

XXIV. SPEECH COMMUNICATION*

Academic and Research Staff

Prof. K. N. Stevens
Prof. M. Halle
Prof. B. Gold
Prof. D. H. Klatt

Prof. A. V. Oppenheim
Dr. Margaret Bullowa
Dr. Paula Menyuk

Dr. A. W. Slawson
N. Benhaim
A. R. Kessler
Eleanor C. River

Graduate Students

R. D. Bruce
T. H. Crystal

W. L. Henke

L. R. Rabiner
R. S. Tomlinson

RESEARCH OBJECTIVES

The general objectives of our research in speech communication are to understand the processes whereby (i) human talkers encode sequences of linguistic units into sound, and (ii) human listeners decode an acoustic speech signal into linguistic units. The experimental approaches that we are following include examination of the mechanism of speech production through acoustic analysis, cineradiographic and other techniques, studies of speech perception, using synthetic speech and speechlike sounds as stimuli, computer simulation of various aspects of the speech-production and speech-recognition processes, and observation of speech perception and production as it develops in children.

K. N. Stevens, M. Halle

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

Development of a 36-channel analog-to-digital conversion system, which will sample the outputs of a frequency-analyzing filter bank and convert the results to provide a logarithmic amplitude scale, is nearing completion.

The system, which includes linear full-wave rectification and smoothing for each channel, sequentially samples the analog data and quantizes it directly into a 6-bit number that is read as decibels below a given reference. Logarithmic quantization is achieved by substituting a digitally controlled attenuator for the usual binary ladder network in conventional A-to-D systems. The principle of such an attenuator has been described in a previous report.¹

Over-all accuracy of the system is expected to be better than ± 1 db over a range of 63 db. Conversion time for each sample is, at present, 25 μ sec. A more detailed description of the system will be given in a subsequent report.

N. Benhaim

*This work is supported principally by the U.S. Air Force (Electronic Systems Division) under Contract AF 19(628)-3325; and in part by the National Institutes of Health (Grant NB-04332-03).

References

1. J. B. Dennis and E. C. Whitman, "Design Considerations for an Improved Vocal Tract Analog," Quarterly Progress Report No. 71, Research Laboratory of Electronics, M.I.T., October 15, 1963, pp. 206-214.

B. SPASS – AN IMPROVED TERMINAL ANALOG SPEECH SYNTHESIZER

This report is a brief presentation of work that is more completely described in the author's thesis.¹

1. Design Objectives

I would like to begin by outlining our objectives in designing SPASS. Basically, we were interested in designing a research tool for generating well-defined stimuli to investigate in detail the various characteristics of speech and related nonspeech sounds. It was imperative that the synthesizer surpass or equal present synthesizers in precision, accuracy, signal-to-noise ratio, and efficiency. We were not primarily interested in designing the output of a bandwidth compression system, or in devising any synthesis-by-rule schemes. These are, of course, still potential extensions of the basic synthesizer, and further work along these lines may possibly be done.

Conventionally, two types of resonance synthesizer configurations have been used. These are the cascade or series connection, and the parallel connection. The differences between these are slight. A cascade synthesizer requires extra circuits for the production of some sounds, but provides a slightly better approximation to the actual transfer function of the vocal tract, and, for most utterances, requires control of fewer parameters. Principally, because of the last reason, we chose the cascade configuration.

Most of the speech synthesizers that have been designed thus far, have used either electrical circuits or digital simulation to realize the required transfer functions. Both methods have advantages and disadvantages. Electrical circuits are fast, and can synthesize speech in real time. But the problem of instructing electrical circuits in how you want them to operate is a singularly difficult task. On the other hand, digital simulation affords a very flexible means of controlling the synthesis, but cannot be done in real time, thereby necessitating a large amount of core memory in which to assemble the waveform or some sort of off-line converter.

In our opinion, the most efficient method of using a computer for synthesizing speech, in most circumstances, is on-line. That is, the experimenter tells the computer what he wants, and the computer immediately produces a result, on the basis of which the experimenter tries something else, and so on. Any other method would be two or three orders of magnitude slower. A computer capable of simulating a speech synthesizer, however, would be prohibitively expensive to operate on-line.

We have taken the next obvious step of combining the good features of both schemes. That is, a collection of electrical circuits is used to generate the required transfer functions, and a small digital computer provides the link between the experimenter and the synthesizer. The combination allows synthesis in real time and extremely flexible control over the synthesizer. Also, the modest cost of a small computer makes on-line operation feasible.

2. Hardware Considerations

We shall now consider some aspects of the hardware design of SPASS. Essentially, the transfer functions of the formant resonators and other circuits are realized by solving the associated differential equations with analog computer components. For example, the transfer function of each resonator is produced with two operational amplifiers connected as integrators in a feedback loop. The resonant frequency and bandwidth are set independently by digitally controlled attenuators² whose attenuation varies in 1/4 db steps over a 64-db range. The attenuation value is set by 8 binary levels from the output of a flip-flop register whose contents are up-dated by control signals from the computer.

The excitation for voiced utterances is supplied by a single-shot multivibrator that generates a single narrow pulse on command from the computer. The pulses are filtered to produce a spectrum that rolls off at 12 db per octave above approximately 200 cps. By having the computer generate each pulse, their relative positions will remain the same on successive repetitions of the utterance. Thus it is possible to exactly reproduce an utterance at any time.

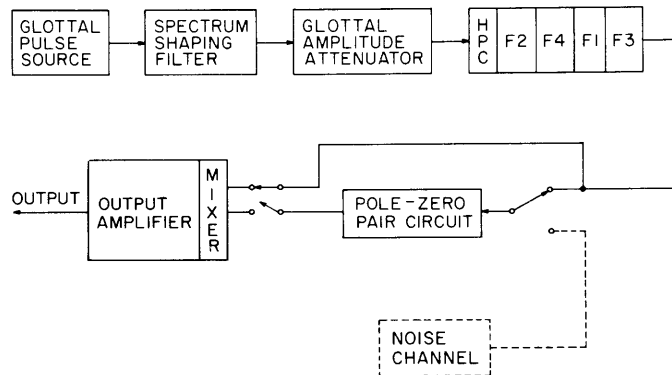


Fig. XXIV-1. Block diagram of SPASS during voiced sounds.

Figure XXIV-1 shows the configuration of the synthesizer for the production of voiced sounds. The signal originates from the glottal pulse source, is shaped by a filter, is

attenuated, and goes into the formant resonators. The signal path is then split, one path going directly to the output, the other first passing through the pole-zero circuit, which is used to introduce zeros into the transfer function for producing nasals. The configuration of the synthesizer is determined by a set of transistor switches some of which are shown here. These switches are set by digital signals according to information sent from the computer. For the production of fricatives, the input to the pole-zero circuit is switched to the noise channel output.

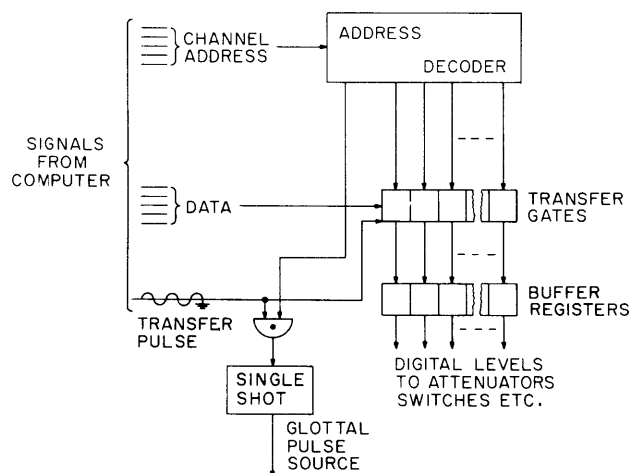


Fig. XXIV-2. Logic block diagram of SPASS.

Figure XXIV-2 is a block diagram of the logic that is necessary to decode and buffer the output of the computer. The operation is completely straightforward. Five digital levels from the computer are decoded into one of 32 possible channel addresses. Eight data levels are then jammed by the transfer pulse into the buffer register selected by the decoded address.

Other logic not shown here signals the computer when an overload occurs, and provides an accurate clock to govern the timing of an utterance.

3. Control Program

Perhaps the most important aspects of the design of SPASS are found in its control program. The synthesizer, alone, is completely useless and a poor control program would not offer much improvement.

The computer for this application must possess a cathode-ray tube display with a light pen, or an equivalent high-speed graphical display device, and some sort of fast-access bulk storage. Furthermore, it should be convenient to attach external i-o devices to the computer. For SPASS we are using a modified PDP-1 computer, built by Digital

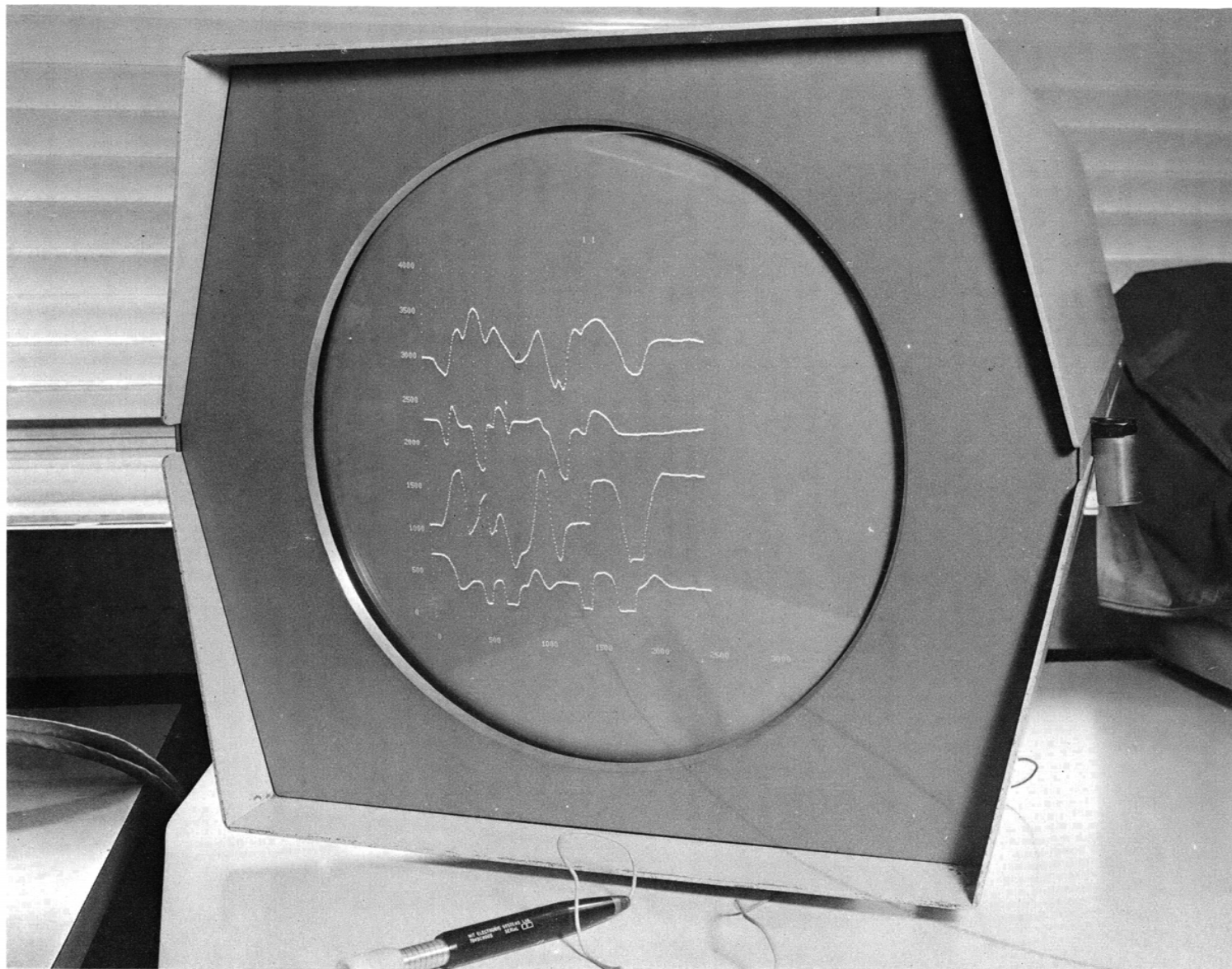
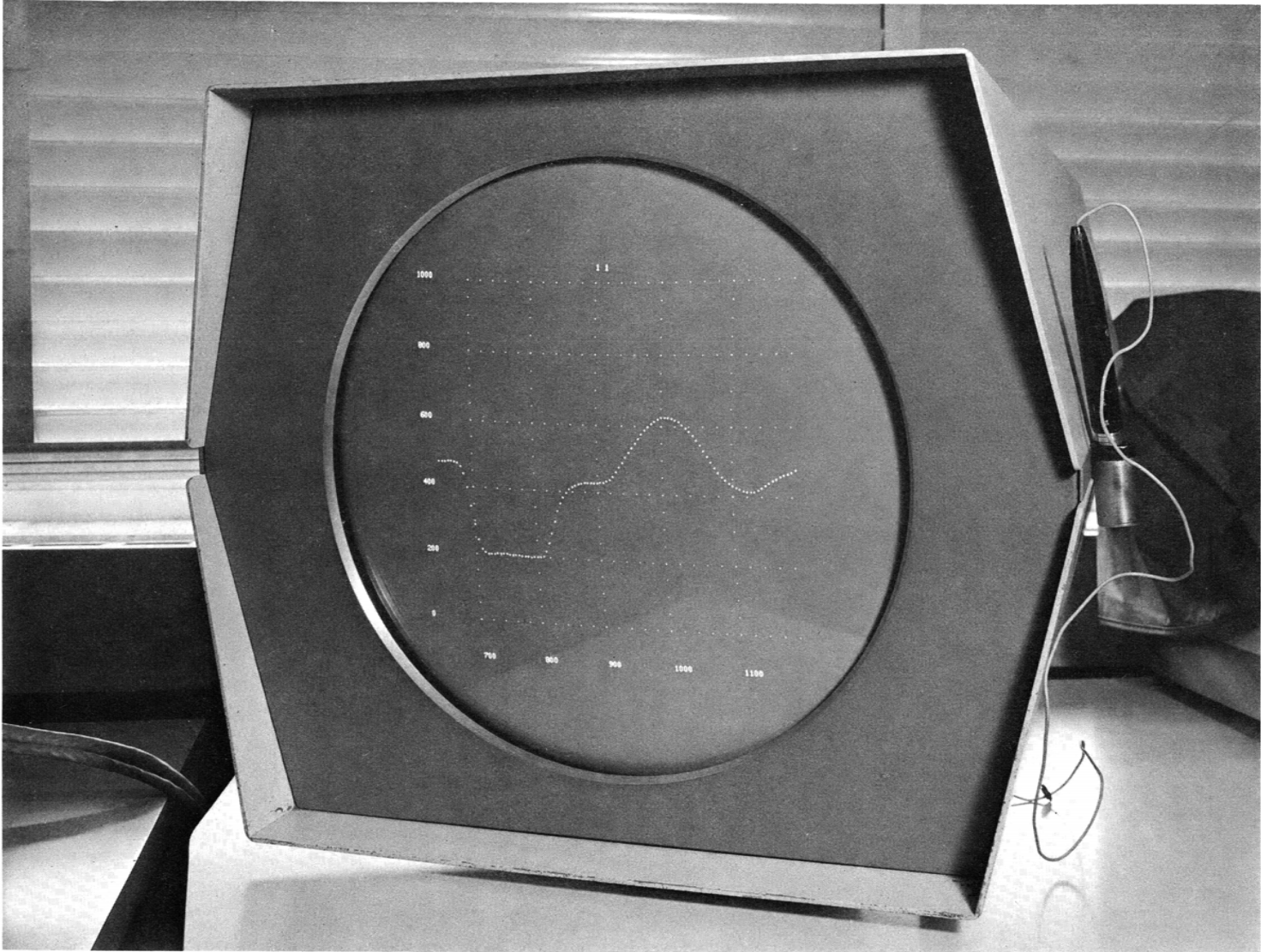


Fig. XXIV-3. First four formants of "Are you a good boy or a bad boy?"



(a)



(b)

Fig. XXIV-4. Correcting data with the light pen: (a) expanded picture; (b) illustrating the correction process.

(XXIV. SPEECH COMMUNICATION)

Equipment Corporation, which has 4096 18-bit words of core memory, a precision CRT display, and a high-speed magnetic drum that can hold 22 core loads of information. Other i-o devices include a photoelectric paper tape reader, paper tape punch, and a typewriter.

Two important requirements were felt to be necessary in the design of the program. First, the experimenter should not need to know how to program or operate a computer. Second, the experimenter is mainly interested in getting something done. Therefore, he should not have to engage in verbal gymnastics with the program to accomplish a simple task.

In SPASS, once the program is loaded, it is in complete control, and all of the program's features are summoned by typing commands to it. The only information about the computer which the experimenter needs is to know what sense switches and a typewriter are, how to put a tape in the tape reader, and a few other mechanical things.

The communication between the user and SPASS consists of typing single-letter commands on the typewriter. Similarly, SPASS replies are extremely brief, just sufficient to let the user know that the program is still alive.

We shall now discuss some of the more important features of the program. The first feature is a versatile graphical display routine. Included in the display is a completely labelled coordinate grid, upon which various input functions are plotted. Figure XXIV-3 is a plot of the 4 formant center frequencies of the utterance, "Are you a good boy, or a bad boy?"

Normally, whenever the program has nothing else to do, it displays data that are currently being worked on. Thus it is always immediately available for inspection. There are commands which cause expansion or compression of either scale, and move the display in any of four directions. An example of an expanded picture is shown in Fig. XXIV-4a. Here we see a display of F1 in the region of /b/ in the first /boy/ of the utterance just mentioned. The data for this utterance was copied from a sound spectrogram of the utterance spoken by the author by drawing curves for 4 formants, pitch, and amplitude with the light pen. In listening to the synthesized utterance, we detected a small error (among others) in this part of the sentence. It sounded something like /buhwoy/. This expanded view shows a plateau in F1 just after the plosive, instead of a continuous rise to the /aw/. It is a trivial task to correct it by redrawing this part of the formant with the light pen. The correction process is illustrated in Fig. XXIV-4b. When the correction is complete, we can go ahead and listen to what this sounds like. To give you an idea of the speed with which the light pen can be used, the data for this utterance (which lasts for a little more than 2 seconds) was drawn in approximately 35 minutes. Ten more minutes time was necessary to make corrections such as the one illustrated in Fig. XXIV-4.

In addition to light-pen input, there are routines for fitting polynomials of any order

to points entered through the typewriter, and the program is able to read input from paper tape.

The great flexibility of SPASS lies in its mapping specification. Essentially what the mapping specification does is to transform the data which have been entered, and which presumably describe, in some way, the utterance to be synthesized, into suitable data for transmission to the SPASS hardware. Since this transformation is very general, it is possible to specify an utterance in terms of variables other than the conventional formant frequencies, pitch, and amplitude. It is also possible to use the mapping specification to systematically vary utterances, interpolate between two utterances, and in general accomplish a result that heretofore would have required a great deal of manual labor.

An example will illustrate how the mapping specification is used. Between two words such as "beal" and "bill" or "bill" and "bell" we can construct a continuum of words by interpolation. We wish to test the ability of human subjects to discriminate between two words at various points along this continuum. We must therefore assemble a suitable test. In this case we propose to present the subjects with a series of stimulus triplets in an A B X form in which the first two words are different and the third is the same as one of the first two. Using a suitable mapping specification, the experimenter can select the interpolated word that he wishes to appear as A, B, or X by entering appropriate numbers as input variables. The program then performs the interpolation, automatically starts the tape recorder, produces the utterance, properly spaces the tape, and stops the recorder. Meanwhile, the experimenter can type the 12, or so, characters that are necessary to specify the next triplet, and the entire series of approximately 90 triplets can be assembled as fast as the computer can perform the arithmetic.

R. S. Tomlinson

References

1. R. S. Tomlinson, "An Improved Solid-State Terminal Analog Speech Synthesizer," S.M. Thesis, Department of Electrical Engineering, M.I.T., 1965.
2. R. S. Tomlinson, "The Design of a Digitally Controlled Attenuator," Quarterly Progress Report No. 74, Research Laboratory of Electronics, M.I.T., July 15, 1964, pp. 185-190.

C. CHILDREN'S LEARNING AND RECALL OF GRAMMATICAL AND NONGRAMMATICAL NONSENSE SYLLABLES

An experiment was undertaken to explore the effect of an aspect of grammaticalness versus nongrammaticalness on the learning and recall of verbal material and the role of maturation in this effect. The examined aspect was the composition of initial consonant

(XXIV. SPEECH COMMUNICATION)

clusters. Table XXIV-1 shows the sets of grammatical (initial consonant clusters permissible in English) and nongrammatical (not permissible) nonsense syllables used in the experiment.

Table XXIV-1. Stimuli sets.

Grammatical						Nongrammatical					
1a	kwoul	glId	drin	stais	trut	2a	kvoul	gzId	dlin	srais	tsut
1b	kwoud	glId	drid	staid	trud	2b	kvoud	gzId	dlid	sraid	tsud
1c	kwId	glId	drId	stId	trId	2c	kvId	gzId	dlId	srId	tsId

Thirty nursery school, kindergarten, first and second grade children (120 in all) were the subjects in the study. The age groupings were 4 years, 5 months to 8 years, 3 months. Each of the children were given three tasks: 1) to learn to associate a nonsense syllable name with a colored circle, all circles having the same color and the name of one member of a set; 2) to recall immediately all members of a set after they were spoken by the experimenter; and 3) to recall each member of a set, one by one, after it was spoken by the experimenter. Each child in the study was given either a grammatical or a nongrammatical set first, and then, one week later, its counterpart so that each child was tested on both types of material. The order of first presentation (grammatical or nongrammatical) was alternated. A typical sequence of 6 children would be presented with the following sets 1a, 2a, 1b, 2b, 1c, 2c, and the seventh child would begin the series again. In addition to these tasks each child was given a set of words beginning with the same consonant clusters found in the set of grammatical nonsense syllables, and asked to recall all members of this set after they were spoken by the experimenter. The set consisted of "glass, stick, drive, queen, truth."

Eighty presentations of stimuli was the maximum number of presentations in the learning task. Ten correct responses in sequence was the criterion for considering the set learned. In the recall task the set was presented maximally 10 times. All recalled responses were tape-recorded and then transcribed. An immediate (as it was produced) transcription of single responses was made for nursery school and kindergarten children. Mean response time was grossly calculated for each task.

For the learning task there were no significant differences because of grammaticality of material in the per cent of children who learned the task or in the per cent of correct responses at any of the age levels. The direction of ease of learning changed, however, in the grade levels, and may have interesting implications for the "proper" age to introduce a second language. At the kindergarten level both per cent of children who

learned the task and per cent of correct responses was higher for nongrammatical nonsense syllables. At the nursery school level, the dichotomization was exactly 50% to 50%, and at later grade levels grammatical material was more easily learned. At any grade level, mean response time for nongrammatical was greater than for grammatical materials; the average difference being 0.8 second. These results are shown in Table XXIV-2.

Table XXIV-2. Learning task.

	Nursery School N = 60	Kindergarten N = 60	1st Grade N = 60	2nd Grade N = 60
Learned (%) [*]				
Grammatical	50	44	52	54
Nongrammatical	50	56	48	46
Correct (%) [*]				
Grammatical	50	53	58	71
Nongrammatical	49	58	55	64
Mean Time (seconds) [*]				
Grammatical	8.1	9.7	10.3	6.8
Nongrammatical	8.5	10.6	10.7	8.1

^{*}Per cent of children learning set, per cent of correct responses, and mean time calculated.

In the recall task a very different pattern emerged. Total recall of a set occurred infrequently, and grade level was an insignificant factor. In all instances of recall it was always a grammatical set that was recalled. There were significant differences in per cent of correct responses at each grade level, with grammatical material being correctly recalled much more frequently. Again, as in the learning task, mean response time was greater for nongrammatical material, the average difference being 0.4 second. These results are presented in Table XXII-3.

Differences in learning and recall between sets were analyzed. The rank order of sets in per cent of correct responses in the learning task was 1a, 2a, 1b, 2b, 2c, 1c, and in the recall task 1c, 1a, 1b, 2a, 2c, 2b. Recall of words showed no significant differences over the age range in per cent of children learning the set or per cent of correct responses. Interestingly, there is a peak in percentage of children learning the set at the kindergarten level.

In summary, the learning task showed no significant differences for grammatical and nongrammatical material, while the recall task showed highly significant differences. The task of reproducing (perhaps articulating) nongrammatical material was much more

(XXIV. SPEECH COMMUNICATION)

Table XXIV-3. Recall task.

	Nursery School N = 60	Kindergarten N = 60	1st Grade N = 60	2nd Grade N = 60
Recalled (%)				
Grammatical	7	3	17	10
Nongrammatical	0	0	0	0
Correct (%)				
Grammatical	47	54	52	61
Nongrammatical	15	23	20	35
Mean Time (seconds)				
Grammatical	6.2	4.5	4.8	4.9
Nongrammatical	6.5	5.5	4.9	5.1

difficult for the children. The percentage of correct responses to grammatical stimuli in the learning task was significantly greater for second graders than for nursery school children, but this was not true for nongrammatical stimuli. The percentage of correct responses to nongrammatical stimuli in the recall task was significantly greater for second graders than for nursery school children, but this was not true for grammatical stimuli. In addition to the overall findings, other factors emerged: 1) the time needed to process nongrammatical material was greater than the time needed for grammatical material, regardless of the task; 2) the pattern of both learning nonsense syllables and recalling words changed over the grade levels with peaks at the kindergarten level for learning nongrammatical sets and recalling words.

A much more qualitative analysis of the data will be carried out to determine age differences, rather than gross educational-level differences, learning and recall curves over trials for different types of material, and pattern of substitutions in recall. The last task will be carried out to examine which features of the consonant clusters are substituted and which are not, and finally to determine, if possible, the hierarchy of feature maintainance in recall of unique morpheme length utterances.

Paula Menyuk

D. DESCRIPTION OF TONGUE SHAPE AND POSITION DURING VOWEL ARTICULATION

The purpose of this work was to develop a simple means for describing the shape and position of the tongue at its target position for vowels. The data were cineradiographic films of a speaker articulating certain utterances. For this study, tracings were made from the films to record the shape of the dorsal surface of the tongue in the midsagittal

plane, and the position of this shape with reference to the mandible and maxilla. This information was assumed to constitute an adequate representation of the tongue.

The description developed was based primarily on seven vowels as spoken by one American speaker of English. Each sample occupied the final position in a bisyllabic utterance of the form /hə'tV/. The frame at the temporal midpoint of the vowel was assumed to represent the closest approximation to the target position that was available. For convenience, the tongue was divided into three parts: tip, body, and base. The tip was the anterior 2 cm of the tongue, the base the posterior 3.5 cm, and the body the remaining length.

Perkell has previously noted that the tongue body has an approximately constant shape, called here the "vowel shape," for all vowels.¹ This observation was the basis for the model developed here. To completely describe the tongue it is necessary only to describe the position of this constant shape and the shape and position of the tip and base. The tip may be determined from the location of the mandible. The base is described by its variations from a fixed shape known as the central-base shape. These variations are at least qualitatively predictable from the location of the body.

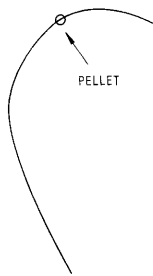


Fig. XXIV-5. Vowel shape and base central position.

The vowel body shape and base central position are determined by the following procedure. The tracings for the different vowel samples are overlaid so that the tongues coincide in the best possible manner. When the films used here were made, a small lead pellet was fixed to the surface of the tongue approximately one-third of its length back from the front extremity. To align the tongue shapes, the pellet locations were overlaid and the tracing rotated for the best possible match. When this had been done for all vowels,

an average shape was constructed. It is shown in Fig. XXIV-5 for the speaker examined here.

To locate the tongue body, it is then necessary to specify three coordinates. Two coordinates locate the position of the pellet in the midsagittal plane, and the third gives the angle of rotation about the pellet from a fixed direction that is required to properly orient the shape with respect to the maxilla. These coordinates for the 7 vowels studied are given in Fig. XXIV-6. The reference is to /a/.

The position of the tongue tip does not need to be specified accurately, but rather a range of positions is possible for each vowel. It is known that the tongue tip is low for all vowels. The anterior extremity is either in contact with or pointing toward a point within a "contact area" on the posterior side of the lower incisors and the floor of the

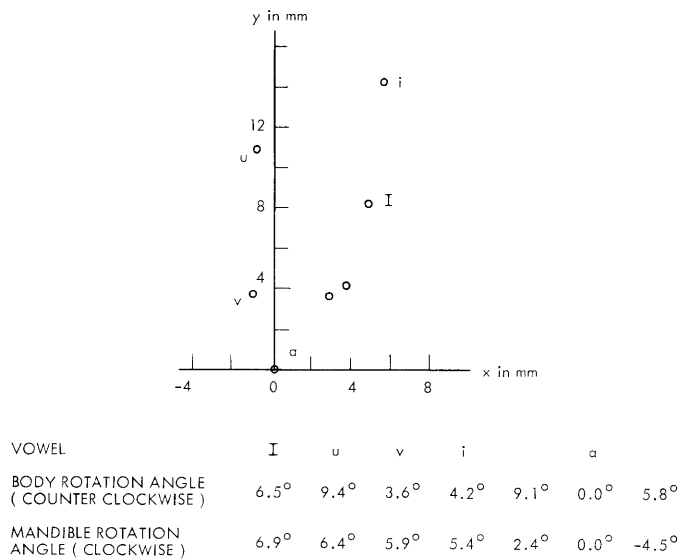


Fig. XXIV-6. Vowel coordinates for 7 vowels produced by one speaker. The reference vowel is /a/. The graph gives x and y coordinates of the pellet on the tongue; in tabular form below are listed the angles of rotation of body, tongue, and mandible.

mouth. The allowable range for the tip is then defined at the posterior end by the anterior end of the body, and at the anterior end by the contact area. The location of the contact area with respect to the tongue and the maxilla varies, however, with the position of the mandible. We found that the position of the tip of the lower medial incisor formed, to a very good approximation, a circle when the maxilla position was fixed. To specify the location of the mandible and therefore the tip range, then, it was merely

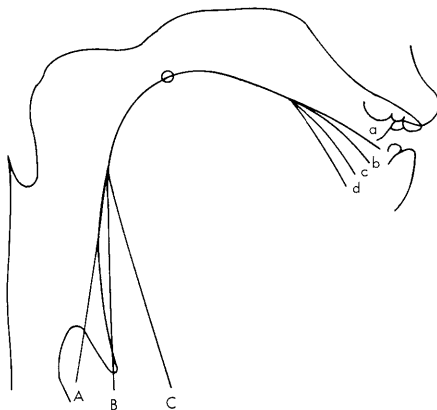


Fig. XXIV-7. Base and tip perturbations for the vowel /I/.

necessary to specify the rotation angle of this tooth from some arbitrary reference direction – in this case the direction for /a/. These values are shown in Fig. XXIV-6.

To complete the description of the tongue, we must show how the base varies away from its central position. Two effects are present. As the tongue rotates counterclockwise, the bottom of the base bulges away from the central position toward the rear. As the height of the tongue increases, certain muscles contract to force the hump upward. This

creates an "S"-shaped deviation from the base central position. While this description is only qualitative, it is presumed that with more data it could be improved upon.

To assess the validity of some of the approximations used in this model, an acoustical study was made. The techniques have been previously reported.² They provided a means for calculating the approximate formant frequencies resulting from the vocal-tract configuration specified by the corresponding tracing. First, variations in the tip were investigated. Three vowels were taken, and the tip position was varied over the allowable range. An example, /I/, is shown in Fig. XXIV-7. The formant frequencies were then calculated and compared with the unperturbed position. A similar study was made by using straight-line approximations for the base, also illustrated in Fig. XXIV-7. The resultant formant frequencies are shown in Table XXIV-4. In interpreting these figures, we assumed that the second formant is of primary importance in the acoustic determination of a vowel. It is apparent that the approximation for the tip has little effect on the second-formant frequency, but the base should be rather accurately specified if a given second-formant frequency is to be achieved.

Table XXIV-4. Formant frequency data for tongue tip and base perturbations shown in Fig. XXIV-7. Frequencies are in cps.

		a	b	c	d	A	B	C
/I/	F ₁	473	485	496	507	489	470	446
	F ₂	1807	1820	1821	1814	1810	1868	1919
	F ₃	2590	2593	2556	2487	2562	2596	2623
/i/	F ₁	297	324	335	344	305	293	273
	F ₂	2039	2069	2081	2090	2008	2071	2166
	F ₃	3049	3140	3175	3197	3066	3079	3090
/a/	F ₁	639	647	657	667	687	669	649
	F ₂	1299	1301	1303	1301	1198	1295	1329
	F ₃	2818	2805	2773	2728	2574	2693	2758

A study has been carried out to assess the use of this description during the transition period of a diphthong. This study gave some indication of the validity of the model in a dynamic description of the vocal tract. A second speaker was also studied, although in less detail than the first, with results similar to those for the first speaker.

The work reported here was done in partial fulfillment of the requirements for the degree of Master of Science. It is reported more fully in a thesis submitted to the Department of Electrical Engineering, M. I. T., August 1965.

J. A. Williams

(XXIV. SPEECH COMMUNICATION)

References

1. J. S. Perkell, "Cineradiographic Studies of Speech: Implications of a Detailed Analysis of Certain Articulatory Movements," a paper presented at 5^e Congrès International d'Acoustique, Liège, September 1965.
2. J. M. Heinz and K. N. Stevens, "On the Description of Area Functions and Acoustic Spectra from Cineradiographic Films of Speech," Quarterly Progress Report No. 74, Research Laboratory of Electronics, M.I.T., July 15, 1964, pp. 192-198.