XIX. SPEECH COMMUNICATION[*]

Academic and Research Staff

Prof. K. N. Stevens
Prof. M. Halle
Prof. W. L. Henke
Prof. A. V. Oppenheim
Dr. S. Berkowitz
Dr. Margaret Bullowa
Dr. J. L. Fidelholtz
Dr. A. W. F. Huggins
Dr. A. R. Kessler
Dr. Emily F. Kirstein
Dr. D. H. Klatt
Dr. Paula Menyuk
Dr. J. S. Perkell
Dr. R. A. Stefanski
Dr. Jacqueline Vaissiere
Mary M. Klatt
R. M. Mersereau
Estill Putney

Graduate Students

T. Baer
D. S. K. Chan
R. E. Crochiere
D. E. Dudgeon
Ursula Goldstein
S. Maeda
B. Mezrich
H. Pines
M. R. Portnoff
L. Tung
V. W. Zue

RESEARCH OBJECTIVES AND SUMMARY OF RESEARCH

The broad aims of our research in speech communication are to gain an understanding of the nature of the processes of human speech production and perception and to learn how these processes are acquired. A practical aim is to utilize knowledge gained through study of speech communication to devise procedures that will permit limited communication between men and machines by means of speech. We are also concerned with the theory and practice of modern digital signal-processing techniques.

Our studies of the mechanism of speech production include examination of the acoustic correlates of various places of articulation for different consonant classes, and further investigations of the mechanisms of vocal-cord vibration, through measurements on excised larynges. We are continuing our work on the measurement of certain temporal and other periodic aspects of speech, and on organizing the timing data into a model that predicts segment durations in sentence material. We are examining the role of timing in speech perception through a series of experiments in which the temporal characteristics of natural sentence material are systematically manipulated, and the influence of these changes on various kinds of listener judgments are assessed.

Two projects are aimed at modeling certain properties of the speech-production process. One is the continued development of a scheme for speech synthesis by rule utilizing a terminal-analog speech synthesizer, and the other is the reactivation of earlier work on modeling of the tongue and its motions during speech production.

Research on language acquisition is concerned with two aspects of the speech-development process. In one study, we examine the acoustic properties of consonants and consonant clusters of children in a period when the child is beginning to acquire phonological rules that specify constraints on these clusters in English. The other study is investigating the properties of the sounds produced in the first few months of life. We are attempting to classify these sounds, and relate these acoustic categories to the behavior and environment of the child.

In addition to these basic studies of speech production and perception there are ongoing projects concerned with the practical objectives of speech recognition and

[*] This work is supported by the U. S. Navy – Office of Naval Research (Contracts N00014-67-A-0204-0064 and N00014-67-A-0204-0069), the National Science Foundation (Grant GK-31353), the National Institutes of Health (Grant 5 RO1 NS04332-10), and the Joint Services Electronics Programs (U. S. Army, U. S. Navy, U. S. Air Force) under Contract DAAB07-71-C-0300.

speaker recognition. The work on speech recognition, part of which is being done in collaboration with, and using the facilities of, the M.I.T. Lincoln Laboratory, is implementing certain new techniques for digital processing of speech signals and is currently applying these techniques to the study of the acoustic characteristics of stop and fricative consonants. We hope to develop algorithms for partial recognition of these segments in sentence contexts. The speaker-recognition studies examine the various sources of inter- and intra-speaker variability in speech sounds. Recent experimental work has isolated a number of acoustic attributes that appear to characterize the voice of an individual.

Digital signal-processing techniques are being used in the design of digital filters that can be used for speech research. Included in this research is a comparison of different techniques for implementing digital filters with respect to the effects of coefficient quantization and multiplication roundoff errors. We are also studying some of the mathematical properties of linear prediction techniques for speech analysis.

Work is progressing on the design and implementation of a general-purpose online (digital) signal analysis language and system to complement the signal synthesis capabilities of the MITSYN system. We hope to be able to analyze and correlate signals from various types and locations of transducers during speech (e.g., accelerometers located in different positions) and to analyze the output of models of speech production. Hardware has been added to our computer system to allow simultaneous two-channel signal inputs, i.e., analog-to-digital conversion. The analysis system is based upon processing primitives or operations such as discrete Fourier transforms, windowing, and block arithmetic operations such as complex conjugate multiplication. These operators can be executed one at a time from a keyboard, or can be assembled into programs for functions such as power spectrum densities, crosscorrelation functions, histogram collection, cepstrum analysis, and cepstrally smoothed spectra. A set of displays is included for showing various representations of time signals and transformed correlates.

K. N. Stevens, M. Halle, W. L. Henke,
A. V. Oppenheim, D. H. Klatt

A. FURTHER THEORETICAL AND EXPERIMENTAL BASES FOR QUANTAL PLACES OF ARTICULATION FOR CONSONANTS

NIH (Grant 5 RO1 NS04332-10)

K. N. Stevens

In previous studies,[1-3] we have examined the acoustic and articulatory correlates of place of articulation for some of the consonants. As the place of maximum vocal-tract constriction is displaced to different positions from the glottis to the lips, the acoustic properties of the sound output in simple consonant-vowel syllables change in a discontinuous fashion. The places of articulation that are used to produce consonants in various languages appear to correspond to the places where the properties of the sound are relatively insensitive to small perturbations in the location of the constriction.

We have shown that this general quantal principle applies for certain broad classes of place of articulation for consonants, particularly the pharyngeal, uvular, and velar consonants.[2] It has not been examined in detail for consonants produced by forming a constriction with the blade of the tongue – the so-called coronal consonants.[4] A number of stop, nasal, and fricative consonants in various languages are produced with a constriction in this region.

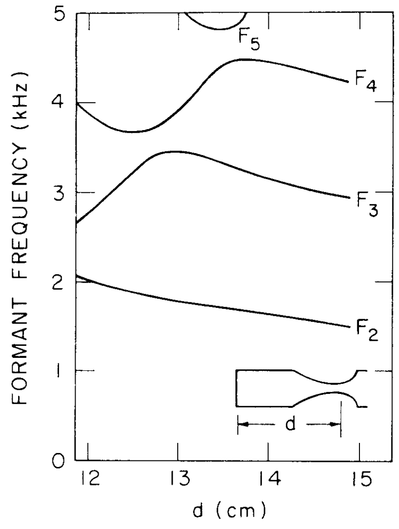

Fig. XIX-1.

Natural frequencies of a constricted tube open at one end and closed at the other, as a function of constriction position. The approximate shape of the tube is shown. The length of the tube is 17 cm, the maximum cross-sectional area is 4 cm^2, and the minimum cross-sectional area is 0.1 cm^2. Curves are labeled with formant number. The first formant (low-frequency) is not shown.

Figure XIX-1 shows the computed natural frequencies of an acoustic tube closed at one end (corresponding to the glottis) and constricted at a point near the open end, at a region where the tongue-tip or tongue-blade constriction is made.[5] The abscissa represents the distance from the glottis to the point of maximum constriction. The curves are labeled with formant number. Formant 1 is very low and is not shown. The portions of the curves that slope up to the right correspond to front-cavity resonances (they increase with decreasing front-cavity length) and those sloping down to the right correspond to back-cavity resonances. Over the region d = 12.5-14 cm, which represents the various configurations for coronal consonants, F2 is in the range 1600-1900 Hz, and F3 at 3000 Hz or above. These ranges are known to be appropriate formant loci for all varieties of coronal consonant production.

Consider the properties of the sound that is generated when a fricative consonant is produced and the position of the constriction formed by the tongue tip is moved gradually from a posterior to an anterior position through d = 12.5 to d = 14 cm. Of especial interest are the two places where the curves have maxima and minima, one around d = 12.8 cm, and the other around d = 13.5 cm. Around one of these regions, F3 and F4 are close together and do not change much with constriction position, and around the other, F4 and F5 are close together and are relatively fixed in position. Near each of these points, neither of the proximate natural frequencies can be assigned exclusively to the back or the front cavity, and it can be shown that both resonances are excited by a turbulence noise source located anterior to the constriction. In between these two regions there is a narrow range where F4 rises rapidly with d, and represents a front-cavity resonance, which is excited by the turbulence noise source.

Figure XIX-2 is a spectrogram of the sound output when the tongue is moved with a continuous gesture in a posterior to anterior direction. We observe the region where F3 and F4 are close together, and the rapid shift to the adjacent region where F4 and F5 are close together. The quantal nature of the relation between sound and articulation is quite evident. The more posterior of the two fricatives is the retroflex [ṣ] that occurs in some languages (but not English), whereas the more anterior version is the [s].

The contrast between these two places of articulation can also be observed in the noise burst at the release of stop consonants, and in the formant transitions for both stop and nasal consonants. A particular attribute of the more posterior stops and nasals is the proximity of F3 and F4 at the release of consonants.

For both anterior and posterior points of articulation the consonant can be produced with the tongue tip exclusively, as in the case of the consonants in Fig. XIX-2, or the blade can be actively involved, as shown in Fig. XIX-3. The upper pictures are sketches of the tongue position for the more anterior consonants with the constriction near the dentialveolar junction, and the lower pictures represent the consonants produced with a more posterior alveolar point of constriction. These contrasting apical (to the left) and laminal (to the right) manners of articulation are well known to phoneticians. The laminal version gives rise to a gradually widening vocal-tract cross-sectional area behind the constriction, while the widening is more abrupt for the apical version.

There appear to be two distinct acoustic consequences of the laminal articulation in

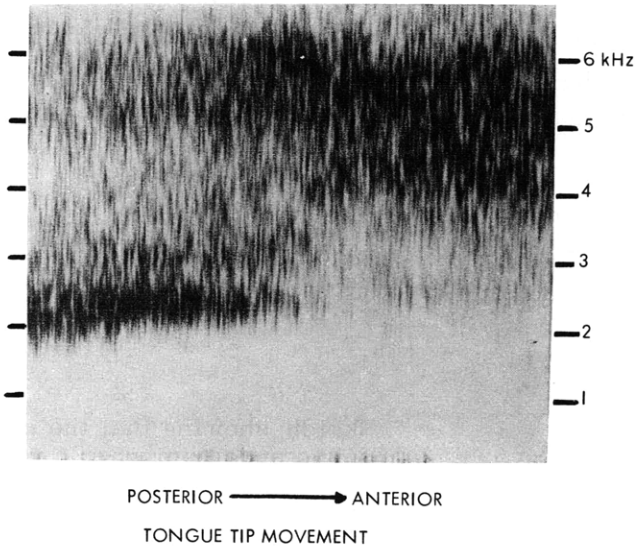

Fig. XIX-2. Spectrogram of the fricative sound produced when the tongue tip is moved slowly from a post-alveolar to a dental position.

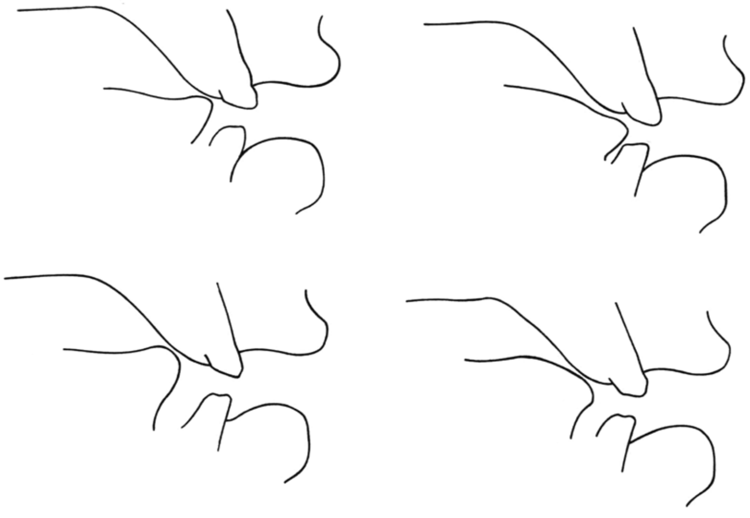

Fig. XIX-3. Sketches of midsaggital sections showing the approximate tongue position for various types of coronal consonants. The two upper sketches show a more anterior constriction position, and the two lower a more posterior constriction position. Apical articulations are to the left, laminal to the right. These sketches were prepared in part from x-ray data and in part from observation and published descriptions of these coronal articulations.

contrast to the apical articulation. One of these consequences, which is of particular importance for fricative consonants, is that the gradual widening or tapering of the vocal tract behind the constriction results in greater coupling of the turbulence noise source near the constriction to the back-cavity resonances. The situation is illustrated schematically in Fig. XIX-4, which shows some high-frequency spectral peaks in the sound output associated with natural frequencies of the part of the vocal tract posterior to the constriction when this part is tapered. Thus for laminal articulations, several resonances are excited over a range of frequencies, above a certain critical frequency determined by the amount of tapering. The English [$\check{s}$] sound is the posterior version of a laminal fricative. The dental [θ] is an anterior version.

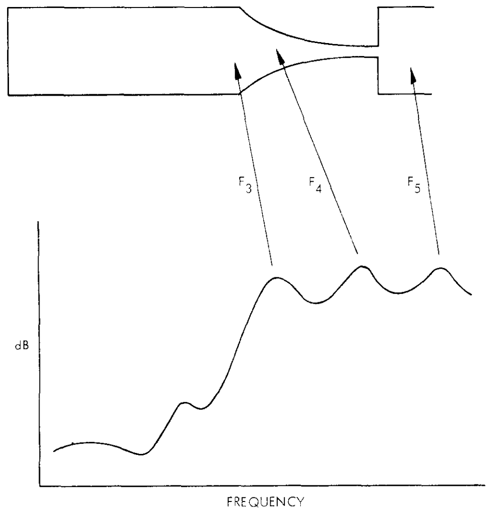

Fig. XIX-4.
Sketch showing that the peaks in the output sound spectrum when a constricted tube is excited by turbulence noise can arise from back-cavity resonances when the tube is tapered.

A second acoustic consequence of the laminal articulation, which plays a role for stops and nasals, apparently arises because the tongue blade, which forms a longer constriction, is released less rapidly than the tongue tip. In the case of a (nonnasal) stop consonant, this means that the burst of frication noise at the release of such a consonant is longer and the voicing onset is delayed. Spectrograms of contrasting laminal and apical stops are shown in Fig. XIX-5. The voice-onset time, i.e., time from stop release to onset of voicing, is just a few milliseconds for the apical version (to the right), and is 20 ms or more for the laminal consonant. There are apparently several languages that have this contrast, and spectrograms that show this acoustic difference for one such language (Isoko) have been published by Ladefoged.[6] We have observed the same difference for the Tamil language. This contrast in voice-onset time appears to exploit a fundamental dichotomy in the auditory response to a sequence of two nonspeech signals with differing properties – in this case a noise burst followed by buzz onset. If the time between onset of the two signals is less than 15-20 ms, psychoacoustic data

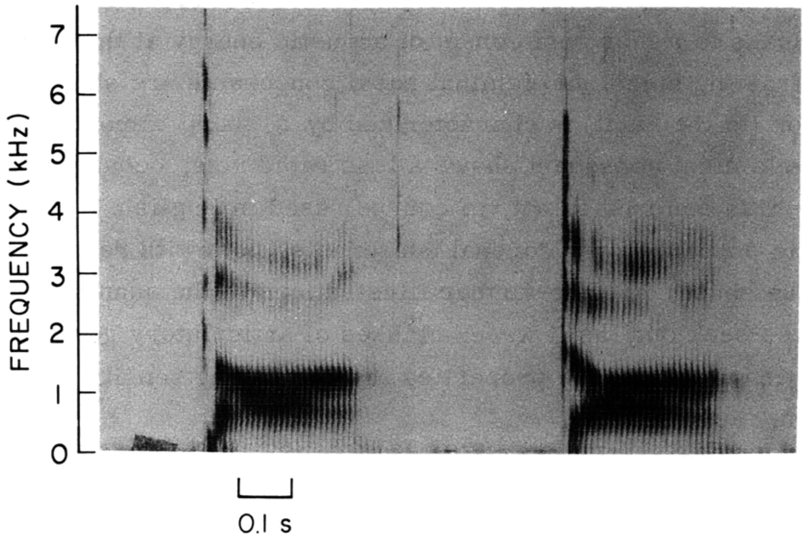

Fig. XIX-5. Spectrograms of the syllables t̪a (left) and ta (right), the former representing the laminal stop.

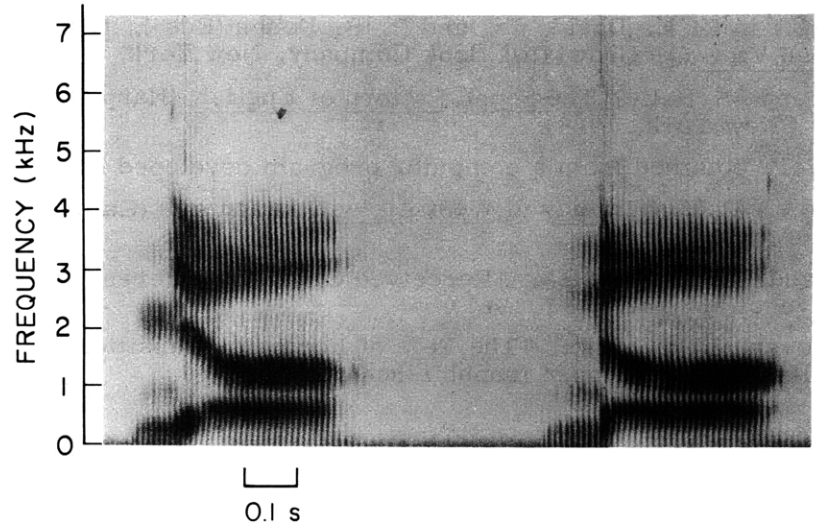

Fig. XIX-6. Spectrograms of the syllables n̪a (left) and na (right), the former representing the laminal nasal consonant.

show that the onsets are judged to be simultaneous; if this time is greater than 20 ms or more, the signals are perceived to be successive.[7, 8]

For nasal consonants, the lower rate of increase in tongue-constriction size for the laminal version leads to a less rapid onset of acoustic energy at the consonantal release. Examples of contrasting apical and laminal nasal consonants are shown in Fig. XIX-6. The apical version (to the right) is characterized by a sharp almost transient rise in energy, while the laminal consonant shows a less rapid rise, occurring over a time of 20 ms or more. This contrast is not, of course, used in English.

These acoustic properties for coronal consonants, both with regard to constriction position and tongue shape, provide further illustrations of the quantal nature of speech events. Languages seek out, as it were, classes of articulatory gestures that lead to distinctive acoustic outputs whose properties are minimally sensitive to perturbations in the articulation.

References

1. K. N. Stevens, "Acoustic Correlates of Place of Articulation for Stop and Fricative Consonants," Quarterly Progress Report No. 89, Research Laboratory of Electronics, M. I. T., April 15, 1968, pp. 199-205.

2. D. H. Klatt and K. N. Stevens, "Pharyngeal Consonants," Quarterly Progress Report No. 93, Research Laboratory of Electronics, M. I. T., April 15, 1969, pp. 207-216.

3. K. N. Stevens, "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data," in E. E. David, Jr. and P. B. Denes (Eds.), Human Communication, A Unified View (McGraw-Hill Book Company, New York, 1972).

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

5. These data were obtained from a computer program developed by W. Henke.

6. P. Ladefoged, A Phonetic Study of West African Languages (Cambridge University Press, London, 1968).

7. I. J. Hirsh and C. E. Sherrick, "Perceived Order in Different Sense Modalities," J. Exptl. Psychol. 62, 423-432 (1961).

8. K. N. Stevens and D. H. Klatt, "The Role of Formant Transitions in the Voiced-Voiceless Distinction for Stops" (unpublished).

B. DURATIONAL CHARACTERISTICS OF PRESTRESSED WORD-INITIAL CONSONANT CLUSTERS IN ENGLISH

NIH (Grant 5 RO1 NS04332-10); ONR (Contract N00014-67-A-0204-0069)

D. H. Klatt

This report describes preliminary results of a spectrographic study of prestressed word-initial consonant clusters. Only information concerning segmental durations will be considered at this time.

There is surprisingly little published information about segmental durations and other acoustic characteristics of consonant clusters in American English,[1-3] although Haggard[4] has studied the durational patterns of selected initial and final consonant clusters in British English. Studies of durational reorganization in clusters are of interest in applied fields where attention to such details may lead to improved speech synthesis by rule and better automatic speech recognition strategies. A less obvious but important reason for interest in consonant clusters is the possibility that durational recoding rules can be used to deduce sequential patterns of articulatory control and other basic information concerning articulatory dynamics during speech production.

1. Experiment

A list of monosyllabic words was constructed to include 5 examples for each of 25 different word-initial clusters. In order to establish a standard or basic duration for each consonant, words beginning with a single consonant were also recorded.

Four monosyllabic words involving the vowels [i ɛ ay u] were selected for each consonant and cluster, and a fifth word was generated by adding a second syllable to the end of one of 4 monosyllabic words. For example, the [str] words were "street, stress, strike, strewn, and stressful."

The word list was randomized and recorded at a moderate speaking rate in an anechoic chamber by 3 adult male speakers. All words were spoken in the frame sentence "Say x instead" in order to produce speaking rates more nearly in line with conversational speech, and to avoid effects of prepausal lengthening in utterance-final position. The frame sentence also permitted the measurement of a silent closure duration for word-initial plosives.

Spectrograms were made of all the phrases. Figure XIX-7 illustrates the measurement technique. The segmental durations were defined with respect to certain acoustic landmarks in the spectrogram.[5] The duration of a plosive included the silent closure interval but not the burst of frication at release. Burst and aspiration were included in the duration of the following phonetic segment. A sonorant-vowel boundary was defined as the time when the second formant passed through a frequency halfway between estimated initial and final frequency values for the transition. If the second formant did not change significantly in frequency, the third formant was used.

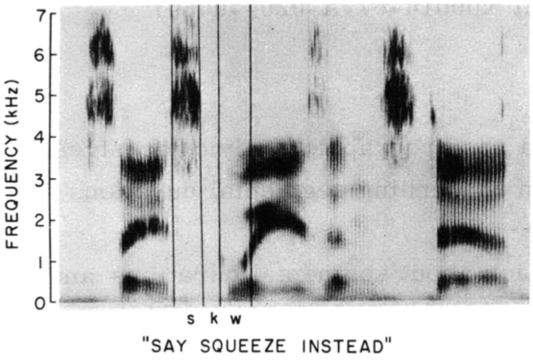

Fig. XIX-7.

Broadband spectrogram used to determine segmental durations.

2. Results

Table XIX-1 gives some indication of speaker variability and/or measurement error in the durational data reported. The word "seat" was repeated at 15 randomized positions in the word list. The standard deviations for the durations of [s] and [i] are presented for each speaker. A standard deviation of 10 ms, or 7%, is probably representative of the data that will be described. The measurement error may in fact be somewhat greater for sonorants because of the added difficulty in determining segmental boundaries.

Table XIX-1. Mean duration and standard deviation of [s] and [i] in the word "seat."

		Duration	
	Speaker	μ (ms)	σ (ms)
[s]	RK	172	10.3
	KNS	169	11.4
	DHK	133	6.0
[i]	RK	116	7.1
	KNS	117	9.0
	DHK	112	7.0

Table XIX-2 presents the average duration of each single consonant obtained from 5 words uttered by each of 3 speakers. These data indicate that labial consonants tend to be longer than dentals, and dentals tend to be longer than velars. The differences

Table XIX-2. Basic durations of the cluster elements as determined from measurements on 5 words that begin with the single consonant.

f	138 ms	s	152		
p	100	t	85	k	78
b	97	d	85	g	78
m	105	n	92		
w	107	l	92		
		r	102		

in closure duration for labial, dental, and velar stops would be less if the frication burst duration were included as a part of the plosive.[6] Also, fricative consonants are longer than stops and sonorants in English. Voiced and voiceless plosives are about of equal length in these data. Presumably, voiced fricatives, which were not included in this study, are shorter in duration than voiceless fricatives.[7]

The cluster data were first examined separately for each of the 4 vowels [i ε ay u]. No statistically significant modification to average consonant duration was found. Thus large differences in vowel duration and differences in vowel features had no average effect on consonant duration.

The addition of a second syllable to the end of a word shortened the stressed vowel 26% and the preceding consonant of the cluster 10%. Earlier consonants in the cluster were nearly unchanged in duration in going from a one-syllable to a two-syllable word.

Differences between individual speakers were examined and we found that the duration data for consonants in singleton and cluster environments were on the average longest for speaker KNS, 4% shorter for RK, and 25% shorter for DHK. Aside from these average differences, no large systematic differences among speakers were observed.

We assumed the absence of any vowel-conditioned or speaker-conditioned effects, and pooled the data across the 5 words containing a given cluster and across the 3 speakers. The leftmost columns of Table XIX-3 present the measured durations for all consonants and consonant clusters studied.

A set of rules describing the way in which the consonantal durations listed in Table XIX-3 change when a consonant is placed in a cluster is offered in Table XIX-4. These rules were derived subjectively and then optimized to minimize the unexplained variance in all of the cluster data. This process was iterated until no new rules of a general nature could be formulated from observing the rightmost columns of Table XIX-3 which indicate the differences between measured and predicted cluster durations.

An example should help to clarify the mechanics of applying the rules presented in Table XIX-4. Consider the [l] in the cluster [spl]. The [l] has a basic duration of 92 ms (see Table XIX-2). General consonantal shortening (rule 1) will shorten this

Table XIX-3. Measured average segmental durations in all consonant clusters that were studied are presented in the leftmost columns. Segmental durations predicted by the rules are also tabulated and the difference between the measured and predicted times in ms are presented in the rightmost columns.

	Duration in ms								
Cluster	Measured			Predicted			Difference		
sp	102	83		99	84		3	-1	
st	110	53		109	66		1	-13	
sk	105	58		109	61		-4	-3	
sm	112	85		111	88		1	-3	
sn	123	70		121	72		2	-2	
sw	125	93		124	90		1	3	
sl	133	67		133	72		0	-5	
fl	118	53		129	66		-11	-13	
pl	100	83		94	92		6	-9	
bl	88	73		91	66		-3	7	
pr	90	97		94	102		-4	-5	
br	95	65		91	74		4	-9	
tw	70	135		70	120		0	15	
tr	62	125		64	121		-2	4	
dr	63	93		64	93		-1	0	
kw	70	117		64	120		6	-3	
kl	75	95		69	98		6	-3	
gl	68	85		69	72		-1	13	
kr	67	108		69	108		-2	0	
gr	70	98		69	80		1	18	
spl	93	88	57	96	81	58	-3	7	-1
spr	97	93	67	96	81	66	1	12	1
str	102	50	80	105	53	85	-3	-3	-5
skw	107	55	87	105	54	81	2	1	6
skr	107	55	72	105	57	72	2	-2	0

Table XIX-4. Rules for predicting the change in basic duration of a consonant in a prestressed word-initial cluster. Total percentage change is obtained by adding the contributions of applicable rules.

Rule	Example
1. General Consonantal Shortening	
In 2-element clusters, $C_1 = -12\%$ and $C_2 = -22\%$	–
In 3-element clusters, $C_1 = -15\%$, $C_2 = -25\%$ and $C_3 = -30\%$	–
2. Sonorant Lengthening if Partially Voiceless	
If C is preceded by a voiceless aspirated stop, C = +28%	k$\underline{r}$
3. Ballistic Shortening	
If C precedes a stop, C = -8%	$\underline{s}$n
If C precedes a voiceless stop, C = -8%	$\underline{s}$k
4. Incompressibility of Labials	
If C is a labial, C = +6%	s$\underline{m}$
If C is adjacent to a labial, C = -6%	$\underline{s}$m
5. Retroflection Following Dental Stops	
If [r] follows a dental stop, [r] = +13%	t$\underline{r}$
If [r] follows a dental stop, dental stop = -13%	$\underline{t}$r

duration 30%. Rule 2 does not apply, since the [p] is not aspirated in [spl]. Ballistic shortening (rule 3) applies only to consonants in front of stops and not to the [l] following the stop [p]. The [l] duration is shortened 6% because rule 4 involves the incompressibility of the adjacent labial [p]. Rule 5 does not apply. Thus the basic duration is shortened 30% and 6%, which gives a total of 36% shorter than 92 ms.

3. Discussion

Consonants are generally shorter in 2- and 3-element clusters than in a simple CV syllable, but one exception to this tendency will be noted. The first rule in Table XIX-4 shortens the first consonant in a 2-element cluster 12%, and the second consonant 22%. Consonants in a 3-element cluster are shortened even more. Early consonants seem to be shortened less than later ones. The addition of another syllable to the end of a word shortened the stressed vowel 26% and the preceding consonant 10%.

The reasons for these and other shortening tendencies in speech have not been clearly established. We are preparing for publication data that suggest that the shortening produced by adding a syllable can best be explained by the fact that the fundamental frequency contour indicating primary stress[8] can then be realized over the entire

2-syllable word if a second syllable is present. Cluster shortening may be attributed to the tendency of speakers to initiate stressed syllables at regular rhythmic intervals independent of the segmental composition of a syllable, but, in view of other evidence, we prefer to give a related explanation based on the notion of information transfer per unit of time: Prestressed consonant clusters represent little more information than does a single consonant because the elements of a cluster sequence are highly constrained.[9] It is then not surprising that speech-production rules have evolved to produce these clusters more rapidly and perhaps with greater coarticulation.

a. Aspiration and Transition Delay

A phonological rule of English states that the voiceless plosives [p t k] are strongly aspirated in prestressed position unless preceded by an [s] in the same word. Thus strong aspiration is present in the words "top" and "tried," but not in "stop" and "stride." The second duration rule in Table XIX-4 indicates that the presence of this aspiration is accompanied by a 28% increase in the duration of a sonorant. The duration of the sonorant is actually longer in words such as "tried" than its basic duration in words such as "ride." Haggard[4] noted this same tendency in British English.

It may be that the aspiration should more properly be assigned to the duration of the plosive and not to the following segment. Peterson and Lehiste[5] found, however, in plosive vowel sequences, that the vowel was only slightly longer if preceded by an aspirated plosive. A more likely explanation is that the formant motions for the sonorant-vowel transition are delayed so that they take place during voicing rather than in the presence of aspiration.

b. Ballistic Closure Effects

During the production of a stop consonant, an articulator may be presumed to make a ballistic closing gesture. Rule 3 in Table XIX-4 indicates that an [s] preceding a stop is shortened 10%. It appears that a rapid ballistic motion impinges on the duration of the preceding phonetic segment. This effect may be explained by assuming that either the command onset for the following stop is initiated somewhat earlier than for other following segments or the articulatory command force is greater in a ballistic gesture, and thus the closing motion is completed earlier.

c. Incompressibility of Labials

Rule 4 indicates that labial consonants cannot be shortened as much as nonlabials, and that this relative incompressibility of labials is accomplished at the expense of adjacent nonlabial consonants. The difference in shortening between dental and velar consonants is not statistically significant. If this rule is attributable to productive rather than to perceptual constraints, it implies that the lips are somewhat more sluggish when subjected to time pressure than are other articulators.[10]

d. Homorganic Incompatible Clusters

The dental consonants [s t d n l] share approximately the same place of articulation. Another consonant that involves the tongue tip is the retroflex consonant [r]. The fifth rule states that a cluster involving a dental consonant followed by [r] is restructured so that [r] is 13% longer and the dental is 13% shorter. Haggard[4] found a similar tendency in his data and proposed the distinction that, although [r]-dental clusters are homorganic, they involve incompatible articulatory gestures.

4. Conclusion

Since we have examined data from only 3 speakers in this investigation, a serious question remains about how general these rules are. Haggard[4] found considerable variation in the data from his 8 speakers of British English. Variability in his cluster-duration measurements was attributable at least in part, however, to the small number of examples elicited from the speakers for each cluster type. The striking feature of the Haggard study is that all of our rules appear as general tendencies in his data too. Coupled with the fact that the rules presented in Table XIX-4 express highly significant deviations from a random distribution, speaker consistency in two dialects of English suggests that the rules express regularities that should be found in data from most speakers of general American English.

It is clear that these data contribute to a better understanding of the physiological constraints and timing controls that should be incorporated in articulatory models of the speech-production process. One might question, however, the perceptual importance of durational recoding rules for clusters. No matter how statistically significant the average data may be, these durational differences are small and may easily be blunted by speaker variability in the production of individual utterances.

Perceptual experiments have been performed which indicate that the larger durational differences induced by the rules are greater than the just-noticeable difference for consonant duration.[11] The larger differences are also considerably greater than the single production standard deviation observed in this experimental paradigm. Thus it appears that duration adjustment rules for clusters would be of value if incorporated in systems for the synthesis of speech by rule or in decoding algorithms for the automatic recognition of speech by machine.

References

1. I. Lehiste, Suprasegmentals (The M. I. T. Press, Cambridge, Mass., 1970).
2. M. F. Schwartz, "Duration of /s/ in /s/-Plosive Blends," J. Acoust. Soc. Am. 47, 1143 (1970).

3. D. H. Klatt, "On Predicting the Duration of the Phonetic Segment [s]," Quarterly Progress Report No. 103, Research Laboratory of Electronics, M. I. T., October 15, 1971, pp. 111-126.

4. M. P. Haggard, "Effects of Clusters on Segment Durations, Speech Synthesis and Perception 5," Psychological Laboratory, Cambridge, England, 1970, pp. 1-50.

5. G. E. Peterson and I. Lehiste, "Duration of Syllabic Nuclei in English," J. Acoust. Soc. Am. 32, 693-703 (1960).

6. J. Kim and P. F. MacNeilage, "Voiceless Consonant Durations in VCV Utterances," Paper presented at the 84th Meeting of the Acoustical Society of America, Miami Beach, Florida, November 1972.

7. C. E. Parmenter and S. N. Trevino, "The Length of the Sounds of a Middle Westerner," Amer. Speech 10, 129-133 (1935).

8. S. Cushing, "English as a Tone Language, The Acoustics of Primary Stress," Quarterly Progress Report No. 92, Research Laboratory of Electronics, M. I. T., January 15, 1969, pp. 351-359.

9. T. R. Hoffmann, "Initial Clusters in English," Quarterly Progress Report No. 84, Research Laboratory of Electronics, M. I. T., January 15, 1967, pp. 263-274.

10. C. V. Hudgins and R. H. Stetson, "Relative Speed of Articulatory Movements," Arch. Neerl. de Phonet. Exper. 13, 85-95 (1937).

11. A. W. F. Huggins, "Just-Noticeable Differences for Segment Duration in Natural Speech," J. Acoust. Soc. Am. 51, 1270 (1972).

C. SPEAKER RECOGNITION AND VERIFICATION USING LINEAR PREDICTION ANALYSIS

ONR (Contract N00014-67-A-0204-0069)

M. R. Sambur

[M. R. Sambur was supported by a Bell Telephone Laboratories Fellowship]

1. Introduction

The selection of features that efficiently characterize a pattern is an important aspect of the problem of pattern recognition. This report summarizes a doctoral study[1] undertaken to determine a set of acoustic features that can be conveniently and automatically extracted from the speech of an individual and are effective for the identification of the speaker.

The investigation was conducted by first determining an initial set of acoustic parameters which, on the basis of theoretical considerations and past experimental work,[2-4] might be suitable candidates for indicating the unique properties of a speaker's vocal apparatus, as well as some aspects of his learned pattern of speaking. The initial selection of features was also made to take advantage of the speech-analysis technique of linear prediction.[5] The list of proposed speaker-characterizing features was then subjected to more detailed scrutiny by an experimental study of repeated utterances produced by several speakers. A probability-of-error criterion was devised to evaluate the relative merits of the features.

2. Experimental Data

To assess the extent of parameter variation over time, the data were specifically collected during 5 different recording sessions. In the first session there were 22 speakers, each of whom repeated a prescribed set of sentences 10 times. The sentences were (i) "Cool shirts please me," (ii) "Pay the man first, please," (iii) "I cannot remember it," (iv) "Papa needs two singers," (v) "Cash this bond, please." Eleven members of the original group returned 3 1/2 years later to provide replications of the experimental data. Between these two recording sessions, 4 members of the group returned on 3 separate occasions to provide further replications.

3. Probability-of-Error Criterion

The obvious goal of a speaker-recognition system is to classify an unknown speaker correctly. Thus the relative merit of a feature should be based upon its contribution to the performance of recognition. In practical terms, if a group of features, G, yields a smaller rate of error than another group of features, then the set G is a better set of features for recognizing speakers. The probability-of-error criterion can be viewed as

3. Measurements Investigated

a. Vowels

Because of their relationship to the shape of a particular speaker's vocal tract, it was argued that the measured formant frequencies and bandwidths of a given vowel spectrum would be important clues to the identity of an unknown speaker. The locations of the real-axis poles and extraneous wideband poles in the vowel spectrum also appeared to be potentially promising speaker-recognition features. This was due to the assumed relationship of these parameters with the shape of the speaker's glottal source spectrum. The recognition potential of these features was examined in the vowels /æ/ (cash) /ɪ/ (this), /i/ (needs) /u/ (two) in sentences (iv) and (v).

A 14-coefficient linear prediction analysis of each vowel was used as a means of extracting formants and glottal "poles". The prediction coefficients were computed in the manner indicated by Atal and Hanauer.[5] The correct application of the linear prediction program was checked by analyzing synthetic speech. The determination of the exact nature of the computed poles (glottal "poles" or formants) was automatically made on the basis of a combination of bandwidth considerations and an anticipation of the expected frequency regions of the first five formants of each analyzed vowel. The features used in the study were the computed poles at the target locations of each vowel in the CVC utterance. The target position was defined as that point in the vowel segment for which the second formant reached a relative maximum.

The error criterion analysis showed that the second and fourth formant frequencies were the most effective identification parameters. In general, the formant frequencies were more significant than formant bandwidths and glottal "poles" for speaker recognition, but the other parameters were sometimes quite important for speaker-verification purposes. In vowels /i/ and /ɪ/, it was also noted that an extraneous wideband pole near 1300 Hz was a stable feature of many of the speakers and thus provided some identification potential.

The error analysis also indicated that the true statistical nature of the feature set was not reflected in only one session of test data. This conclusion was reached in all 92 measurements that were made.

b. Nasals

The nasals /n/ and /m/ were examined in the words needs and remember. The nasal spectrum is closely tied to the nasal cavity and certain attributes of this spectrum have been effectively incorporated in speaker-recognition schemes.[2] A 14-coefficient linear prediction analysis was used in this study to try to characterize the nasals in more precise fashion than has been done previously. The accurate pole-zero data obtained by

Fujimura[11] for nasals indicated the competence of the linear-prediction method in extracting pole locations, except in regions of pole-zero interplay.

The use of the error criterion showed that the nasal parameters were an especially rich source of recognition and verification features. The most promising measurements were the value of the formant frequency near 1000 Hz in /n/ and the value of the third or fourth resonances (1700-2300 Hz) in /m/.

c. Strident Consonants

The formant structure of the strident consonants is influenced by anatomical details around and forward of the alveolar ridge, and hence should display some recognition and verification potential. The stridents were examined in the words this (/š/) and cash (/š/). These sounds were analyzed by first sampling the waveform at 20 kHz and then computing a 10-coefficient spectrum in the middle of the frication region of the strident. The error criterion showed that the stridents were not as significant as the nasals and vowels sounds in characterizing a speaker. The value of the formant near 4300 Hz in /s/ and the value of the formant near 3400 Hz in /š/ did, however, provide some identification potential.

d. Fundamental Frequency

Fundamental frequency parameters have been found to be valuable recognition features by previous investigators.[2,8] In this study, we examined the F0 contour in sentence (v) by appealing to a stylized model of the pitch contour (Fig. XIX-8). The slope parameters RF0, F1F2 and F2F0 were measurements used to indicate the shape of an individual's pitch contour. The average fundamental frequencies associated with the voiced sections were also used as recognition parameters.

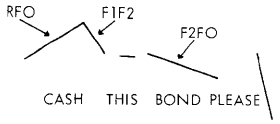

Fig. XIX-8. Stylized intonation pattern for the F0 contour.

The error analysis showed that the average frequency parameters were more important than the slope features in speaker recognition. The error analysis also indicated that the fundamental frequency of a speaker was quite variable across 5 recording sessions, and this variability diminished the total identification effectiveness of the measurements.

e. Timing Measurements

The learned behavior of an individual is reflected in the temporal aspects of his speech and provides a source of identification parameters.[3] The timing measurements investigated were the slope of the second formant in the diphthong /aɪ/ and the duration of the frication and aspiration noise of the plosive /k/ in cash.

The duration of the noise in the production of this /k/ ranged from 40 ms for one individual to 127 ms for another. This parameter was also quite stable within and across recording sessions and turned out to be an effective feature. The F2 slope in /aɪ/ was also quite variable among speakers and demonstrated excellent identification.

4. Conclusions

The error criterion was used to order the entire set of 92 parameters that were investigated. The first 38 of these features are listed in Table XIX-5.

Table XIX-5. Ordering of features.

Feature	Speech Event	Feature	Speech Event
1. NF2	/n/	20. THISF0	F0
2. UF3	/u/	21. MANF2	/m/ in man
3. IF2	/ɪ/	22. MANB3	/m/ in man
4. K	duration of /k/	23. EEF1	/i/
5. REMF3	/m/ in remember	24. EEF4	/i/
6. NF6	/n/	25. EEF3	/i/
7. REMF4	/m/ in remember	26. SHF2	/š/
8. CASHF0	F0	27. AEF2	/æ/
9. IF4	/ɪ/	28. AEF4	/æ/
10. AI	F2 slope in /aɪ/	29. AEF1	/æ/
11. REMFI	/m/ in remember	30. SF2	/s/
12. AVF0	F0	31. UF4	/u/
13. SF3	/s/	32. IF1	/ɪ/
14. UF2	/u/	33. BONDF0	F0
15. EEF2	/i/	34. REMF6	/m/ in remember
16. NF1	/n/	35. IF5	/ɪ/
17. MANF4	/m/ in man	36. MANB4	/m/ in man
18. UF1	/u/	37. AEF3	/æ/
19. NF3	/n/	38. SHF1	/š/

When interpreting this ranking of features, it is important to keep in mind that the ordering is established in accordance with the measurements of a given group of speakers; the speech characteristics of another group may result in a different ordering of features. For example, a group composed of both female and male speakers may result

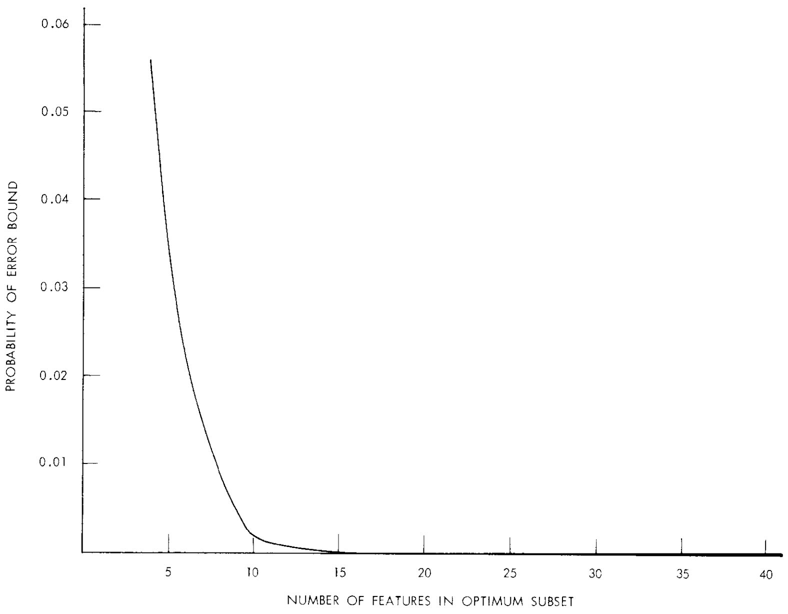

Fig. XIX-9. Error analysis for 38 features used in the identification scheme.

in higher relative ranking of fundamental frequency information than the group investigated here. In any case, the ranking shown in Table XIX-5 affords a general idea of what features are important in recognizing an unknown speaker.

Figure XIX-9 shows theoretical error bounds as a function of number of features. To test how accurately these error bounds reflect the performance of a set of features, an actual speaker recognition experiment was conducted. In the identification experiment only the five "best" features listed in Table XIX-5 were used, and it included data from speakers gathered in 320 tests during many sessions. Only 1 error was made in 320 speaker identification tests. The error rate of .003 was well within the predicted bound of 0.03.

This study furnishes a promising indication that effective speaker recognition and verification systems for certain applications can be designed. It is important, however, that the test data be collected during many sessions so that the true statistical nature of the measurement set will be reflected.

References

1. M. R. Sambur, "Speaker Recognition and Verification Using Linear Prediction Analysis," Ph.D. Thesis, Department of Electrical Engineering, M.I.T., September 1972.
2. J. J. Wolf, "Efficient Acoustic Parameters for Speaker Recognition," J. Acoust. Soc. Am. 51, 2044-2056 (1972).
3. O. Tosi, H. Oyer, W. Lashbrook, C. Pedrey, Julie Nicol, and E. Nash, "Experiment on Voice Identification," J. Acoust. Soc. Am. 51, 2030-2043 (1972).

4. K. N. Stevens, "Sources of Inter- and Intra-Speaker Variability in Acoustic Properties of Speech Sounds," Proc. VIIth International Congress of Phonetic Sciences, Montreal, Canada, August 21-28, 1971 (Mouton and Company, The Hague, Netherlands, in press).

5. B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am., **50**, 637 (1971).

6. G. S. Sebestyen, *Decision-Making Processes in Pattern Recognition* (The Macmillan Company, New York, 1962).

7. G. S. Sebestyen and A. K. Hartley, "Study Program of Pattern Recognition Research," Report No. AFSRL 6265, Litton System, Inc., Waltham, Massachusetts, December 31, 1961. (AD 273235.)

8. W. S. Mohns, "Statistical Feature Evaluation in Speaker Identification," Ph.D. Thesis, Department of Electrical Engineering, North Carolina State University, Raleigh, North Carolina, 1969.

9. P. D. Bricker, R. Gnanadesikan, M. V. Mathews, S. Pruzansky, P. A. Tukey, K. W. Wachter, and J. L. Warner, "Statistical Techniques for Talker Identification," Bell Syst. Tech. J. **50**, 1427-1454 (1971).

10. J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, (John Wiley and Sons, Inc., New York, 1965).

11. O. Fujimura, "Analysis of Nasal Consonants," J. Acoust. Soc. Am. **49**, 541 (1962).