XXVIII. SPEECH COMMUNICATION*

Academic and Research Staff

Prof. K. N. Stevens	Prof. A. V. Oppenheim	C.-W. Kim
Prof. M. Halle	Dr. Margaret Bullowa	N. Benhaim
Prof. W. L. Henke	Dr. Paula Menyuk	J. S. Perkell
Prof. D. H. Klatt	Dr. J. Suzuki†	Eleanor C. River
	K. Fintoft‡	

Graduate Students

J. K. Frediani	L. R. Rabiner	M. Y. Weidner
A. J. Goldberg	R. S. Tomlinson	J. J. Wolf

RESEARCH OBJECTIVES

The objective of the research in speech communication is to gain an understanding of the processes whereby (a) discrete linguistic entities are encoded into speech by human talkers, and (b) speech signals are decoded into meaningful linguistic units by human listeners. Our general approach is to formulate theories or hypotheses regarding certain aspects of the speech processes, obtain experimental data to verify these hypotheses, and simulate models of the processes and compare the performances of the models and of human talkers or listeners. Research in progress or recently completed includes: observations of the acoustic and articulatory aspects of speech production in English and in other languages through spectrographic analysis; study of cineradiographic data and measurement of air-flow events; study of the perception of speech sounds by children and examination of the acoustic properties of the utterances of children; computer simulation of articulatory movements in speech; investigation of the mechanism of larynx operation through computer modeling and acoustic analysis; examination of new procedures for analysis of speech signals using deconvolution techniques; experimental studies of the perception of vowel sounds; speech synthesis by rule with a computer-simulated terminal analog synthesizer; a re-examination of the system of features used to describe the phonetic segments of language; and the development and improvement of interface equipment for spectral analysis of speech with a computer and for synthesis of speech from computer-generated control signals.

K. N. Stevens, M. Halle

A. REAL-TIME SPECTRAL INPUT SYSTEM FOR COMPUTER ANALYSIS OF SPEECH

On-line operation of a real-time spectral input system for computer analysis of speech was achieved during the period covered by this report. The system, mentioned in a previous report,[1] was used with a bank of 36 bandpass filters and a PDP-1 computer to analyze recorded utterances played back in real time. A block diagram of the complete analyzing configuration is shown in Fig. XXVIII-1.

---

*This work was supported principally by the U.S. Air Force (Electronic Systems Division) under Contract AF19(628)-5661; and in part by the National Institutes of Health (Grant 5 RO1 NB-04332-04).

†On leave from Radio Research Laboratories, Tokyo, Japan.

‡On leave from Norges Laererhogskole, Trondheim, Norway.

Operation of the system is, at present, completely under program control. When in data-taking mode, the program continually pulses the real-time analyzer, thereby causing it to read and convert the output of each channel. Channel stepping is performed

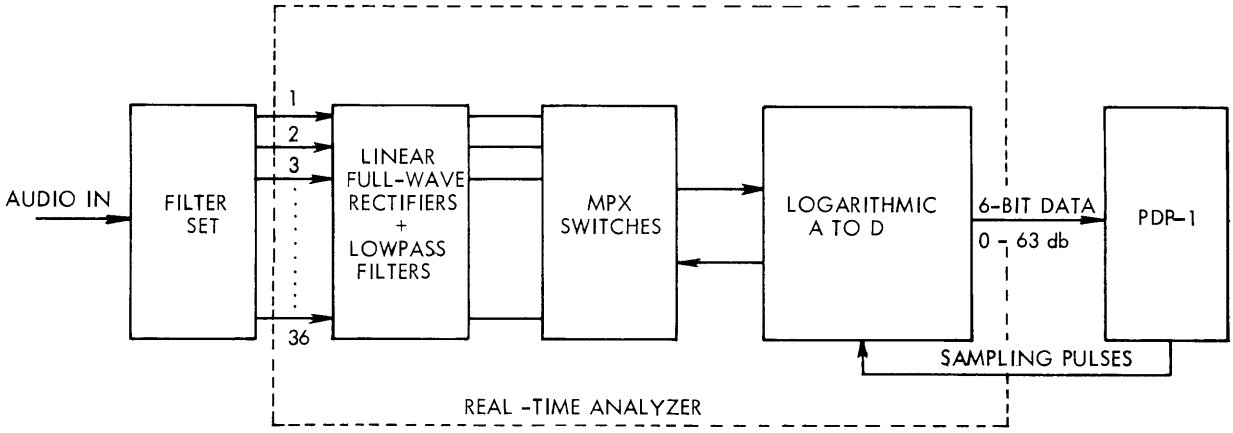

Fig. XXVIII-1. Diagram of the real-time analyzer as used with analyzing filter bank and PDP-1 computer.

at the end of each conversion by the internal logic of the analyzer. Digitized channel information is sent back to the computer and stored in the core. When the sum of the outputs of three selected channels rises above a set threshold, the program recognizes the onset of speech. Termination of speech is similarly recognized. The beginning and end thresholds and the sampling rate are program parameters. At present, 4000 words of data may be stored, representing approximately 3.4 seconds of speech at a 10-msec sampling rate. The program can display any given 36-channel spectrum sample and also each spectrum sample in sequence throughout the utterance. A display of selected channels outputs as a function of time is also available.

N. Benhaim, Eleanor C. River

References

1. N. Benhaim, "Real-Time Spectral Input System," Quarterly Progress Report No. 80, Research Laboratory of Electronics, M.I.T., January 15, 1966, p. 197.

B. CHILDREN'S PERCEPTION OF A SET OF VOWELS

In experiments comparing identification and discrimination functions for vowels in isolation and in consonantal context, it has recently been found that vowels in context tend to be perceived in a categorial fashion. Discrimination functions are characterized by peaks at the phoneme boundaries, whereas isolated vowels form a perceptual continuum. It is felt that these results support the theory that experience with the

generation of speech movements and with simultaneous observation of the acoustic consequences of these movements plays an important role in shaping the process whereby speech is perceived.[1]

In the case of a child the acoustic consequences of his speech movements differ quite radically from the consequences produced by the adults in his environment. The question asked in this experiment was: Are the phoneme boundaries for a set of vowels in consonantal context the same for the child as for the adult. Six children and 5 adults constituted the population of this study. The children, three boys and three girls, ranged in age from 5 to 11 years.

Each subject listened through earphones to 90 stimuli consisting of random presentation of 9 different synthetically produced CVC syllables. Two of the 9 syllables (steps 2 and 8) formed a typical version of b/i/l and b/I/l spoken by an American male. Five additional stimuli were produced by computing a set of interpolated formant contours that were equally spaced between b/i/l and b/I/l (steps 3-7). Two more were produced by extrapolating one step before b/i/l and one step after b/I/l (steps 1 and 9). These step sizes were equal to those between the interpolated stimuli. This produced 9 different syllables. (For a more detailed discussion of the stimuli see Stevens.[1])

Three black and white drawings pasted to a black surface were placed before each subject and identified by the experimenter as b/i/l (a truck), b/I/l (a bird's bill) and b/ɛ/l (a church bell). The subjects were asked to point to the picture that the speaker was naming. The percentage of judgments for each subject as /i/, /I/ or /ɛ/ for each of the 9 steps was then computed and a mean for adults and children was obtained. Table XXVIII-1 gives the numerical results.

Table XXVIII-1. Per cent of judgments /i/, /I/, /ɛ/ as a function of step.

Step	ADULTS %	CHILDREN %	ADULTS %	CHILDREN %	ADULTS %	CHILDREN %
	i		I		ɛ	
1	100	100	0	0	0	0
2	100	97	0	3	0	0
3	79	90	20	9*	1	0
4	62	73	36	27	2	0
5	15	24	81	75	3	0
6	0	5	97	95	3	0
7	0	1	95	98	5	1
8	0	0	86	85	14	15
9	0	0	42	32	58	68

*$p > .05$.

There were no significant differences between adults' and children's judgments on each step except for step 3 as computed by chi-square comparison. The tendency is for the adults to begin to perceive /I/ "sooner" in the continuum than the children, and for the children to do so with /ɛ/, although not significantly so. Figure XXVIII-2 shows that phoneme boundaries for the vowels in this experiment are the same, in terms of steps, for the children and for adults.

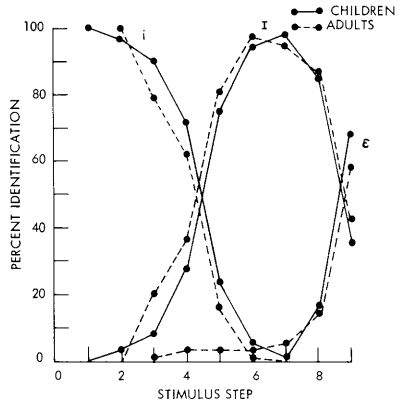

Fig. XXVIII-2. Phoneme boundaries for vowels in the experiment.

The answer to the question that was raised, therefore, is yes for the set of vowels in this experiment. This leaves us, however, with the task of trying to explain how the child can identify vowels in much the same way as adults when he produces these vowels so that they are, in terms of formant frequency, a poor match with those of the adults. We might also ask how the adult identifies the vowels produced by children.

There are several possibilities concerning the perceptual cues that may be in operation. Although the child's vowels do not match the adult's, his set of vowels are clearly differentiated from each other in formant frequencies. Furthermore, the direction that differences take for F_1 and F_2 between vowels is the same for adults and children. Also, there is no overlap between the vowels produced by children and adults. That is, for example, the /I/ produced by the child is not like the /i/ produced by the adult in terms of formant frequencies. Therefore, a system of distinctive differences exists between the vowels produced by children, as well as between the vowels of children and adults. (The data will be reported in detail elsewhere.[2]) Two further speculations would be that acoustic characteristics other than formant frequencies provide cues for identification and that articulatory gestures used by children to produce the vowels might be analogous to those used by adults. The first speculation is now being examined from data obtained in a previous experiment.[3] We have, at this time, no data on the articulatory gestures of children. We have the task of identifying those parameters by which the

child matches what he produces to what the adult produces and those by which the adult matches what he produces to what the child produces so that there is mutual understanding, which does, in fact, exist.

We have found that the child identifies certain vowels in consonantal context categorially, as does the adult, and that the boundaries of these vowels are strikingly similar for both children and adults. We would like to explore this question with other speech sounds; primarily, with those sounds that create difficulty developmentally, such as w, r, l, y.

Paula Menyuk

References

1. K. N. Stevens, "On the Relations between Speech Movements and Speech Perception, a paper presented at the 18th International Congress of Psychology, Moscow, August 1966.
2. Paula Menyuk, "Children's Production and Perception of a Set of Vowels" (in preparation).
3. Paula Menyuk, "Cues Used in the Perception and Production of Speech by Children," Quarterly Progress Report No. 77, Research Laboratory of Electronics, M.I.T., April 15, 1965, pp. 310-313.

C. ARTICULATORY ACTIVITY AND AIR FLOW DURING THE PRODUCTION OF FRICATIVE CONSONANTS

One of the principal objectives of research in speech is to understand the mechanism underlying the control of the speech-generating system. Several kinds of experimental observations can be made in order to investigate the nature of this process. Among these, air-flow and pressure measurements can provide useful information concerning the activities of the various articulatory structures. The purpose of this report is to describe one result of a larger study that has been reported on elsewhere.[1] Measurements of air flow during speech production have led to certain conclusions about the manner in which voiceless fricatives are produced in intervocalic position.

A face mask incorporating a linear flow resistance and a pressure transducer was used to measure the volume velocity of the air stream expelled from the lungs during speech production.[1-3] Figure XXVIII-3 is an example of the graphic record for the utterance "Say the word /hə'fɑf/ again." A double peak in the air flow occurs for each /f/ phone as indicated by the arrows. This type of double peak is characteristic of the voiceless fricatives /f, θ, s, š/ in the context of this frame sentence for all 5 speakers studied.

The double peak is probably a consequence of the relative timing of laryngeal and articulatory gestures. In Fig. XXVIII-4 a tracing of average air flow is compared with

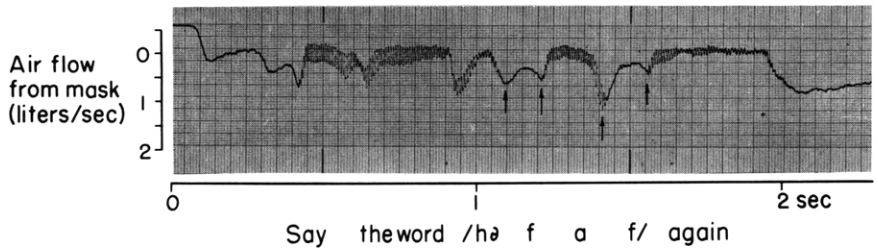

Fig. XXVIII-3. Example of the air-flow record (inverted) from the graphic recorder for the utterance "Say the word /hə'fɑf/ again." (Speaker: KNS.)

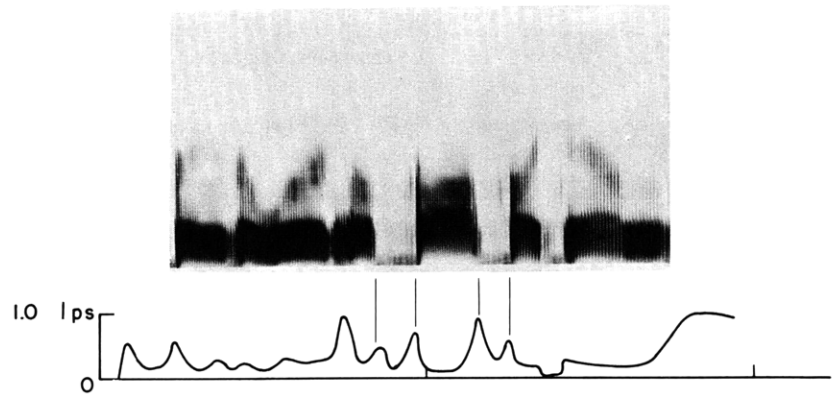

Fig. XXVIII-4. Spectrogram and tracing of average air flow utterance "Say the word /hə'fɑf/ again." The lines indicate the times of cessation and initiation of voicing in the syllable /fɑf/.

a spectrogram of the utterance to indicate the times of voicing onset and cessation. The lines that mark these times occur approximately at air-flow peaks.

An interpretation of these results in articulatory terms is suggested in Fig. XXVIII-5. During the unstressed vowel/ə/, the glottis begins to open while the vocal cords continue to vibrate. At the first peak in air flow (indicated by the first dashed line) the lower teeth begin to make contact with the upper lip, and the constriction that is formed causes a rise in mouth pressure. As a result, vocal-cord vibration ceases rather abruptly. The supraglottal articulator continues to constrict until vocal-tract resistance reaches a maximum value. The articulator then begins to move away in anticipation of the next phone, thereby lowering vocal-tract resistance. Air flow through the glottis thus increases, the vocal cords approximate as a consequence of the reduced pressure, and vocal-cord vibration begins (at the point indicated by the second dashed line). Finally, the vocal cords assume a mode of vibration which is characteristic of the vowel, with higher glottal resistance. A similar pattern of the air flow occurs in the

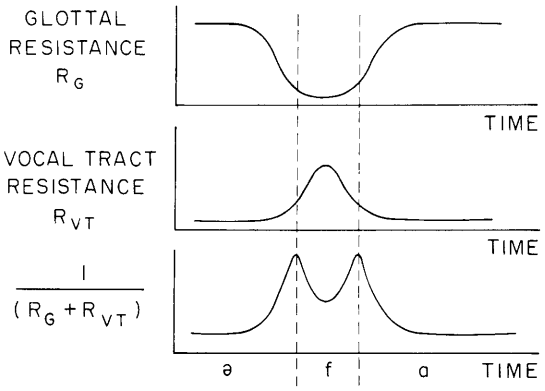

Fig. XXVIII-5. Interpretation of the articulatory events causing a double peak in the air-flow trace for a voiceless fricative in intervocalic position. Total resistance to flow is assumed to be the sum of glottal resistance and vocal-tract resistance. Dashed lines indicate times of cessation and initiation of voicing.

final /f/ of the utterance illustrated in Fig. XXVIII-4.

The voiced fricatives, /v ð z ž/ display similar air-flow traces, except that air flow is reduced relative to that of the voiceless fricatives, and the double peak is less pronounced. Voicing occurs throughout these sounds, but the flow becomes higher than for a vowel in spite of the turbulence-producing supraglottal constriction. Thus the laryngeal mode of vibration for voiced fricatives must differ from that occurring during vowels; the vocal cords probably remain separated during a vibratory cycle, with the result that there is an appreciable DC component to the flow.

In the course of the study,[1] data were gathered on other consonants of English. Some consonant clusters were recorded and found to have air-flow traces exhibiting coarticulation effects. Word stress was found to have some effect on air flow. The data suggest certain limits on the speed of reaction and coordination of larynx and vocal-tract structures during speech production.

The experimental data reported here were obtained at the Harvard School of Public Health, in collaboration with Dr. Jere Mead of its Department of Physiology.

D. H. Klatt

References

1. D. H. Klatt, K. N. Stevens, and J. Mead, "Studies of Articulatory Activity and Air Flow during Speech," Proc. Conference on Sound Production in Man, New York Academy of Sciences, 1966 (in press).

2. J. F. Lubker and K. L. Moll, "Simultaneous Oral-Nasal Air Flow Measurements and Cinefluorographic Observations during Speech Production," The Cleft Palate Journal, Vol. 2, pp. 257-273, July 1965.

3. N. Isshiki and R. Ringel, "Air Flow during the Production of Selected Consonants," J. Speech Hearing Res. 7, 233 (1964).