XIX. SPEECH COMMUNICATION*

Academic and Research Staff

Prof. K. N. Stevens	Dr. Mary C. Bateson	Dr. D. H. Klatt
Prof. M. Halle	Dr. Margaret Bullowa	Dr. Paula Menyuk
Prof. W. L. Henke	Dr. H. Funkenstein	Dr. J. S. Perkell
Prof. A. V. Oppenheim	Dr. A. W. F. Huggins	A. R. Kessler
	Dr. Emily F. Kirstein	

Graduate Students

T. Baer	R. M. Mersereau	H. A. Sunkenberg
R. E. Crochiere	B. Mezrich	R. N. Weinreb
D. E. Dudgeon	M. R. Sambur	V. W. Zue
	J. S. Siegel	

RESEARCH OBJECTIVES AND SUMMARY OF RESEARCH

The broad aim of our research in speech communication is 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. Several projects directed toward these goals are active at present.

Studies of the mechanism of speech production have included examination of turbulence noise generation and of glottal activity during consonants, and these studies are leading toward a revised specification of the distinctive features related to laryngeal configurations. Projects on the modeling of the speech production process continue, with work on improved rules for the synthesis of certain consonants in stressed syllables, using a terminal analog synthesizer, and with the development of systematic procedures for specifying segment durations. In addition to the modeling of speech production our interests have broadened to include the development of techniques for on-line computer-aided music composition and synthesis.

Modern digital signal-processing techniques are being applied to speech in several projects. These include development of a method of frequency warping, for computation of spectra with unequal resolution using the fast Fourier transform. This technique is being applied to the problem of helium speech translation. Work on procedures for computer generation of speech spectrograms is also in progress.

Studies of the perception of voiced sounds characterized by time-varying fundamental frequency (F_0) are in progress, with projected experiments on F_0 contours of the type that occur in tone languages. In a series of experiments, we are examining the timing and rhythmic aspects of speech, through studies of the perception of sentences in which certain segment durations have been altered in a systematic way.

Research on lagnuage acquisition includes studies of the sounds that are used by infants in the first few months of life in interactive "conversational" situations and in other situations that can be reasonably well specified from the context, using previously acquired synchronized tape and film of children in their natural environment. Attempts are being made to classify these sounds according to the situations in which they occur

*This work was supported in part by the U.S. Air Force Cambridge Research Laboratories under Contract F19628-69-C-0044; and in part by the National Institutes of Health (Grant 5 RO1 NS04332-09) and M.I.T. Lincoln Laboratory Purchase Order CC-570.

and according to their acoustic characteristics. Experimental studies attempting to delineate the capabilities of children to produce and perceive sounds that are distinguished on the basis of certain prosodic features continue, and these experiments will be broadened to include investigation of particular segmental features as they occur in different phonetic contexts.

Most of the research projects described above make use of a digital computer facility comprising a PDP-9 computer with various peripheral items. The hardware and software capabilities of this facility are undergoing continual evolution. Current and recent activities include the addition of an in-house designed inexpensive but flexible display processor with a stroke-type character generator, and the specification and implementation of semantic extensions to a FORTRAN programming system oriented toward programming for on-line graphics and sonics. These extensions include machinery for the manipulation of "data structures" that greatly facilitate the design of graphics-using systems. Some general-purpose systems implemented using this extended FORTRAN system include DYNAMO – a general continuous system simulator, and MITSYN – a music synthesis oriented system. MITSYN includes independently useful subsystems for the on-line graphical specification of signal-processing networks built up by the interconnection of signal-processing primitives, and a parameter notation and editor used to create files of parameter values vs time for control of systems (often sound synthesis) which include temporally varying parameters as input.

A facility for presenting arbitrary sequences of prerecorded audio stimuli is in the final stages of development. The stimuli can be monotic or dichotic, and can be synthetic, natural, or edited-natural waveforms. The addition of disk-storage to our PDP-9 computer has made it possible to present the sequences in real time, and thus run subjects on-line, with the result that adaptive procedures can be used, in addition to the more mundane application of producing experimental tapes.

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

A. ANALYSIS OF VOCAL-TRACT X-RAYS USING INTERACTIVE COMPUTER GRAPHICS

1. Introduction

Present knowledge of speech articulation is largely based on still and motion pictures of vocal-tract x-rays. Still pictures are appropriate for the study of isolated, sustained speech sounds, but motion pictures are required to investigate articulatory dynamics. Cineradiography samples the positions of the articulators at a rate corresponding to the cinematic time base – typically in the range 24-200 frames per second. The analysis of cineradiographic films generally is tedious and time-consuming, since most applications require that the projected images of individual cineradiographs be traced by hand. Even when the manual tracing is complete, further labor must be invested to obtain useful articulatory data. For example, it may be desirable to measure the vocal-tract cross dimensions, calculate the rates of articulatory movement, or compare various shapes and positions assumed by an articulator.

This report describes some preliminary applications of interactive computer graphics to the analysis of vocal-tract cineradiographs. Our objective was to reduce

the investment of human time and to provide for a convenient method of data storage. The programming language[1] is an extended version of FORTRAN IV that, among other things, affords the use of flexible data structures, dynamic free storage allocation, and man-machine interaction by means of pen and tablet, push buttons, toggles, and knobs. The computer facility has been described by Henke in a previous report.[2]

The computer-aided analysis involves two programs: VTGRIN (vocal-tract graphical input) and VTDISP (vocal-tract display). VTGRIN provides for the input of the vocal-tract configurations to the computer. The user draws the vocal-tract profile with the pen and tablet, and a corresponding computer-generated image is displayed simultaneously on the screen of a cathode-ray tube (CRT). When the drawing for a given frame is complete, the necessary data for reconstruction of the vocal-tract image are stored in a file. The second program, VTDISP, is used to read the data from the file and to construct the vocal-tract display. This program also includes procedures for vocal-tract analysis, the formation of composite displays (two or more vocal-tract configurations displayed simultaneously) and display embellishing (labeling).

2. Vocal-Tract Graphical Input (VTGRIN)

The cineradiographs (or cineradiograph tracings) to be used for graphical input can be rear-projected through the tablet, overlaid on the tablet surface (tracings only), or

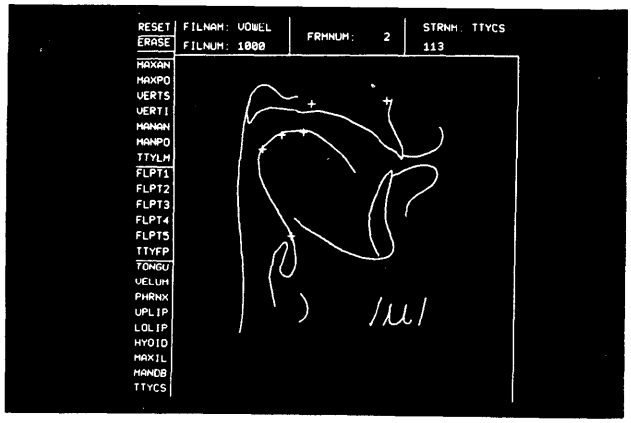


Fig. XIX-1.

CRT display generated by VTGRIN, showing the vocal-tract configuration for the vowel /u/. The mnemonics on the left-hand border constitute a command menu for error correction and the selection of articulatory objects. The display title (top), carries the file, frame and structure identification. It also serves as a command area for file-write instructions.

projected onto the phosphor of a CRT that stands directly in front of, and perpendicular to, the tablet. The operation of the program is independent of the method selected, since the tablet always serves as the input device. The user gains a major advantage from the method of CRT projection, however, because this method allows him to view the projected x-ray image and the computer-generated image in the same plane. Hence errors in tracing are readily apparent as discrepancies between processed and unprocessed vocal-tract profiles.

Upon program initiation, the teletype requests a file name, file number, file comment string, and frame number, and incorporates this information in its display title at the top of the picture. The program is then ready to accept graphical input. A photograph of a CRT display generated by VTGRIN is shown in Fig. XIX-1.

VTGRIN can be regarded as "articulation-oriented" insofar as separate graphical structures are created for different articulators or articulatory segments. The "names" of the various articulatory objects are selected from the command menu, which echoes pen-sensitive program-control areas of the tablet, that is, areas in which pen contacts evoke specific program actions. The user monitors the position of the pen on the tablet by means of a cursor in the CRT display. Activation of the pen when its cursor is in the vicinity of an articulatory name causes the selection of that name for object identification. In other words, the pen cursor is used to choose the desired mnemonic from the command menu. The current articulatory name and a corresponding identification number appear in the right-hand corner of the display title.

The program recognizes two types of articulatory objects: point objects, and continuous objects. Point objects are of two kinds: landmarks, or bony structure reference points; and fleshpoints, or radiopaque markers attached to fleshy surfaces. Point objects consist of a single pair of x and y coordinates, indicated in the display by a cross having the coordinate pair as center. In Fig. XIX-1, two landmarks lie within the outline of the maxilla, and four fleshpoints appear on the lingual contour.

The following mnemonics in the command menu serve as landmark names.

MAXAN:	maxilla, anterior point
MAXPO:	maxilla, posterior point
VERTS:	vertebrae, superior point
VERTI:	vertebrae, inferior point
MANAN:	mandible, anterior point
MANPO:	mandible, posterior point
TTYLM:	miscellaneous landmarks having identification numbers that are specified by teletype input or supplied by the system.

The miscellaneous landmarks (TTYLM) allow for identifications not anticipated by the more explicit names. Unlike the other landmark mnemonics, TTYLM can be selected repeatedly within a frame.

The mnemonics for fleshpoints are FLPT1, FLPT2, FLPT3, FLPT4, FLPT5 and TTYFP. If more than 5 fleshpoints are to be identified in a given frame, the mnemonic TTYFP is selected and unique identification numbers are input on the teletype or supplied by the system. In this way, a virtually unlimited number of fleshpoints can be specified in a single frame.

The second type of articulatory objects, continuous objects, are the outlines of

vocal-tract structures. These objects are associated with a variable number of data points and are realized in the display as line-incremented graphical structures, with the following mnemonics.

TONGU:	tongue
VELUM:	velum
PHRNX:	posterior pharyngeal wall
UPLIP:	upper lip
LOLIP:	lower lip
HYOID:	hyoid bone
MAXIL:	maxilla
MANDB:	mandible

The mnemonic TTYCS, like the TTYLM and TTYFP, enables identification of an arbitrary number of additional objects in any given frame. Moreover, this menu item can be used for the input of freehand script (e.g., the phonetic symbol in Fig. XIX-1) and the drawing of straight-line segments.

The two commands RESET and ERASE, which appear in the upper left-hand corner of the display, enable the user to correct errors. RESET aborts all of the data for the current frame, requests a frame number, rebuilds the display title, and waits for graphical input. ERASE deletes the graphical structure only for the current articulatory object and then accepts new graphical input for that object.

The display title (Fig. XIX-1), reflects another tablet command area. A pair of vertical lines divides the display title into three blocks which correspond to program-control blocks on the tablet. Pen contacts in these blocks are used (i) to close the data file and return to initial program status, (ii) to prepare for a new frame whose identification number is to be supplied by teletype input, or (iii) to prepare for a new frame whose identification number is the current frame number incremented by one. With each of these commands, the data that have been entered for the current frame are written in file storage.

3. Vocal-Tract Display (VTDISP)

VTDISP is designed to complement VTGRIN and comprises routines for display from file, translation and rotation, vocal-tract measurement, and labeling. The user commands are given by pushbutton interrupts and toggle settings. Also, knobs, or potentiometer controls, are employed to adjust continuous or multivalued variables.

To initiate the program, the identification of the desired file and frame are entered by the teletype. A default value of zero may be used for the frame number, in which case the number of the first frame in the file is assumed. Once the file has been opened, the file identifiers and comments are read from the file and written on the

teletype. Then the file is searched for the desired frame and, if found, the data are read and the vocal-tract image is constructed.

The user commands are outlined below. The primary program functions (those written in upper-case letters) are called by pushbutton interrupts, whereas the secondary functions are requested by toggle switches. In order to illustrate the program actions, representative displays generated by VTDISP are shown in Figs. XIX-2 through XIX-7.

RESET: Recalls the initial program status; aborts all current graphical structures.

SCALE: Establishes a conversion factor (tablet units to centimeters) for scaling purposes. A known life-sized distance is entered with the pen and tablet.

ANALYZE: Generates a transverse grid for vocal-tract analysis and a menu for measurement parameters (Fig. XIX-2). This grid is regarded as preliminary, and alternative patterns (conforming more closely to an axial line with a series of normals) are under consideration. The grid shown in Fig. XIX-2 can be adjusted to individual

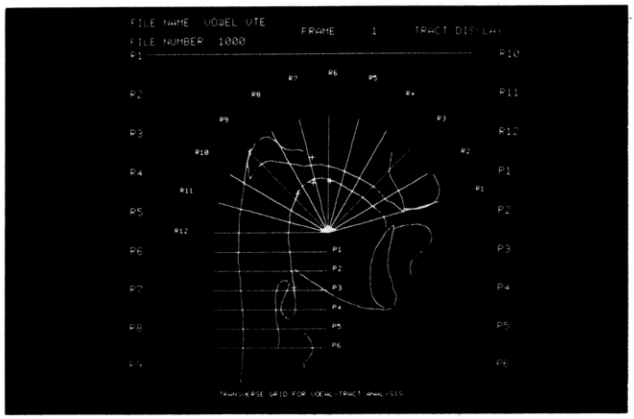


Fig. XIX-2. CRT display created by the ANALYZE function of VTDISP. The display includes the vocal-tract configuration for the vowel /i/, a transverse grid built of radial and parallel lines, and a menu for selection of the analysis parameters.

vocal-tract configurations by means of translation, rotation, and changes in the spacing of the parallel grid lines (lines P1-P6). Depending upon the status of the toggle switches, the measurement parameters displayed on the left- and right-hand margins of the CRT can be selected with the pen and tablet or updated automatically. To determine the vocal-tract cross dimensions, the user marks the intersections of the grid lines and the boundaries of the vocal tract with the pen and tablet. The computer calculates the metric equivalents of the recorded measurements and generates a plot of the vocal-tract cross dimensions (Fig. XIX-3). Note that the labial extension of the

vocal tract is neglected in the present analysis procedure.

TRANSLATE: Effects translation of the vocal-tract image on the CRT. After the user has responded to a teletype request for the reference-point structure number (necessarily a point object), horizontal and vertical positioning of the vocal-tract image can be adjusted with the pen and tablet.

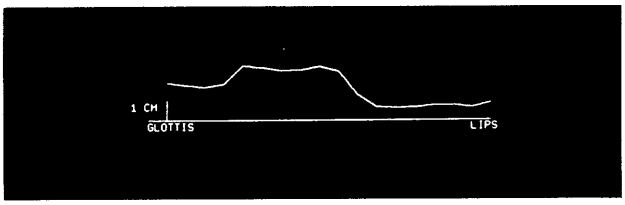


Fig. XIX-3. Display of vocal-tract cross dimensions for the vowel in Fig. XIX-2. These dimensions are determined by the intersections of the grid lines with the boundaries of the vocal tract.

STORE: Saves the current vocal-tract image and requests, by the teletype, a display intensity for the next frame. Hence, a composite of two or more frame images can be constructed by using different display intensities to keep the separate images distinct.

REGISTER: Sets reference-point coordinates for registration (translation and rotation) of future frames. The teletype requests a display intensity for the next frame, and, depending upon the status of the toggle switches, the current vocal-tract image is saved or destroyed. Like the STORE function, this command is useful for the construction of composite displays. The displays in Figs. XIX-4, XIX-5, and XIX-7 were built with the REGISTER command used.

NEXT FRAME: Reads the data for the next successive frame in the file, builds the new vocal-tract image, and displays the new image with any images preserved by the commands STORE or REGISTER. The composite of two vocal-tract images shown in Fig. XIX-4 was generated by calling consecutively the two commands REGISTER and NEXT FRAME. Registration of the two frames was based on the two landmark reference points that lie within the maxillary border. Two different display intensities were employed in order to make the articulatory positions for vowel /i/ distinct from those for vowel /u/. The four crosses on each lingual contour represent the positions of radiopaque markers placed on the tongue. The markers are valuable in studying the relative movements of different portions of the tongue. Because the tongue is capable of changes in both position and shape, a point parametrization of this structure constitutes a convenient and informative method of analysis.

An application of point-parametrized vocal-tract data to the study of articulatory dynamics is illustrated in Figs. XIX-5 and XIX-6. Figure XIX-5 shows the positions

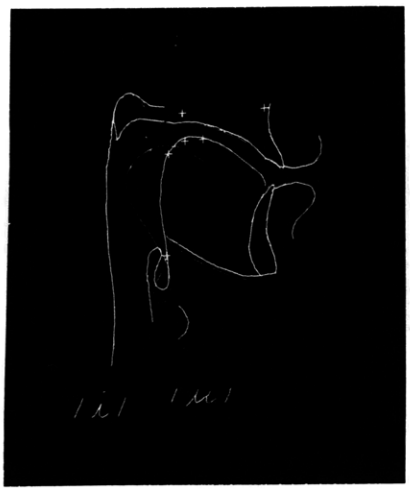


Fig. XIX-4. Composite display generated by VTGRIN for the vocal-tract configurations of the vowels /i/ (greater intensity) and /u/ (lesser intensity). The two frame images were registered with respect to the landmark reference points within the maxillary border.

of three points on the tongue and a point on the hyoid bone as recorded at 40-ms intervals during the word "coy." Direction lines indicating the course of movement for each point are presented in Fig. XIX-6. From top to bottom, the movement paths pertain to points on the dorsal aspect of the tongue, on the back of the tongue, on the tongue root, and on the midpoint of the anterior projection of the hyoid bone. Note that during the release

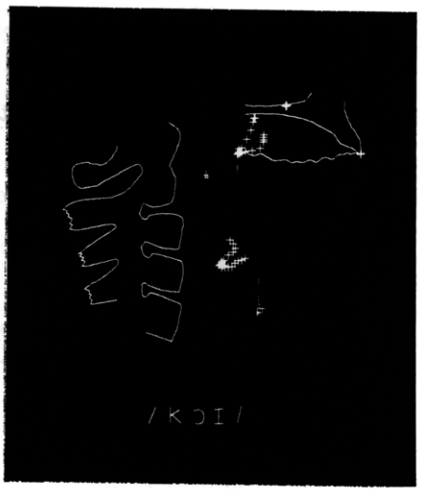


Fig. XIX-5. Composite display showing the displacement patterns of 3 points on the tongue and a point on the hyoid bone during the utterance "coy." The positions of the four points are recorded at intervals of 40 ms. See Fig. XIX-6 for an illustration of the direction of motion.

of the dorsal stop /k/, the tongue-and-hyoid-bone assembly moves downward, but very little hyoid-bone movement occurs for the dipthong segment. The displays in Figs. XIX-5 and XIX-6 were produced with the functions REGISTER and NEXT FRAME,

and two display intensities were used in the case shown in Fig. XIX-5. The two crosses positioned on the maxilla served as reference points for frame registration.

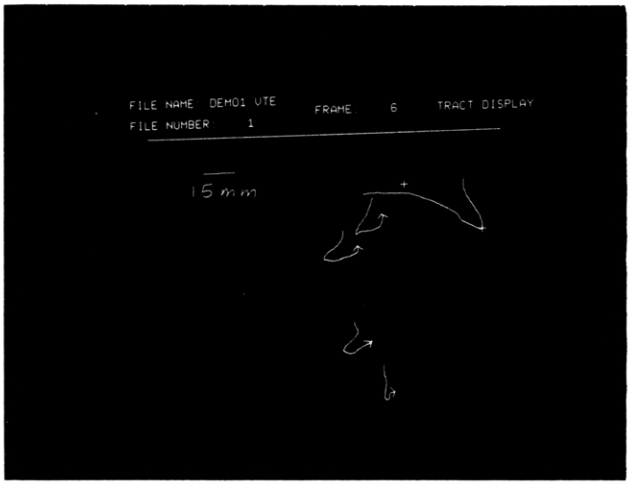


Fig. XIX-6. Lines showing the movement paths of the four points in Fig. XIX-5.

Fig. XIX-7.

Labeled composite display for vocal-tract configurations corresponding to the initial and final elements in the word "we." Labels were affixed to the display by means of the TEXT function in VTDISP. Registration of the two frame images was based on the landmark reference points lying within the maxillary border.

SEEK FRAME: Requests a new frame identification number from the teletype, searches the current file for the desired frame, reads the data, builds the vocal-tract image, and displays the new image with any previously stored images. The composite display in Fig. XIX-7 was created with the commands REGISTER and SEEK FRAME. The data file from which these frames were taken contained the vocal-tract configurations recorded at 40-ms intervals during the phrase "we saw." The initial and final

elements of the word "we" that are shown in Fig. XIX-7 were separated by a period of 240 ms, so several frames in the file were skipped to produce this composite display. Again, two display intensities were employed.

SEEK FILE: Closes the currently open file and returns to the initial program status (except that stored frame images are preserved).

TEXT: Accepts strings of text from a keyboard located near the tablet and displays them on the CRT. The position of each text line is fixed in the display with the pen and tablet. This function is useful for labeling, as in the case of Fig. XIX-7, which contains a title and phonetic symbols affixed to each vocal-tract configuration.

During this research, the author benefitted from many discussions with W. L. Henke, whose cooperation is gratefully acknowledged.

R. D. Kent

References

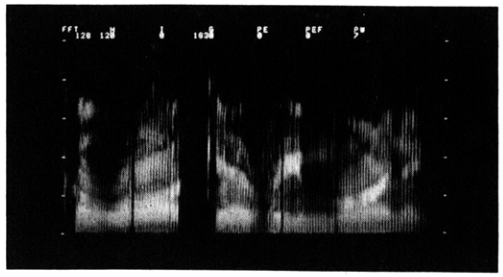
1. W. L. Henke, "An Interactive Computer Graphics and Audio System sans Large Budgets and Great Fuss," Quarterly Progress Report No. 98, Research Laboratory of Electronics, M.I.T., July 15, 1970, pp. 126-133.
2. W. L. Henke, "Speech Computer Facility," Quarterly Progress Report No. 90, Research Laboratory of Electronics, M.I.T., July 15, 1968, pp. 217-219.

B. AN APPLICATION OF THE CEPSTRUM AS A MEASURE OF THE AMPLITUDE OF A SIGNAL

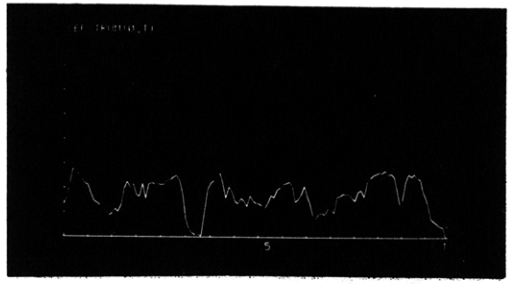
The cepstrum of a signal is the inverse Fourier transform of the log magnitude of the Fourier transform of that signal. If we consider two time functions such that one is a constant multiple of the other, then their cepstra are identical except at the origin. Each cepstrum has an impulse at the origin, the area of which varies with the gain of that signal. This suggests the possibility of using the cepstrum value at zero to measure signal amplitude.

To test this hypothesis, the following experiment was tried. A speech waveform was multiplied by a time window of finite width whose position T could be varied along the speech waveform. The cepstrum value at zero was then computed for several values of T and the results were plotted. The total energy of the speech in the time window was also computed so that this amplitude measure could be compared with the cepstrum.

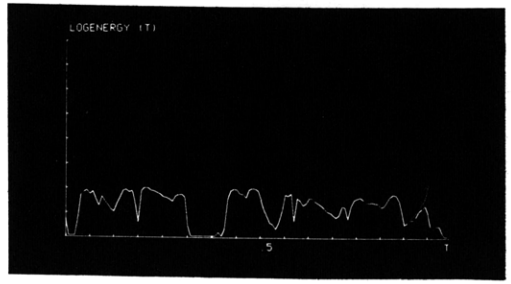
In Figs. XIX-8 through XIX-12 these two quantities are compared; the formats of the figures are identical. In each figure (a) is a spectrogram of each speech segment, (b) is the cepstrum evaluated at quefrency = 0 as a function of window position, and (c) is the logarithm of the energy as a function of window position. In all three curves the same time axis is used.



(a)



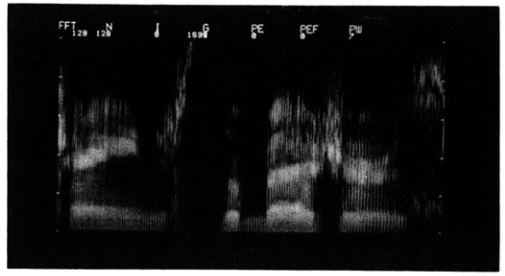
(b)



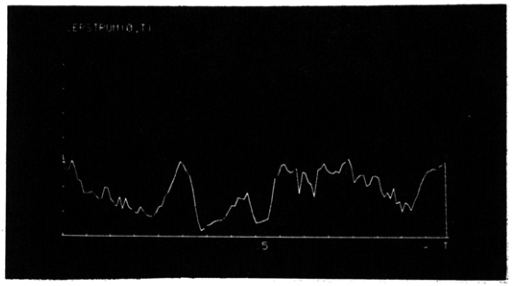
(c)

Fig. XIX-8.

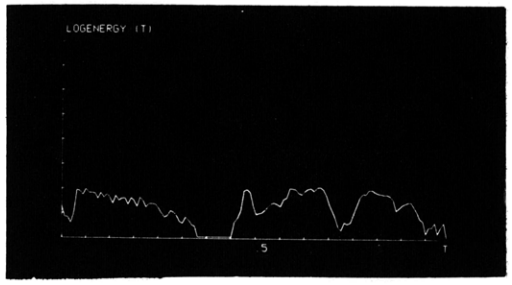
(a) Spectrogram, (b) cepstrum(0, T), and (c) log(energy(T)) for the sentence "... took a walk every morning."



(a)



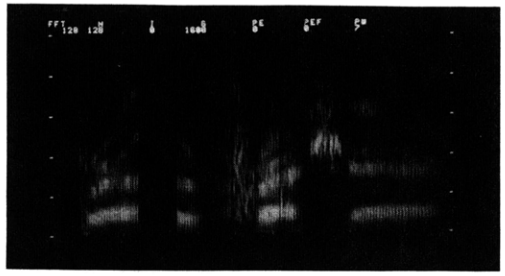
(b)



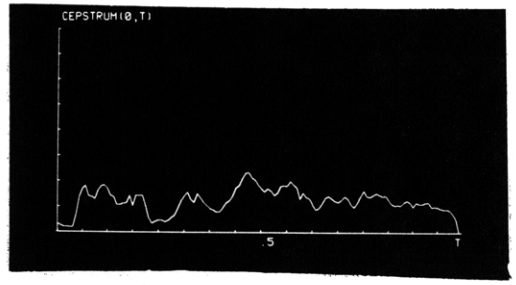
(c)

Fig. XIX-9.

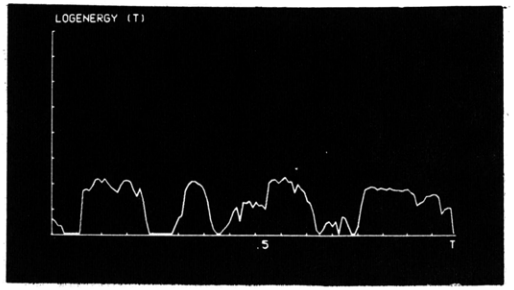
(a) Spectrogram, (b) cepstrum(0, T), and (c) log(energy(T)) for the sentence "... changed the measures."



(a)



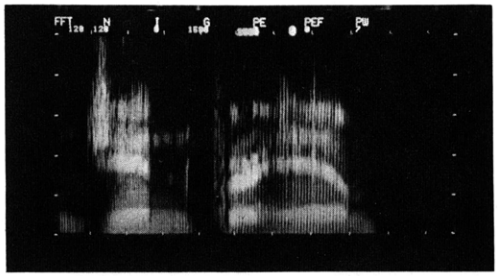
(b)



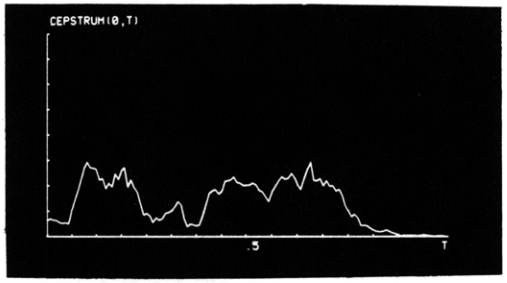
(c)

Fig. XIX-10.

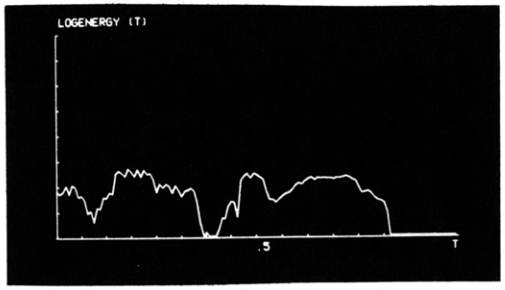
(a) Spectrogram, (b) cepstrum(0, T), and (c) log(energy(T)) for the sentence "... breath of fresh air."



(a)



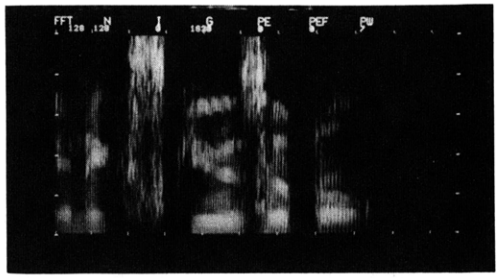
(b)



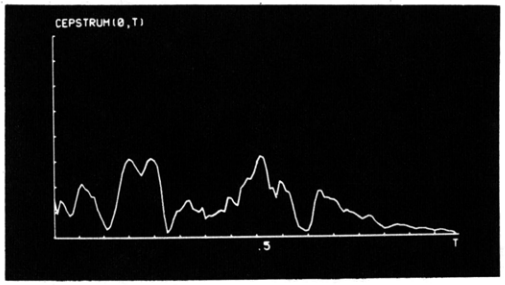
(c)

Fig. XIX-11.

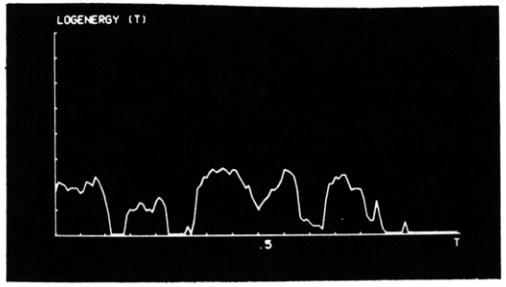
(a) Spectrogram, (b) cepstrum(0, T), and (c) log(energy(T)) for the sentence "... gin for him."



(a)



(b)



(c)

Fig. XIX-12.

(a) Spectrogram, (b) cepstrum(0, T), and (c) log(energy(T)) for the sentence "... inexcusable."

It will be noticed that in Fig. XIX-8 the two curves have similar gross features. In Fig. XIX-9 they are similar except for the ⟨g⟩ of "changed" and the ⟨s⟩'s in "measures." For these fricatives the cepstrum shows definite peaks, whereas the energy curve has minima. In Fig. XIX-10 the curves agree except for the ⟨f⟩ and ⟨sh⟩ of "fresh," in Fig. XIX-11 they are similar except for the ⟨g⟩ of "gin," and in Fig. XIX-12 they agree except for the ⟨s⟩ of "inexcusable."

In conclusion, the short-time cepstrum can be useful as a measure of amplitude. When used on some fricatives (voiced and unvoiced) it yields a larger amplitude than would be obtained from using the energy. It is a potentially useful tool to aid in speech segmentation and speech analysis.

R. M. Mersereau, A. V. Oppenheim

C. SYNTHESIS OF SUNG VOWELS

Techniques for synthesizing speech by a machine or by a computer simulation of the speech mechanism have reached the point where intelligible and natural-sounding words and sentences can be generated. Few attempts have been made, however, to simulate the singing voice, and informal efforts in this direction have usually led to unnatural, mechanical-sounding tones. The purpose of this study is to develop procedures for synthesizing acceptable sung tones by systematically manipulating some of the parameters of a vowel synthesizer. One goal of the synthesizer experiments is to gain insight into the attributes of a singer's voice that contribute to superior quality, as estimated by experienced judges of the singing voice.

1. Vibrato Experiment

One of the features of a good sung tone (as judged by trained listeners) is a periodic fluctuation of the fundamental frequency about a mean frequency. This vibrato produces modulations in the amplitudes of harmonics within the resonance peaks of a vowel, and it has been argued that this results in a desirable or pleasing sound.[1]

The vibratos of some famous concert and operatic singers like Caruso have been measured and often found to exceed a semitone, with a rate of 6-8 periods of fluctuation per second.[2] One might expect, therefore, that introducing a vibrato of 3% up and down from the fundamental frequency with a period of ~160 ms could make synthesized vowels sound more lifelike. In fact, imposing such a vibrato on the fundamental frequency of the TASS speech synthesizer[3] has produced human-sounding vowels that more closely resemble singing than speaking.

We decided, therefore, to explore the effect of various vibrato rates and depths for a particular vowel produced by the synthesizer, keeping all other parameters constant. Our general approach was to synthesize a series of stimuli and to obtain listener

judgments of the quality of the sounds. The first five formant frequencies for the vibrato experiment were, respectively, 730, 1150, 2500, 3500, and 4500 Hz, and the vowel had the quality of [a].

Two separate experiments were performed, the first with a uniform vibrato period of 150 ms and depths ranging from zero to 8% deviation in both directions, and the other with depth fixed at ±2% and vibrato period varying from 100 ms to 200 ms for different stimuli, corresponding to rates between 10 and 5 per second, respectively. (A vibrato depth of ±2% was found to produce the highest quality ranking in the depth test.)

Two sets of stimuli were used for each test, having fundamental frequencies of 100 Hz and 200 Hz for the depth test (near the first two G's below middle C on the piano) and 150 Hz and 200 Hz (near D and G) for the rate test. The sounds were all 0.8 s in duration, gradually rising to full amplitude in the first 0.15 s and falling for the last 0.15 s.

In a pilot experiment, an A-B comparison format was used for the depth tests. The tones were presented to listeners in pairs and the subjects chose the one from each pair that sounded more lifelike or as if it were produced by the better singer. Each stimulus was paired with every other stimulus of the same fundamental frequency (and vibrato rate), and each pair was also heard with the order of the two stimuli reversed. The order of presentation of the pairs was random. The A-B test was necessarily lengthy and caused undesirable fatigue in the listeners.

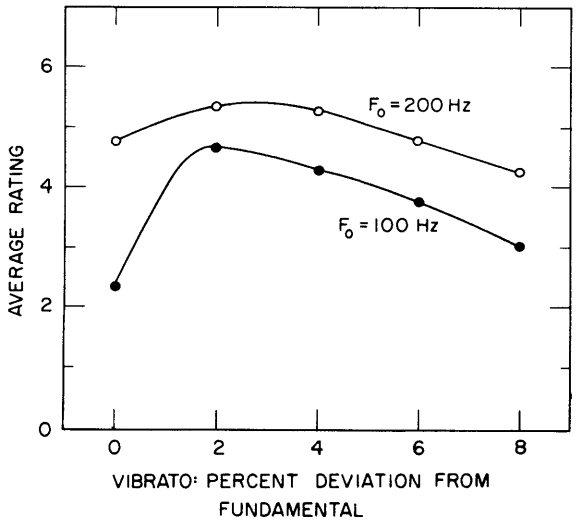


Fig. XIX-13.
Average ratings of listeners for vowels produced with various amounts of vibrato. Ratings are on a scale from 0 (least acceptable) to 9 (most acceptable). Each point represents average ratings based on 11 judgments. Vibrato period, 150 ms.

Another format proved to be more effective, and was followed in subsequent experiments. The stimuli were presented one at a time, spaced at 6-s intervals, and the

subjects were asked to rate the sounds on a 0-9 scale, nine being the closest to good human singing. Each stimulus was heard twice in the course of the test; the order was random. This procedure was also employed on the rate tests. Seven subjects judged the stimuli for the depth tests (with 4 subjects repeating the test to give a total of 11 presentations of the test stimuli), and 3 subjects were used for the rate tests. The subjects in these tests had some musical experience, although it had not been directly with training in singing.

The data from the vibrato tests are displayed in Figs. XIX-13 and XIX-14. Average ratings associated with each stimulus were calculated. As can be seen, the most pleasing vibrato for the subjects tested was found to be a deviation of 2% in each direction at a rate slightly higher than 6 periods per second. It is interesting to note that

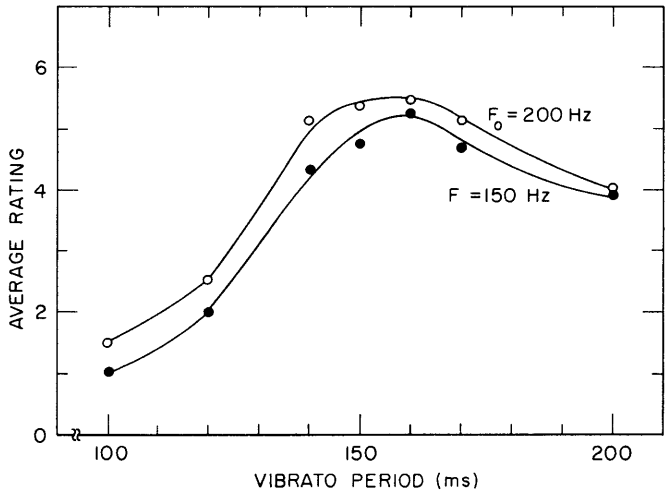


Fig. XIX-14. Average ratings of listeners for vowels produced with various periods of vibrato. Ratings are on a scale from 0 (least acceptable) to 9 (most acceptable). Each point represents average ratings based on 18 judgments. Depth of vibrato is ±2% from fundamental.

in both speed and depth tests, the tones with the higher fundamental frequency were slightly more pleasing than the corresponding ones with a lower fundamental. If this is found to be the case in general and not merely for the particular notes that were used (G, D, and G), one might conclude that this particular synthesizer configuration is more suited to singing [a] in the tenor range than in the bass register.

2. Fourth-Formant Experiment

Researchers speak of a "singing formant," meaning a concentration of energy in the 2.5 kHz to 3 kHz region often found in the voices of accomplished singers. The energy

in this region can be increased by lowering F_4, since then the F_3 and F_4 peaks come closer together. This observation suggested an experiment to investigate the effect of F_4 frequency on the quality of synthesized sung tones. Listeners were presented with synthesized stimuli having different values of F_4 in an attempt to determine an optimal value for a given configuration of the first three formants. A test similar in format to the vibrato tests was prepared, with stimuli of uniform vibrato (±2%, 6 times per second) and the same formant pattern as before, except that F_4 was varied in steps from 2.7 kHz to 4 kHz. The sounds with low fourth formant were noticeably louder than the others, so the levels were adjusted to give equal VU meter readings in preparing the test recording.

Table XIX-1. Ratings by one listener of sung vowels with various fourth-formant frequencies according to "placement of voice in the head." A rating of 8 was farthest forward and 1 was farthest backward in the head. Data are averages of 3 judgments. A rating of 4 was judged to provide the best quality sung vowel.

F_4	Average of Three Ratings
2700 Hz	7.3
2800	6.0
2900	5.7
3000	4.7
3100	4.0
3200	4.0
3300	3.3
3500	3.0
4000	1.5

Making judgments on these stimuli was found to be more difficult than in the vibrato tests, and, as a result, listeners were more carefully selected. Two students of voice and one voice teacher participated in several hours of listening and discussion. We felt that a group effort might yield more meaningful results in a shorter time. Qualities usually considered to be metaphorical by investigators, such as focus, placement, and head resonance were discussed and later proved useful to the singers in rating the stimuli. Each stimulus occurred 3 times in the test; after some practice runs (without feedback as the location of F_4) the subjects began to achieve some degree of consistency from one repetition to the next and between the different occurrences of the same sound. Judgments of a sound were considerably influenced by the immediately preceding sound in the test. There was general agreement, however, that stimuli with F_4 values at the upper and lower ends of the range were less acceptable than those in the range 3000-3300 Hz.

One of the subjects rated the tones from 1 to 8 with regard to voice placement in the head, trying to imagine where a human singer might direct his voice to obtain a sound character similar to each sample of the test. With reasonable consistency, these numbers were inversely related to the position of the fourth formant, with 8 being the most forward placement, corresponding to low values of F_4. This listener said that the best sounds were those in the middle, with those rated 4 being generally superior to those rated 5. Items with ratings higher or lower than 4 or 5 were less acceptable, and those assigned a rating of 6 or 7 were judged to have a nasal quality. The ratings assigned in this fashion by this observer to each of the three occurrences of each stimulus are listed in Table XIX-1, and seem to indicate that the fourth formant is too near the third when it is below 3 kHz, and not close enough when it approaches 4 kHz. This result, of course, is only valid for the particular vowel [a] used in the test, and might be expected to be different for other vowels.

3. Conclusions

This study showed that meaningful judgments could be made of computer-synthesized vowel sounds designed to imitate human singing. Some of the sounds were very machinelike, but others were sufficiently realistic that trained singers could judge them by imagining the physical sensations that a person would feel in singing them. The results suggest that the variables considered (i.e., vibrato of the fundamental frequency and position of the fourth formant) can be adjusted to make isolated sung vowels that are acceptable to people experienced in evaluating singing.

As an incidental feature of this study, the synthesizer was made to sing scales and arpeggios which sounded quite realistic by changing from pitch to pitch, as the human voice does, and fluctuating a few percent at a vibrato rate near 6 per second while sustaining any pitch. The synthesis of such passages could be expanded into another study involving judgments by trained singers.

G. L. Gibian

References

1. P. B. Oncley, "Frequency, Amplitude and Waveform Modulation in the Vocal Vibrato," paper presented at 80th Meeting of the Acoustical Society of America, Houston, Texas, 3-6 November 1970.

2. W. Vennard, Singing: The Mechanism and the Technic (Carl Fischer, Inc., New York, 1967).

3. W. L. Henke, "TASS—Another Terminal Analog Speech Synthesis System," Quarterly Progress Report No. 95, Research Laboratory of Electronics, M.I.T., October 15, 1969, pp. 73-81.

4. J. Sundberg, "Articulatory Differences between Spoken and Sung Vowels in Singers," Speech Transmission Laboratory Quarterly Progress and Status Report, Royal Institute of Technology, Stockholm, Sweden, 15 April 1969, pp. 33-46.