	4	11	9	12	10	16	11	17	14
f	27	22	27	23	13	12	10	19	20	14	16	16	15	3	13	18
θ	17	18	18	13	15	21	12	14	20	14	23	6	14	9	12	14
s	18	17	23	11	18	21	17	11	24	15	15	16	11	13	17	5
ʃ	16	20	27	17	13	37	14	10	21	7	20	18	9	8	16	15
b	12	11	24	15	19	15	12	24	20	19	24	12	15	11	18	17
d	16	24	18	13	15	15	14	22	25	21	25	17	18	13	15	25
g	11	20	29	9	18	18	15	26	30	14	18	14	16	20	24	22
v	9	17	18	11	7	12	9	25	14	13	15	15	19	11	12	17
ð	16	11	10	7	6	14	10	20	17	18	15	7	17	12	18	18
z	18	18	15	9	13	19	7	22	14	9	21	12	23	10	22	12
ʒ	8	16	17	14	12	15	7	22	18	8	15	11	15	11	18	13
m	19	24	15	14	14	14	8	14	15	12	13	8	11	6	25	28
n	11	18	20	6	9	18	9	14	14	13	9	8	10	12	33	32

TABLE II. Confusion matrix for $S/N = -12$ db and frequency response 200–6500 cps.

	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	51	53	65	22	19	6	11	2		2	3	3	1	5	8	5
t	64	57	74	20	24	22	14	2	3	1	1	2	1	1	5	1
k	50	42	62	22	18	16	11	4	1	1	1	2			4	2
f	31	22	28	85	34	15	11	3	5		8	8	3		3	
θ	26	22	25	63	45	27	12	6	9	3	11	9	3	2	7	2
s	16	15	16	33	24	53	48	3	5	6	3	1	6	2		1
ʃ	23	32	20	14	27	25	115	1	4	5	3		6	3	4	2
b	4	2	2	18	7	7	1	60	18	18	44	25	14	6	20	10
d	3		1	4	7	4	11	18	48	35	16	24	26	14	9	12
g	3	1	1	1	4	5	7	20	38	29	16	29	29	38	10	9
v		1	1	12	5	4	5	37	20	23	71	16	14	4	14	9
ð		1	4	17	2	3	2	53	31	25	50	33	23	5	13	6
z	6	1	2	2	6	14	8	23	29	27	24	19	40	26	3	6
ʒ	3	2	2	1		6	7	7	30	23	9	7	39	77	5	14
m		1			1	1		11	3	6	8	11		1	109	60
n	1			1		1		2	2	6	7	1	1	9	84	145

Table III. Confusion matrix for $S/N=-6$ db and frequency response of 200–6500 cps.

	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	80	43	64	17	14	6	2	1	1		1	1			2	
t	71	84	55	5	9	3	8	1				1	2		2	3
k	66	76	107	12	8	9	4					1			1	
f	18	12	9	175	48	11	1	7	2	1	2	2				
θ	19	17	16	104	64	32	7	5	4	5	6	4	5			
s	8	5	4	23	39	107	45	4	2	3	1	1	3	2		1
ʃ	1	6	3	4	6	29	195		3							1
b	1			5	4	4		136	10	9	47	16	6	1	5	4
d							8	5	80	45	11	20	20	26	1	
g					2			3	63	66	3	19	37	56		3
v				2		2		48	5	5	145	45	12		4	
ð					6			31	6	17	86	58	21	5	6	4
z					1	1	1	7	20	27	16	28	94	44		1
ʒ								1	26	18	3	8	45	129		2
m	1							4			4	1	3		177	46
n					4			1	5	2		7	1	6	47	163

Table IV. Confusion matrix for $S/N=0$ db and frequency response of 200–6500 cps.

	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	150	38	88	7	13											
t	30	193	28	1												
k	86	45	138	4	1		1									1
f	4	3	5	199	46	4		1				1			1	
θ	11	6	4	85	114	10					2					
s		2	1	5	38	170	10			2						
ʃ		3	3			3	267									
b				7	4			235	4		34	27	1			
d									189	48		4	8	11		
g									74	161		4	8	25		
v				3	1			19		2	177	29	4	1		
ð								7		10	64	105	18			
z									17	23	4	22	132	26		
ʒ									2	3	1	1	9	191		1
m								1							201	6
n												3		1	8	240

Table V. Confusion matrix for $S/N=+6$ db and frequency response of 200–6500 cps.

	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	162	10	55	5	3							1				
t	8	270	14													
k	38	6	171	1												
f	5	1	2	207	57			3			1					
θ	5	1	2	71	142	3					2	2				
s		1		1	7	232	2			1						
ʃ						1	239									
b				1	2			214			31	12				
d									206	14		9	1	2		
g								11	64	194		4	2	1		
v				1	1			14		2	205	39	5			1
ð								2		4	55	179	22	2		
z									3	10	2	20	198	3		
ʒ									3	4			2	215		
m															217	3
n									1						2	285

TABLE VI. Confusion matrix for $S/N=+12$ db and frequency response of 200–6500 cps.

	p	t	k	f	θ	s	ʃ	b	d	g	v	ð	z	ʒ	m	n
p	240		41	2	1											
t	1	252	1	1						1						
k	18	3	219													
f				225	24			5			2					
θ	9		1	69	185			3				1				
s						232										
ʃ							236									
b					1			242			24	12	1			
d									213	22			1			
g					1				33	203		3				
v								6			171	30			1	
ð					1			1		3	22	208	4			1
z									2	4	1	7	238			
ʒ														244		
m												1			274	1
n																252

TABLE VII. Confusion matrix for $S/N=+12$ db and frequency response of 200–300 cps.

	p	t	k	f	θ	s	ʃ	b	d	g	v	ð	z	ʒ	m	n
p	47	61	68	15	11	17	9	3	3	1		1	2	2	3	1
t	59	63	64	19	15	14	13	3	4	1		5	2	2	2	2
k	37	47	56	10	13	15	10	1	2	1		2		1		1
f	21	29	21	38	37	47	19	2	2	1		2	2	3	3	1
θ	13	23	25	23	39	54	39	2	2	1		5	1		4	5
s	16	25	10	29	52	65	34	1	4	2	4	5	1	1	1	2
ʃ	15	33	23	18	28	70	41	1	1			7	3	1	1	2
b		1	1	8	8	5	3	98	28	17	38	19	9	2	8	7
d	1		1	11	7	12	5	70	84	33	12	10	24	9	1	
g	4	1	2	7	5	13	8	56	74	33	13	15	21	13	6	1
v		2	1	1	2	1	1	44	34	18	77	34	36	14	2	1
ð	1				3		1	22	16	19	45	46	45	23	11	8
z	2	3	2	2	4	3	2	15	15	20	46	35	64	21	2	
ʒ	1	1		1	2		1	11	15	24	54	42	70	39	2	5
m			1	1	2	2		1	3	3	4	5	1	4	161	60
n	1	3	2	1	1	1	2	1	3	2	2	4	2	2	133	108

TABLE VIII. Confusion matrix for $S/N=+12$ db and frequency response of 200–400 cps.

	p	t	k	f	θ	s	ʃ	b	d	g	v	ð	z	ʒ	m	n
p	72	68	90	20	15	4	1	2	4	1		1				2
t	73	72	74	20	8	6	3	1	2	2		2		1		
k	63	74	127	9	7	5	2			1		1	1	1		1
f	7	7	10	63	69	41	8	3	1	1	1	3		1	1	
θ	5	8	11	60	85	45	14	2	4	2	6	5	1			
s	1	6	5	19	49	125	60	5	2	1	2	9	4			
ʃ	2	6	8	8	22	69	89	2	4	1		3	5	1		
b		1	1	19	14	5		134	20	13	14	11	4	1	2	1
d			2		1	6	4	19	120	23	2	3	11	3		2
g			2	1		5	1	11	116	59	8	7	11	4	1	2
v		1		1	1	2		25	4	8	111	55	18	2	2	2
ð		1	1	6	5	1		43	16	15	75	66	23	11	1	4
z	2		2	1	5	5	2	21	20	17	18	33	91	25	1	1
ʒ					4		2	1	27	29	11	16	83	78	1	
m								12	3		1				219	57
n					1	1		12	3	1	1	2			99	120

TABLE IX. Confusion matrix for $S/N = +12$ db and frequency response of 200–600 cps.

	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	115	43	70	10	3	2						1				
t	69	63	71	4	4							1				
k	59	49	134	4	1						1					
f	2	3	2	126	89	11	1	2			1	8	1		1	1
θ	2	1	1	103	97	35	7	2	1		5	1				1
s	3	3		34	88	93	26	4	1			7		1		
ʃ	3	6	12	7	31	98	87	1	2	1	2	1	1			
b			1	10	5	1		201	13		13	4				
d		1		1	1	6	1	29	169	39	3	3	6	5		
g				1		7		12	99	97		4	8	11		1
v				5	2			14	1	2	141	57	9	4	1	
ð								10	6	10	109	90	31	7	1	
z						1	2	3	15	30	17	42	116	22		
ʒ			1				1		10	21	8	17	110	116		
m						1							1		215	39
n				1											119	120

TABLE X. Confusion matrix for $S/N = +12$ db and frequency response of 200–1200 cps.

	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	165	46	31	3	1				1				1			
t	91	83	68	4	1	2			1			2				
k	48	55	147	2	3							1				
f	16	4	3	146	60	3	2	11			1	2				
θ	4	3		109	76	17	2	12	1			2	1	1		
s	2	1	1	43	83	83	11	3		1	1	7				
ʃ	1	6	2	12	41	86	90		6	4		4				
b				14	5			223	4		5	1				
d	1				1	3	4	4	173	37		2	1	2		
g	1					1			102	107	1	2	7	7		
v	2	2		2	1			23	1	2	163	62	14	3	1	
ð				1		3	2	27	6	32	87	107	36	7		
z	1							4	12	48	10	15	114	39		1
ʒ							1		3	35	1	16	60	134	2	
m	1									1					229	9
n															5	247

TABLE XI. Confusion matrix for $S/N = +12$ db and frequency response of 200–2500 cps.

	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	215	29	26	5	1											
t	74	91	47													
k	15	16	201													
f	6		1	186	31	2		3				7				
θ	1	5	1	93	81	25	1	1		2	2	4				
s	1	3	1	31	78	142	9	1		1		5				
ʃ		1	1			23	210			1						
b				11	6	1		206	4		11	1				
d							1	1	217	30			1	6		
g				2		1	1	1	54	169		1		3		
v				1	2	1		36		1	178	39	9	1	1	
ð				3	6	2		14		17	58	146	45	1		
z						2			17	40	7	24	122	20		
ʒ				1			5		5	9			11	265		
m															242	18
n															2	242

TABLE XII. Confusion matrix for $S/N=+12$ db and frequency response of 200–5000 cps.

	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	228	7	7	1			1									
t		236	8													
k	26	5	213													
f	6	1	1	194	35			3			1	3				
θ		2	2	96	146	2		2	1		1	8				
s		2		1	31	204	1	1	9	4		7				
ʃ						1	243									
b				13	12			207	2	3	19	8				
d									240	9				3		
g								1	41	199			2	1		
v				3	3			20		2	182	47	2			1
ð					7			10	3	22	49	170	19			
z					1			3	8	24	2	22	145	3		
ʒ							1		2				13	264		
m															213	11
n																248

TABLE XIII. Confusion matrix for $S/N=+12$ db and frequency response of 1000–5000 cps.

	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	179	9	44	6	3					2	1					
t		272	3					1								
k	15	1	227					1	1		2					1
f	12	1		162	28	3	1	34			6		1		4	
θ	8	2	7	39	125	13	2	6	2	1	4	19	3		1	
s				3	28	200		2	1	1	4	6	9	1		1
ʃ						1	221							2		
b	2			9	10	1		130		6	74	24			16	
d		2					1		195	35	6	2	2	8		5
g				2					48	151		3	4	5		11
v	1			28	8			48	1	3	145	33	3		17	1
ð	1			1	14			8	11	12	31	116	26	5	21	6
z		1			2	24	2	1	19	7	3	31	163	4	2	1
ʒ				1			20		2	2				207		
m	3		2	5	4	1		10			6				224	1
n			1	1	1			1	8	4	2	1	1	1		207

TABLE XIV. Confusion matrix for $S/N=+12$ db and frequency response of 2000–5000 cps.

	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	94	32	26	15	6	3	1	10	4	4	13	12	1	5	3	3
t	7	223	3	3	1		3		7	1	1	1		5	1	
k	24	25	126	4	7	4	2	3	6	15	1	3	1	2	7	2
f	38	7	19	72	24	5	2	24	3	12	28	11	4	3	12	4
θ	22	7	11	20	63	27		19	8	13	22	26	16		12	10
s	2	9	1	5	23	148			4	3	3	4	44	6		8
ʃ	1	1					208	1					1	28		
b	15	5	5	37	12	2		72	7	8	40	30	4		40	7
d	2	6	7		2			4	192	19	4	6	3	2	2	23
g	2	1	3	1	8	4	1	8	44	122	10	6	6	1	3	20
v	17	1	12	13	7		1	39	5	14	42	23	2	4	32	12
ð	5		6	9	20	5		17	16	19	17	64	20	1	36	25
z	3	2	2	5	8	44		5	22	7	1	13	99	5	7	9
ʒ							37			4				199	4	
m	10	4	3	8	7		1	9	5	10	10	16	2		113	26
n	2		2		3	2		1	20	11	3	7	6	3	4	192

TABLE XV. Confusion matrix for $S/N = +12$ db and frequency response of 2500–5000 cps.

	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	69	30	37	26	16	4	4	21	9	18	13	12	9	3	7	10
t	4	164	9	2	2	2		1	4	4	1	2	2		3	
k	20	35	76	9	11	5	6	3	5	25	5	3	15	11	7	4
f	27	8	7	24	28	7	8	15	8	14	34	14	6	2	11	11
θ	15	19	7	20	49	10	8	12	16	16	13	20	10	5	16	16
s	6	8	2	1	19	160	4		16	10	8	11	27	2	7	11
ʃ	1	1	2	1	5	1	204	1				1	2	44		1
b	23	4	10	13	17		2	48	17	17	34	28	10	1	28	12
d	1	7	6	5	4	2	1	1	128	16	8	6	5	13	5	16
g	6	3	16	5	6	5	2	17	39	85	11	13	6	7	6	13
v	22	6	6	26	18	3	3	33	12	9	32	28	7	2	18	7
ð	21	11	9	16	28	4	2	35	14	22	20	44	10	2	24	22
z	4	5	1	2	9	60	5	1	27	21		12	86	6	2	3
ʒ	2	4	2			3	49	1	7	1	2	1	5	167		
m	18	3	7	11	16	8	2	13	16	12	16	21	3	1	68	37
n	8	4	12	7	9	2		10	22	17	13	8	5	4	16	119

TABLE XVI. Confusion matrix for $S/N = +12$ db and frequency response of 3000–5000 cps.

	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	31	15	15	15	14	11	6	19	11	8	15	15	5	9	12	19
t	11	184	16	6	5	5	5	8	9	3	4	2	5	3	6	4
k	15	35	50	7	16	7	2	14	14	24	7	9	8	9	8	7
f	19	12	12	15	19	8	2	25	16	25	15	12	6	2	17	11
θ	15	14	13	13	30	15	3	15	24	12	14	17	10	3	14	20
s	4	4	8	11	8	140	4	7	8	6	6	11	35	7	2	7
ʃ		6	2	3	1	4	177	1	2	2	1	6	1	23	7	
b	17	13	11	25	23	8	1	27	13	19	25	13	5	6	17	13
d	14	23	15	11	11	4	3	15	63	25	14	10	13	6	19	14
g	14	15	17	17	12	8	1	23	39	45	14	10	13	7	17	16
v	19	19	22	18	20	8	10	35	18	16	19	21	7		28	16
ð	19	13	12	12	24	8	6	22	24	15	24	21	10	5	33	16
z	9	21	9	7	17	59	6	6	11	13	10	15	41	4	10	14
ʒ	4	6	1	5	1	11	51	3	3	7	1	10	9	128	7	5
m	16	7	14	11	19	5	4	31	16	17	17	10	10	6	58	19
n	16	7	12	6	16	7	6	14	29	16	13	22	7	4	19	58

TABLE XVII. Confusion matrix for $S/N = +12$ db and frequency response of 4500–5000 cps.

	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	26	21	23	16	24	20	4	15	16	14	20	9	10	9	16	9
t	10	141	12	3	4	4	3	5	11	5	7	11	4	5	8	3
k	16	34	25	14	11	13	8	20	20	8	18	13	20	10	12	22
f	9	9	22	18	18	6	6	18	17	9	17	19	9	3	27	13
θ	16	21	25	5	20	10	2	29	23	24	27	28	11	5	16	10
s	8	5	15	7	11	138	7	6	4	11	13	7	34	5	6	7
ʃ	3	3	7	1	1	12	190	1	4	2	2	4	6	26	6	4
b	12	8	23	11	18	13	9	26	14	18	21	14	11	6	16	16
d	24	26	28	16	19	8	4	19	18	19	13	11	6	3	16	14
g	12	16	17	14	21	11	10	12	17	21	18	19	7	10	22	13
v	21	11	17	15	24	12	8	19	15	14	33	23	6	3	23	16
ð	18	19	15	16	20	7	5	24	16	16	22	28	9	11	24	10
z	8	12	8	8	7	64	5	12	10	9	12	17	51	11	6	8
ʒ	5	18	10	8	9	11	57	5	4	5	9	11	15	85	9	7
m	8	13	20	13	15	14	7	18	8	16	16	17	12	2	15	18
n	20	15	15	18	15	7	6	19	20	12	17	15	12	4	21	16

TABLE XVIII. Confusion matrix at $S/N=-12$ db with a 200–6500-cps channel.

	p	*t*	*k*	*b*	*d*	*g*	Sum
p	117	58	115	14	10	2	316
t	74	101	103	8	4	6	296
k	105	109	153	5	8	4	384
b	13	9	10	217	45	26	320
d	3	4	5	47	200	117	376
g	3	11	8	45	147	94	308
							2000

is given in Table XVIII. There are 882 entries on the main diagonal, so $A=0.441$. If we group the consonants $|pk|$, $|t|$, $|b|$, and $|dg|$, there are 1366 correct responses, so $A'=0.683$. If we again group $|ptk|$ and $|bdg|$, there are 1873 correct responses, so $A''=0.9365$. Now if we wish to reconstruct the matrix from these three articulation scores, we would first divide the 882 correct responses equally among the six diagonal cells, which gives 147 observations per cell. When we add the four cells for $|pk|$ and $|dg|$ to the diagonal cells, the count increases from 882 to 1366, so the additional 484 observations must be divided equally among the four additional cells, which gives 121 per cell for $|pk|$ and $|dg|$ confusions. When we add the eight remaining cells for the $|ptk|$ and $|bdg|$ groups, the count increases from 1366 to 1873, so the additional 507 observations must be divided evenly among those eight cells, which gives 63.4 per cell. The remaining 127 observations are then divided equally among the 18 cells remaining in the lower left and upper right quadrants, which gives 7.1 per cell. In this way the generalized, three-valued articulation score gives a reasonably clear picture of the distribution of errors.

The procedure just described can lead to serious errors if the stimulus frequencies are quite disparate. For example, if one stimulus is presented much more often than any other, it will contribute more to the total number of correct responses and then the equipartition of correct responses among the diagonal cells will be in error. In such cases the original data matrix should first be corrected to the frequencies that would presumably have been obtained if the stimuli had been equally frequent. This correction is made by multiplying the entries in each row by n/kn_i, where n_i is the frequency of occurrence of the ith stimulus ($i=1,2,\cdots,k$) in a sample of n observations. Then the "articulation scores corrected for stimulus frequencies" are calculated for the revised matrix. To reconstruct the data matrix, the corrected frequencies should be partitioned as before and then each row multiplied by kn_i/n in order to remove the correction and regain the original stimulus frequencies. Whenever an experimenter employs some unusual (nonuniform) distribution of stimulus frequencies, this fact should be stated explicitly in order to avoid misinterpretations of the articulation scores so obtained.

Some such generalization of the articulation score seems essential in order to preserve the data on clustering of errors. In our own analysis of the data, however, we have preferred a somewhat more elaborate statistical analysis. We have presented this simpler technique for the reader who feels that the information measures we have employed are too abstract or do not permit a simple reconstruction of the original matrix. Having pointed out this simpler technique, however, we shall make little use of it in the following discussion.

LINGUISTIC FEATURES

For many years linguists and phoneticians have classified phonemes according to features of the articulation process used to generate the sounds. These features of speech production are reflected in certain acoustic characteristics which are presumably discriminated by the listener. When we begin to look for reasonable ways to group the stimuli in order to summarize the pattern of confusions, it is natural to turn first to these articulatory features for guidance. In order to describe the 16 consonants used in this study we adopted the following set of features as a basis for classification.

(1) *Voicing*. In articulatory terms, the vocal cords do not vibrate when the consonants |*ptkfθsʃ*| are produced, and they do vibrate for |*bdgvðzʒmn*|. Acoustically, this means that the voiceless consonants are aperiodic or noisy in character, whereas a periodic or line-spectrum component is superimposed on the noise for voiced consonants. In addition, in English the voiceless consonants seem to be more intense and the voiceless stops have considerable aspiration, a sort of breathy noise between the release of pressure and the beginning of the following vowels, and may be somewhat briefer than the voiced stops. Thus the articulatory difference is reflected in a variety of acoustic differences.

(2) *Nasality*. To articulate $|m|$ and $|n|$ the lips are closed and the pressure is released through the nose by lowering the soft palate at the back of the mouth. The nasal resonance introduced in this way provides an acoustic difference. In addition, $|mn|$ seem slightly longer in duration than their stop or fricative counterparts and somewhat more intense. Also, the two nasals are the only consonants in this study lacking the aperiodic component of noisiness.

(3) *Affrication*. If the articulators close completely, the consonant may be a stop or a nasal, but if they are brought close together and air is forced between them, the result is a kind of turbulence or friction noise that distinguishes |*fθsʃvðzʒ*| from |*ptkbdgmn*|. The acoustic turbulence is in contrast to the silence followed by a pop that characterizes the stops and to the periodic, almost vowel-like resonance of the nasals.

(4) *Duration*. This is the name we have arbitrarily adopted to designate the difference between |*sʃzʒ*| and the other 12 consonants. These four consonants are

long, intense, high-frequency noises, but in our opinion it is their extra duration that is most effective in setting them apart.

(5) *Place of Articulation.* This feature has to do with where in the mouth the major constriction of the vocal passage occurs. Usually three positions, front, middle, and back, are distinguished, so that we have grouped |*pbfvm*| as front, |*tdθsðzn*| as middle, and |*kgʃʒ*| as back consonants. Although these three positions are easy to recognize in the production of these sounds, the acoustic consequences of differences in place are most complex. Of the various accounts of the positional feature that have been given, the work done by the Haskins Laboratory[1,2] seems to provide the best basis for an interpretation of our data. For the voiced stops |*bdg*| the most important acoustic clue to position seems to be in the initial portion of the second formant of the vowel |*a*| that follows; if this formant frequency rises initially, it is a |*b*|, but if it falls it is |*d*| or |*g*|. Since the vowel formant is relatively audible, the front |*b*| is easily distinguished from the middle |*d*| and the back |*g*|. The latter two positions are much harder to distinguish and probably cannot be differentiated until their aperiodic, noisy components become sufficiently audible so that high-frequency noise can be assigned to middle |*d*| and low-frequency noise to back |*g*|. For the voiceless stops |*ptk*|, however, the story is different because the transitional portion of the second formant occurs during the period of aspiration, before vocalization has begun, and is correspondingly much harder to hear. The plosive part of the voiceless stops is relatively intense, however, so that the high-frequency noise of middle |*t*| distinguishes it from the low-frequency noise of front |*p*| and back |*k*|. The distinction between |*p*| and |*k*| is slightly harder to hear because it seems to depend upon hearing the aspirated transition into the second vowel resonance. What acoustic representation there is for place of articulation of the fricative sounds is even more obscure. Probably the middle |*sz*| are distinguished from the back |*ʃʒ*| on the basis of the high-frequency energy in |*sz*|. The distinction between front |*fv*| and middle |*θð*|, however is uncertainly attributable to slight differences in the transition to the following vowel. The distinctions between |*f*| and |*θ*| and between |*v*| and |*ð*| are among the most difficult for listeners to hear and it seems likely that in most natural situations the differentiation depends more on verbal context and on visual observation of the talker's lips than it does on the acoustic difference. In any event, when we summarily assign these consonants into three classes on the basis of "articulatory position," we are thereby concealing a host of difficult problems. The positional feature is by all odds the most superficial and unsatisfactory of the five features we have employed.

TABLE XIX. Classification of consonants used to analyze confusions.

Consonant	Voicing	Nasality	Affrication	Duration	Place
p	0	0	0	0	0
t	0	0	0	0	1
k	0	0	0	0	2
f	0	0	1	0	0
θ	0	0	1	0	1
s	0	0	1	1	1
ʃ	0	0	1	1	2
b	1	0	0	0	0
d	1	0	0	0	1
g	1	0	0	0	2
v	1	0	1	0	0
ð	1	0	1	0	1
z	1	0	1	1	1
ʒ	1	0	1	1	2
m	1	1	0	0	0
n	1	1	0	0	1

In Table XIX a digital notation is used to summarize the classification of these 16 consonants on the basis of these five features. From Table XIX it is easy to see in what ways any two of the consonants differ.

Now if we apply the groupings given in Table XIX to the data matrices in Tables I–XVII, we can obtain a set of articulation scores, one score for each feature. For example, we can group the voiceless consonants together *versus* the voiced consonants and so estimate the probability that the voicing feature will be perceived correctly—the articulation score for voicing. The necessary summations for each feature for every table have been made and are given in Table XX.

A COVARIANCE MEASURE OF INTELLIGIBILITY

The recent development of a mathematical theory of communication has made considerable use of a measure

TABLE XX. Frequencies of correct responses in Tables I–XVII.

Condition	*S/N*	Band	All	Voice	Nasal	Frict	Durat	Place
1	−18	200–6500	313	2286	3200	2032	2600	1439
2	−12	200–6500	1080	3586	3742	2610	3095	1842
3	−6	200–6500	1860	3877	3921	3202	3429	2386
4	0	200–6500	2862	3977	3992	3706	3780	3099
5	6	200–6500	3336	3985	3998	3861	3910	3472
6	12	200–6500	3634	3985	3997	3916	3980	3691
7	12	200–300	1059	3725	3864	2922	2905	1717
8	12	200–400	1631	3801	3939	3402	3388	2088
9	12	200–600	1980	3903	3991	3696	3475	2341
10	12	200–1200	2287	3891	3994	3641	3526	2616
11	12	200–2500	2913	3927	3999	3778	3673	3224
12	12	200–5000	3332	3920	3999	3811	3853	3522
13	12	1000–5000	2924	3735	3861	3566	3801	3476
14	12	2000–5000	2029	3208	3573	3087	3689	2992
15	12	2500–5000	1523	2857	3472	2871	3552	2587
16	12	3000–5000	1087	2527	3283	2601	3390	2227
17	12	4500–5000	851	2283	3267	2463	3260	1927
Random guessing			250	2031	3125	2000	2500	1406

[1] Liberman, Delattre, and Cooper, Am. J. Psychol. **65**, 497–516 (1952).

[2] Liberman, Delattre, Cooper, and Gerstman, Psychol. Monographs **68**, No. 8, 1–13 (1954).

of covariance between input and output. This measure has been defined in terms of the mean logarithmic probability (MLP). If the input variable is x, which can assume the discrete values $i=1,2,\cdots,k$ with probability p_i, then the measure of the input is

$$\mathrm{MLP}(x)=E(-\log p_i)=-\sum_i p_i \log p_i.$$

If the logarithm is taken to the base 2, then the measure can be called the number of binary decisions needed on the average to specify the input, or the number of bits of information per stimulus. A similar expression holds for the output variable y, which can assume the values $j=1,2,\cdots,m$. Similarly, the number of decisions needed to specify the particular stimulus-response pair is $\mathrm{MLP}(xy)$, where p_{ij} is the probability of the joint occurrence of input i and output j. A measure of covariance of input with output is given by

$$T(x;y)=\mathrm{MLP}(x)+\mathrm{MLP}(y)-\mathrm{MLP}(xy)$$

$$=-\sum_{i,j} p_{ij} \log \frac{p_i p_j}{p_{ij}}.$$

$T(x;y)$ is often referred to as the transmission from x to y in bits per stimulus. The relative transmission is given by

$$T_{\mathrm{rel}}(x;y)=T(x;y)/H(x).$$

Since $H(x)\geq T(x;y)\geq 0$, the ratio varies from 0 to 1; if the transmission is poor and the response is not closely correlated to the stimulus, then $T_{\mathrm{rel}}(x;y)$ will be near zero, but if the response can be predicted with considerable accuracy from the stimulus, then $T_{\mathrm{rel}}(x;y)$ will be near unity.

In practice the true probabilities are not known and must be estimated from the relative frequencies obtained in a finite sample taken during the experiment. The maximum likelihood estimate of $T(x;y)$ is obtained by using n_i/n, n_j/n, and n_{ij}/n in place of p_i, p_j, and p_{ij}, respectively, where n_i is the frequency of stimulus i, n_j is the frequency of response j, and n_{ij} is the frequency of the joint occurrence of stimulus i and response j in a sample of n observations. In Tables I–XVII the cell entries are the n_{ij}, row sums give n_i, column sums give n_j, and n is 4000. Like most maximum likelihood estimates, this estimate will be biased to overestimate $T(x;y)$ for small samples; in the present case, however, the sample is large enough that the bias can safely be ignored.

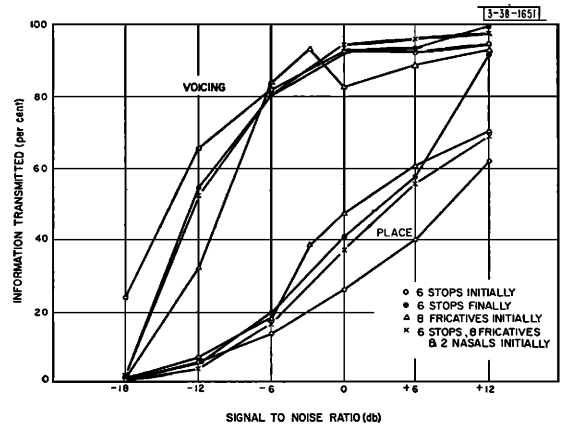

FIG. 1. The relative information transmitted about voicing (top four curves) and place (bottom four curves) is plotted as a function of signal-to-noise ratio in decibels. The four curves for each feature were obtained from four independent experiments using different test vocabularies. Voicing information is transmitted at signal-to-noise levels 18 db below those needed for place information.

The covariance measure of intelligibility can be applied to the several linguistic features separately in just the same way that the articulation score for each feature was obtained for Table XX. For example, we can construct a fourfold confusion matrix by grouping the voiceless sounds together as one stimulus and the voiced sounds as the other and then tabulating the frequency of voiceless responses to voiceless stimuli, of voiced responses to voiceless stimuli, of voiceless responses to voiced stimuli, and of voiced responses to voiced stimuli. For this 2 by 2 confusion matrix we can calculate the covariance of response with stimulus in the same way as described above and so measure the transmission of information about voicing. Similar measures can be calculated for nasality, affrication, duration, and position.

This breakdown of the confusion matrix into five smaller matrices and the measurement of transmission for each one of these five separately is equivalent to considering that we are actually testing five different communication channels simultaneously.[3] Of course, the five channels will probably not be independent. Some interaction or "cross talk" is to be expected, in the sense that knowing one feature may make some other feature easier to hear. However, the impressive thing to us was that this cross talk was so small and that the features were perceived almost independently of one another.

At first thought one might expect that if all five channels were independent, then the sum of the information transmitted by the separate channels should equal approximately the transmission calculated for all five taken together in the whole 16 by 16 matrix. This first thought would be true except for one fact; the inputs to the five channels are not independent and, therefore, even if the channels themselves are independent, the amounts transmitted through each channel will be related.

In Table XXI the average amounts of information in bits per stimulus that the listeners received are presented for the composite channel and for the five subchannels individually for all 17 conditions of masking and filtering. The last row in the table gives the amounts

[3] W. J. McGill, Psychometrika 19, 97–116 (1954).

TABLE XXI. Amounts of information transmitted in bits per stimulus in Tables I–XVII for composite channel and for each feature separately.

Condition	*S/N*	Band	All	Voice	Nasal	Frict	Durat	Place
1	−18	200–6500	0.061	0.021	0.008	0.000	0.001	0.001
2	−12	200–6500	0.959	0.516	0.264	0.069	0.087	0.058
3	−6	200–6500	1.834	0.797	0.397	0.279	0.249	0.249
4	0	200–6500	2.797	0.944	0.495	0.620	0.483	0.578
5	6	200–6500	3.226	0.951	0.543	0.782	0.636	0.856
6	12	200–6500	3.546	0.956	0.555	0.853	0.751	1.090
7	12	200–300	1.155	0.623	0.371	0.159	0.042	0.025
8	12	200–400	1.686	0.709	0.457	0.393	0.218	0.125
9	12	200–600	2.159	0.821	0.520	0.614	0.272	0.231
10	12	200–1200	2.379	0.805	0.523	0.583	0.281	0.359
11	12	200–2500	2.828	0.852	0.544	0.702	0.419	0.721
12	12	200–5000	3.185	0.847	0.521	0.730	0.581	0.936
13	12	1000–5000	2.643	0.638	0.350	0.506	0.520	0.872
14	12	2000–5000	1.582	0.273	0.160	0.229	0.426	0.499
15	12	2500–5000	1.053	0.130	0.083	0.143	0.348	0.296
16	12	3000–5000	0.624	0.048	0.023	0.067	0.235	0.143
17	12	4500–5000	0.455	0.014	0.002	0.045	0.193	0.068
Maximum possible			4.000	0.989	0.544	1.000	0.811	1.546

that would be transmitted if no mistakes at all occurred (on the assumption that all 16 syllables occurred equally often). The degree of redundancy in the input is indicated by the fact that the sum of the transmissions for the five channels is 4.890 bits, whereas the composite channel can transmit only 4 bits. This difference means that some of the input information is going through more than one channel. However, for the conditions and phonemes tested, the sum for the five channels can be used to give a rough approximation for the composite channel if the sum is corrected by the factor 4/4.89. If all of the features were transmitted equally well, this correction factor would be exact, but in most cases it is only an approximation.

The fact that the measures for the separate channels can be summed in a simple manner to give an approximate value for the total transmission is of considerable practical significance. This perceptual independence of the several features implies that all we need to know about a system is how well it transmits the necessary clues for each feature; measurements for the individual features can be made much more quickly and easily than can a measurement for the composite channel, and the correction factor for the input redundancy depends entirely on the input vocabulary and not upon an experimental test.

In the following we shall discuss the relative transmission measures. The relative measure is computed from Table XXI by dividing each entry in that table by the maximum value given at the bottom of each column. The advantage of the relative measure is that it permits an easy comparison of one channel with another. Differences in transmission due simply to the fact that the input to one channel was greater than the input to another channel are removed when we examine the relative efficiency of the two channels. We ask simply, what fraction of its input did each channel transmit? The ratio of transmitted to input information provides us with a normalized measure of stimulus-response covariation.

DISCUSSION

In Fig. 1 the normalized covariance measure—relative transmission in percent—is plotted as a function of the signal-to-noise ratio for two linguistic features, voicing and place of articulation, for the data presented in Tables I–VI. In Fig. 2 a similar plot is shown for the features of nasality, affrication, and duration. In addition to the data in Tables I–VI, the results of three smaller studies are also plotted on the same graph. In one of these smaller studies only the six stop consonants |*p*|, |*t*|, |*k*|, |*b*|, |*d*|, and |*g*|, were used initially before the vowel |*a*|. In a second study these same six stop consonants occurred finally after the phonemes |*ta*|. And in the third study only the eight fricative consonants |*f*|, |*θ*|, |*s*|, |ʃ|, |*v*|, |ð|, |*z*|, and |ʒ| were used initially before the vowel |*a*|. Both voicing and place of articulation are involved in these three smaller test vocabularies, so the relative transmission for these two features can be compared in Fig. 1 with the results obtained from the complete set of 16 consonants. Duration was also tested with fricative sounds and this function is added in Fig. 2. The comparisons show a gratifying degree of agreement from one study to the next.

The glaringly obvious statement that must be made about Figs. 1 and 2 is that voicing and nasality are much less affected by a random masking noise than are the other features. Affrication and duration, which are so similar that a single function could represent them both, are somewhat superior to place but far inferior to voicing and nasality. Voicing and nasality are discriminable at signal-to-noise ratios as poor as −12 db whereas the place of articulation is hard to distinguish at ratios less than 6 db, a difference of some 18 db in efficiency.

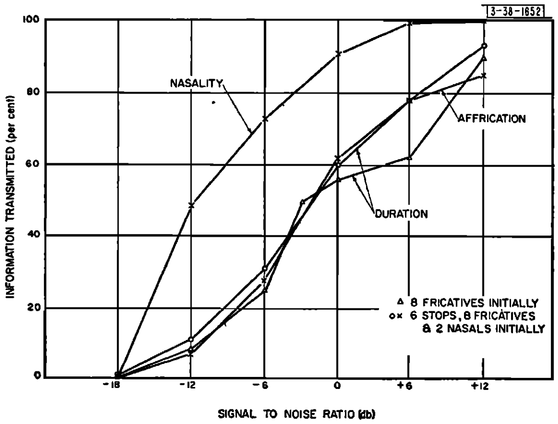

FIG. 2. The relative information transmitted about nasality, affrication, and duration is plotted as a function of signal-to-noise ratio in decibels. The two curves for duration were obtained from independent experiments using different test vocabularies. Nasality and voicing are equally discriminable.

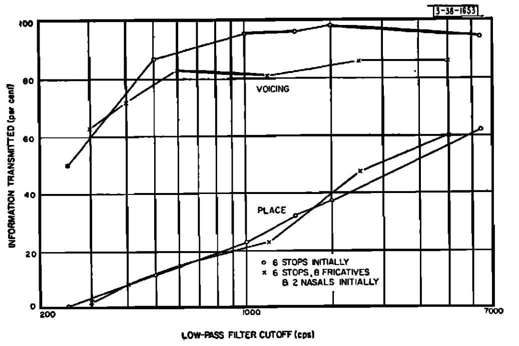

FIG. 3. The relative information transmitted about voicing and place is plotted as a function of the cutoff frequency of the low-pass filter. The two curves for each feature were obtained from independent experiments. The relation between voicing and place is the same for low-pass filtering as for masking with random noise (see Fig. 1).

In Figs. 3 and 4 similar functions are drawn for the results given in Tables VII–XII for low-pass filters. An additional small study with just the six stop consonants is also represented in Fig. 3. Figure 3 looks much like Fig. 1; voicing is greatly superior to place of articulation. Figure 4 is similar to Fig. 2, except that the results for affrication and duration are now somewhat different. These comparisons show that there is a considerable correspondence between masking by random noise and filtering by low-pass filters. This correspondence seems reasonable if we think of the high-frequency components of speech as relatively weak and therefore most susceptible to masking by the uniform spectrum of the noise. That is to say, the uniform noise spectrum should mask high frequencies more than low, so it is in effect a kind of low-pass system.

Whereas low-pass filtering and noise have much the same effect on speech perception, high-pass filtering presents a totally different picture. In Fig. 5 the relative transmissions calculated from Tables XII–XVII are plotted for all five features as a function of the filter cutoff frequency. With a minor exception for duration, all features deteriorate in about the same way as the low frequencies are removed. Duration holds up somewhat better, probably because |s|, |ʃ|, |z|, and |ʒ| are characterized in part by considerable high-frequency energy. This homogeneity reflects a fact that can be seen from visual inspection of Tables XIII–XVII; the errors do not cluster or fall into obvious patterns in the confusion matrix, but seem to distribute almost randomly over the matrix. When an error occurs with high-pass filtering, there is little chance of predicting what the error will be. Thus we find an important difference between high- and low-pass filtering; low-pass filters affect the several linguistic features differentially, leaving the phonemes audible but similar in predictable ways, whereas high-pass filters remove most of the acoustic power in the consonants, leaving them inaudible and, consequently, producing quite random confusions. Of course, this difference must be tempered by the fact that a random noise was used along with the filters, so that the noise acted "with" the low-pass filter to eliminate high frequencies but "against" the high-pass filter in such a way as to produce a narrow band-pass system. However, casual observations made since these tests were completed convince us that the difference cannot be explained entirely in this way and that, even without noise, audibility is the problem for high-pass systems and confusibility is the problem for low-pass systems.

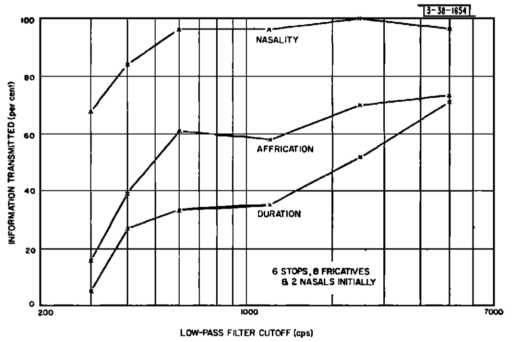

FIG. 4. The relative information transmitted about nasality, affrication, and duration is plotted as a function of the cutoff frequency of the low-pass filter. Nasality is somewhat more discriminable than voicing.

An important application of data on filtered speech has been to divide the frequency scale into segments making equal contributions to intelligibility. The high-pass and low-pass functions are plotted on the same graph and the frequency at which the two functions cross is said to divide the frequency scale into two equivalent parts; the frequencies above the crossover are exactly as important as the frequencies below the crossover frequency. We have observed this traditional method of analysis in Fig. 6 where the solid functions are the articulation scores and they are seen to cross at about 1550 cps. This frequency is somewhat lower than one would expect for female talkers, but the test vocabulary used here may not permit valid comparisons with other research.

We would like to argue that the meaning of these crossover points is apt to be a bit tricky. In the first place, the point depends crucially upon the test materials, in the sense that we can obtain very different crossover points for the different linguistic features: 450 cps for nasality, 500 cps for voicing, 750 cps for affrication, 1900 cps for place of articulation, and 2200 cps for duration. What crossover point we get depends on how we load the test vocabulary with these different features. In the second place, high- and low-pass filters do different things to speech perception, as we pointed out previously. If we plot the relative amount of information transmitted, instead of the articulation score, we obtain the dashed functions shown in Fig. 6. The crossover point for the information measure is about 1250 cps, a good 300 cps lower than for the articulation score.

By the same argument as before, there is as much information above 1250 cps as there is below. Why do these two measures give different divisions of the frequency scale? The answer lies in the fact that low-pass errors are more predictable and so carry some information, whereas high-pass errors are more random and contain no hint about what the true message might have been. Relative to the articulation scores, therefore, the high-pass information is smaller and the low-pass information is greater; the relative shifts move the crossover point downward in frequency. Which of these two crossover points is the more meaningful? Here the answer depends upon what use is to be made of the voice communication system. If isolated words, numerals, station call letters, etc. are the only messages, then a miss is as good as a mile; there is no redundancy in the message to enable the listener to correct an error, so the percentage of messages correctly received is what we want to know. On the other hand, if connected discourse in all its notorious redundancy is sent over the system, a listener can detect perceptual errors on the basis of context and can correct them more easily if they are consistent and predictable; then the transmission measure is what we want to know. However, if we arrive at a position where we must weight the frequency scale one way for isolated words and another way for conversational speech, the beautiful simplicity that makes the traditional crossover argument so attractive seems spurious. Our own intuitions would lead us to search for a different line of attack on the problem.

It may be possible to evaluate voice communication systems more adequately if we explore the implications of the multiple-channel argument used to analyze our data. It is not obvious that things will be any simpler if we must replace a single complicated channel with a dozen simpler channels in our theoretical model of speech perception. However, transmission of the separate features may be easier to relate to the system parameters. Even if a completely automatic computational procedure cannot be developed along multiple-channel lines, a short series of relatively simple articulation tests may suffice to determine the necessary parameters. In any event, the development and standardization of tests for the individual features would seem to have considerable value for the diagnosis both of inefficient equipments and of hard-of-hearing people.

One advantage of a multichannel approach to speech perception is that the message, as well as the equipment, is included in the analysis. Given any specific vocabulary of speech signals, we can calculate the relative importance of each feature for distinguishing the alternative signals and so derive a weighting factor for each channel. If the messages are coded properly into those channels or features that the system handles well, considerable advantage may be gained. For example, a low-pass system would perform best for speech signals that were distinguishable on the basis of voicing and nasality.

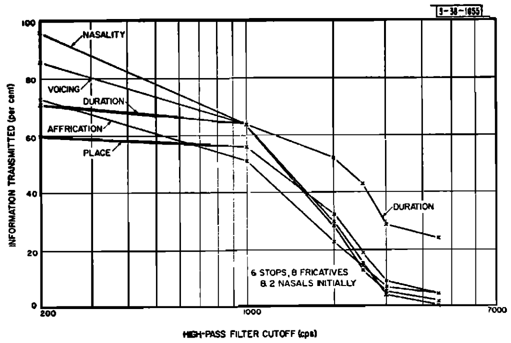

FIG. 5. The relative information transmitted about all five features is plotted as a function of the cutoff frequency of the high-pass filter. The effect of eliminating the low frequencies is the same on all features except duration.

A set of rules for developing an optimally distinguishable vocabulary for a given communication system would be rather complex and involved. There is, however, a very simple procedure for testing any given vocabulary. If the relative efficiencies of the system for the several features are known, we may know that some features will not be transmitted and cannot be used to distinguish two signals. Any two phonemes that differ only with respect to such missing features can be regarded as equivalent stimuli for the listener. Now suppose that we take any one of such a set of equivalent stimuli and use it wherever any of the set occurs; for example, if $|p|$, $|t|$, and $|k|$ are indistinguishable, we might use $|t|$ for all three. When all the speech signals are rewritten with $|t|$ wherever $|p|$, $|t|$, or $|k|$ occurred before and similar substitutions are made for all other sets of equivalent stimuli, the rewritten signals will approximate what the listener will hear. If we now alphabetize the rewritten signals, we will probably find some that are identical. These are the signals that will be confused and we can then take steps to eliminate such confusions.

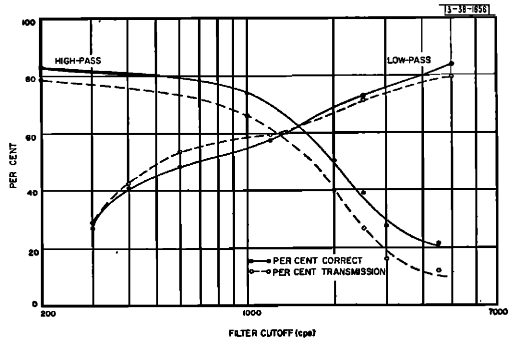

FIG. 6. Both the articulation score and relative information transmitted are plotted as a function of the frequency cutoff for both high-pass and low-pass filters. The crossover points are different for the two measures.

For example, if we look at Figs. 3 and 4 to see what happens when frequencies above 1000 cps are filtered out of the speech, we find that the features of place and duration are effectively absent and that voicing, nasality, and affrication are doing all the work. In other words, the filter has effectively deleted the last two columns in Table XIX. With those two columns gone there are really just five distinguishable phonemes left: |*ptk*|, |*fθsʃ*|, |*bdg*|, |*vðzʒ*|, and |*mn*|. Replace these by, say, |*t*|, |*s*|, |*d*|, |*z*|, and |*n*|, respectively. Now when we rewrite the vocabulary of speech signals with just these five consonants instead of the original 16, we will discover which signals are transformed into indistinguishable forms by the filter. Insofar as possible, no two signals should be the same in their rewritten versions. The basic idea behind this procedure is that redundancy in the input signals will be most effective in reducing errors if we insure that frequent confusions do not transform one permissable signal into another permissable signal.

We have explored the validity of this substitution scheme for just those conditions described in the preceding example. Sentences and longer texts were rewritten with the indicated substitution of five for 16 phonemes. Such rewritten passages are appropriately called "elliptic" English, the ellipsis referring to the omission of two features, place and duration. With a little practice it was possible to speak the elliptic passages at normal rates and with normal intonation. Over a high quality communication system the elliptic speech was intelligible but sounded a little as though the talker had a marked dialect or speech defect. Then the low-pass filters were introduced. When all the frequencies above 1000 cps were removed (the conditions for which the substitutions were designed), the ellipsis could no longer be detected. Elliptic speech sounded just the same as normal speech under these conditions of distortion. A similar result was obtained with a masking noise at signal-to-noise ratios of about 0 db. The illusion is quite compelling and this demonstration that we could duplicate the effects of noise or distortion by deleting certain features of the speech increased our confidence in a multichannel model of speech perception.

An interesting sidelight on elliptic speech is provided by the art of ventriloquism. A ventriloquist talks without moving his lips. The consonants |*p*|, |*f*|, |*b*|, |*v*|, |*m*|, and |*w*| are normally produced with lip movements and so pose a problem. A variety of solutions are possible; these sounds are avoided or omitted or produced out of the side of the mouth, or made in alternative ways (especially |*f*| and |*v*|). In most of the older books on ventriloquism, however, a system of substitutions is proposed; |*k*| for |*p*|, |*g*| for |*b*|, and |*n*| for |*m*| are common suggestions. These substitutions should be especially satisfactory for the "voice in a box" trick, where the high frequencies should be attenuated in passing through the walls of the box and the confusion of sounds would be expected to occur naturally.

The place of articulation, which was hardest to hear correctly in our tests, is the easiest of the features to see on a talker's lips. The other features are hard to see but easy to hear. Lip reading, therefore, is a valuable skill for listeners who are partially deafened because it provides just the information that the noise or deafness removes.