XV. SPEECH COMMUNICATION*

Prof. K. N. Stevens	Dr. A. W. F. Huggins	V. V. Nadezhkin
Prof. M. Halle	Dr. B. E. F. Lindblom†	Y. Kato‡
Prof. J. B. Dennis	Dr. S. E. G. Öhman†	J. A. Rome
Prof. J. M. Heinz	A. M. Advani	R. S. Tomlinson
Dr. A. S. House	Jane B. Arnold	E. C. Whitman
	W. L. Henke	

A. DESIGN OF A DIGITALLY CONTROLLED ATTENUATOR

The development of an improved dynamic vocal-tract analog for speech synthesis has made necessary the design of a wide-range digitally controlled attenuator.[1] This report describes a circuit realization of such an attenuator and discusses the design considerations in some depth. Also included in this report are some experimental data on its performance.

The circuit employs transistor switches with resistive voltage dividers as the basic element of attenuation. Eight stages of attenuation under the control of 8 external binary signals provide from 0 to 63 3/4 db attenuation in 1/4 db steps. Each stage has two states – attenuating and nonattenuating. The attenuation for each stage when on is an integral power of 2 db independently of the state of the other stages. By cascading these sections, the total attenuation is simply the sum (in decibels) of the attenuation of each stage.

A simplified attenuator section is shown in Fig. XV-1. The source impedance is

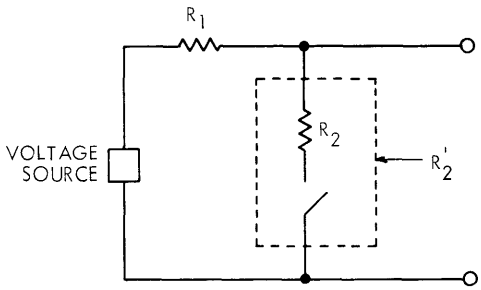


Fig. XV-1. Elementary stage of attenuation.

*This research is supported in part by the U.S. Air Force (Electronic Systems Division) under Contract AF 19(628)-3325; in part by the National Science Foundation (Grant GP-2495); in part by the National Institutes of Health (Grant MH-04737-04 and Grant NB-04332-02); and in part by the National Aeronautics and Space Administration (Grant NsG-496).

†On leave from the Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden.

‡On leave from the Nippon Electric Company Limited, Tokyo, Japan.

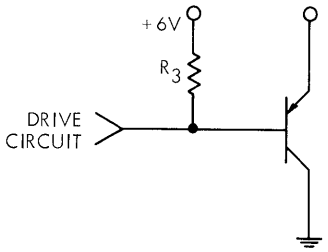


Fig. XV-2. Transistor switch.

zero, and the load impedance is infinite, so the attenuation is $\frac{R_2'}{R_1 + R_2'}$. R_2' is made to vary from some finite value to a nearly infinite value by means of the switch shown. When the switch is open the attenuation is very nearly one (0 db); when closed, the attenuation is very nearly $\frac{R_2}{R_1 + R_2}$.

The switches shown in Fig. XV-1 are very simple transistor switches. An elementary circuit is shown in Fig. XV-2. The transistor is operated in the so-called inverted connection with collector grounded, since this reduces the difference in DC output between the on and off states.[2] When there is no drive, the base is brought to +6 volts through R_3 reverse biasing both junctions. There is then a very high impedance from emitter to ground. This is the off state of the switch. As long as the input voltage does not exceed +6 volts, this mode of operation is applicable, but if it does exceed this, distortion results because the emitter junction becomes forward biased for part of the cycle.

When the drive circuit forward-biases the switch, the transistor saturates. There is then a very low impedance from emitter to ground. This is the on state of the switch. The impedance is low, but not zero, so the attenuation is $\frac{R_2 + R_s}{R_1 + R_2 + R_s}$, where R_s is the saturation resistance of nearly the ideal $\frac{R_2}{R_1 + R_2}$, and variations in attenuation with variations in R_s are minimized. R_s is also nonlinear and produces distortion that is monotonically related to the ratio of signal current to drive current. This ratio should be as small as possible to minimize distortion. This requires a large drive current and/or small signal current.

The noise introduced by the transistor switches is of four kinds. One of these is caused by the discrete nature of the attenuation, and three are caused by circuit deficiencies, The first is simply the modulation products caused by modulating the signal with a step, or a series of steps. The amplitude of these modulation products is directly proportional to the amplitude of the step and the signal amplitude. The 1/4 db steps cause the

modulation products to be inaudible with complex signals such as speech. With simpler signals such as sinusoids, however, they are audible.

The second kind of noise also results from modulation. The cause is different delays in the switching of various stages of the complete attenuator. For example, if when switching from 31 3/4 db to 32 db the 32-db stage is slower or faster than the other stages, the attenuation will momentarily be 0 db or 63 3/4 db. The effect is that of modulating the signal with a rather large narrow pulse. This can best be minimized by keeping all switching delays as small as possible. The major source of delay is the time required for the switching transistor to come out of saturation. This time for a given transistor is directly proportional to the drive current. Therefore the drive current should be minimized to minimize this noise.

The third kind of noise is caused by capacitive coupling of the driving signal from base to emitter of the transistor switch. The area of the noise pulses so generated is equal to the product of the charge that flows from the emitter and R_1. The amount of charge is rather complexly related to the voltage swing at the base, the drive current, and the transition and diffusion capacitances of the emitter junction. Measurements indicate a typical range of 200-500 nano-volt-seconds into 1000 ohms or 200-500 μcoul for 2N1307 transistors. The above-mentioned four items should all be minimized to reduce noise. Cost considerations rule out high-speed transistors, and reducing the base voltage swing requires that the signal swing be correspondingly reduced. Therefore it is necessary to minimize the drive current to minimize this noise.

The fourth kind of noise is caused by a change in DC level at the emitter between the on and off states. This phenomenon is adequately discussed elsewhere.[2] This noise is minimized by minimizing R_1 and minimizing the drive current.

The unity gain buffer amplifier (Fig. XV-3), which is used to isolate successive stages, consists of two class A complementary emitter followers. Although one transistor could have been used, the base-emitter drop must be cancelled. Otherwise the DC output would be chopped up by the following attenuator stages, thereby making considerable noise. It is impossible to cancel the drop for all temperatures, because of its large variation with temperature. By using matched transistors, the base-emitter drops very nearly cancel each other for all temperatures. A drift of less than 1 mv has been achieved over a fairly wide temperature range. Most of this drift is due to differential cooling of the two transistors. To minimize the effect of differential cooling, the total cooling is lowered by limiting the power dissipation to less than 30 mw. In order to do this, the bias current must be lowered. This reduces the maximum signal current allowed.

D-C coupling is necessary to eliminate low-frequency phase shift. Our application uses the attenuator in a high-gain feedback loop, and excessive phase shift could lead to disastrous oscillations. The output impedance of the buffer amplifier is very nearly

$R_4/2$ ohms. This impedance serves as the R_1 indicated in Fig. XV-1.

Most of the considerations discussed above have conflicting circuit requirements. For example, to minimize the capacitively coupled noise requires a small R_1 and small drive current. The first requires high bias currents in the emitter followers, with the result that there is greater drift. It also causes greater signal currents and more distortion. The second does not help any because reducing the drive current will also increase distortion. The final circuit shown in Fig. XV-4 represents an effective compromise of these conflicting requirements.

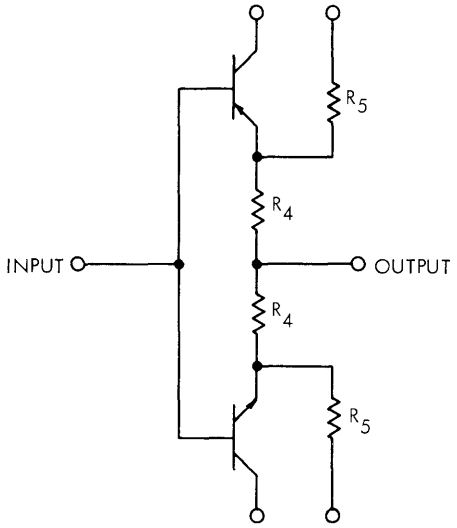
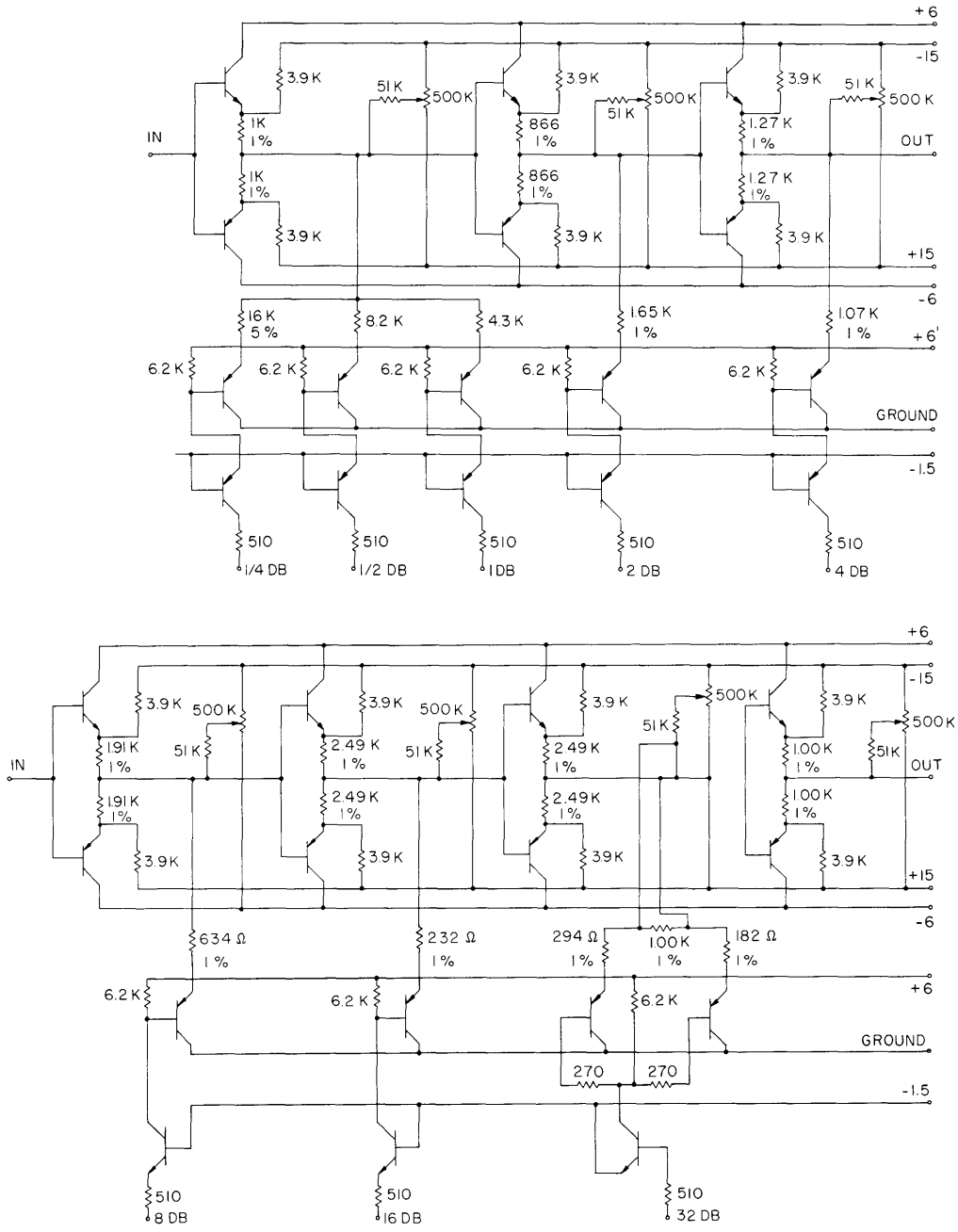


Fig. XV-3. Buffer amplifier.

The complete circuit is shown in Fig. XV-4. Its operation should be fairly obvious from the preceding discussion. The first three stages operate with no intervening buffer amplifiers, since the source impedance for one stage is not significantly altered by the state of the other stages. The 32-db stage consists of 2 sections. This is done to maintain a reasonable size resistance in the lower arm of the divider. This improves the accuracy of the stage and reduces distortion.

One transistor per stage is used to drive the switches. This transistor converts logical levels of 0 and -3 volts to appropriate driving signals. The first 7 stages have a grounded-base configuration to allow a gain-independent drive without requiring clamping or saturation of the driving transistor. The last stage must drive two transistors and is therefore operated in a grounded-emitter configuration. This requires a control signal of the opposite sense for the last stage, but this is usually available from the external equipment.

The variable resistors associated with each stage permit slight variations in DC



NOTES:- ALL TRANSISTORS PNP - 2N 1307
NPN - 2N 1306

ALL RESISTORS 5% 1/2 WATT UNLESS OTHERWISE NOTED

ALL 1% RESISTORS 1/4 WATT

TRANSISTORS IN EMITTER FOLLOWER PAIRS SHOULD HAVE V_{BE} MATCHED WITHIN #3MV AT 4MA AND MOUNTED TO MAINTAIN BOTH TRANSISTORS AT NEARLY THE SAME TEMPERATURE

NOMINAL MAXIMUM SIGNAL LEVEL 5V PEAK (+ OR -)

APPROXIMATE POWER DISSIPATION 14 WATTS

Fig. XV-4. Complete digital attenuator. Output of upper block is to be connected to input of lower block.

level to be cancelled out. This allows the fourth type of noise mentioned above to be reduced.

The following experimental data were measured on a prototype of this attenuator.

Phase shift at 20 kc, -4°

Harmonic distortion at 5.5 volts peak input, 0.05 per cent

S/N at maximum signal and any average attenuation, >40 db

Maximum total drift at output from 20°C to 30°C ambient, 15 mv

Input impedance, >180 KΩ

Output impedance, 510 Ω

Minimum attenuation, 0.29 db.

R. S. Tomlinson

References

1. J. B. Dennis, Speech synthesis, Quarterly Progress Report No. 67, Research Laboratory of Electronics, M.I.T., October 15, 1962, pp. 157-162.

2. C. L. Searle, A. R. Boothroyd, E. J. Angelo, Jr., and D. O. Pederson, SEEC Notes I, Elementary Circuit Properties of Transistors (John Wiley and Sons, Inc., New York, 1962), Chapter 10.

B. AN ARTIFICIAL PALATE FOR CONTINUOUS ANALYSIS OF SPEECH

An artificial palate incorporating electrical contact-sensing elements has been designed and constructed. The research problem was to develop a device capable of sensing the regions of lingua-palatal occlusion as a function of time during continuous speech. In many respects the device is similar to that described by Kuzmin.[1]

The palate itself is of conventional structure consisting of a thin plastic material formed to adhere closely to the palate of the subject. Preliminary to its construction, a cast of the subject's upper dental arch and palate was made. Eighteen metal contacts (diameter, 1/16 inch) are imbedded at various locations throughout one lateral half of the artificial palate, and each electrode is connected to a separate fine insulated wire that leads out of the corner of the mouth to the recording apparatus. The other lateral half of the artificial palate is effectively short-circuited to the tongue by means of a single exposed wire that traverses back and forth across its surface.

When the tongue touches a given contact on the artificial palate a signal from an oscillator is connected to an amplifier and relay driver, and a relay is closed. This relay connects the output of an audio oscillator to one channel of a tape recorder. A different relay and an oscillator of a different frequency are associated with each contact, and the oscillator outputs are summed. Thus the pattern of recorded tones represents the temporal pattern of lingua-palatal occlusion for the electrode locations. The

second channel of the tape recorder is used to store the speech signal, and a sound spectrogram is made of the combined two-channel output of the tape recorder.

A preliminary experiment was performed with the artificial palate to show how the points of contact for a postdental consonant in intervocalic position are influenced by the vowel environment. The results are in general agreement with descriptions of coarticulation derived from standard palatographic and acoustical techniques, and indicate that the present method is capable of providing quantitative data on speech articulation in a relatively rapid and convenient form.

This study is reported in greater detail in a thesis submitted to the Department of Electrical Engineering, M.I.T., in partial fulfillment of the requirements for the degree of Bachelor of Science, May 1964.

J. A. Rome

References

1. Y. I. Kuzmin, Paper G35, Proc. Fourth International Congress on Acoustics, Copenhagen, 1962.

C. PICK-UP OF THROAT-WALL VIBRATIONS FOR THE SYNTHESIS OF SPEECH

An investigation has been made of a relatively new technique for obtaining information concerning the human glottal excitation, reported previously by Sugimoto and Hiki[1] and by Porter.[2] Vibrations are picked up at the wall of the throat and this signal is processed to obtain "pitch" information in the form of a pulse during every fundamental time period. These pulses are then fed into a "buzz" generator where they produce a waveform simulating the glottal excitation pulses for use as excitation for a speech synthesizer.

Interesting features of this study include the use of a compact ceramic transducer to pick up the signal and a simple, yet fairly effective, pitch extractor. Reliability of the total system is acceptable though not absolute. The major obstacle is the difficulty in maintaining good coupling between the transducer and the throat wall, with the result that the amplitude of the signal picked up varies considerably as the position of the transducer shifts. This level variation can in turn cause the output to miss occasional pulses if the amplitude of the input falls below a certain level. Also, the device occasionally produces spurious pulses on transient inputs.

This study is reported in greater detail in a thesis submitted to the Department of Electrical Engineering, M.I.T., in partial fulfillment of the requirements for the degree of Bachelor of Science, May 1964.

A. M. Advani

References

1. T. Sugimoto and S. Hiki, On the Extraction of the Pitch Signal Using the Body-Wall Vibration at the Throat of the Talker, Proc. Speech Communication Seminar, Stockholm, Vol. 1, C-10, 1962.

2. H. C. Porter, Extraction of Pitch from the Trachea, Research Note AFCRL-63-24, Air Force Cambridge Research Laboratories, Bedford, Massachusetts, 1963.

D. EFFECTS OF CONTEXT IN AUTOMATIC SPEECH RECOGNITION

A study has been carried out to examine the importance of using contexual information in automatic speech recognition. Two points were considered: first, improvement in accuracy of identification of vowels when information about their context is used; second, whether the positions of the formants during the transition between a consonant and a vowel can be used to identify the consonant reliably.

The syllables studied consisted of isolated CV combinations of the six consonants /b/, /d/, /g/, /ℓ/, /r/, and /w/ with the four vowels /i/, /ɪ/, /ɛ/, /æ/. The recognition system was realized on a digital computer. The accuracy of vowel recognition without contextual information for one talker was 83 per cent, and approximately 20 per cent of the errors were corrected when contextual information was used. The consonants were correctly identified 79 per cent of the time.

This study is discussed in greater detail in a thesis submitted to the Department of Electrical Engineering, M.I.T., in partial fulfillment of the requirements for the degree of Bachelor of Science, May 1964.

O. Philbrick

E. DERIVATION OF AREA FUNCTIONS AND ACOUSTIC SPECTRA FROM CINERADIOGRAPHIC FILMS OF SPEECH

In the study of speech events investigators have been interested in formulating descriptions both at the acoustical level and at the articulatory level, and in interpreting observations at one of these levels in terms of descriptions at the other. At the articulatory level, the type of description that bears a direct mathematical relation to the acoustic output is a specification of the cross-section area of the vocal tract at each point along its length – the so-called area function. The description of articulatory activity in terms of area functions is inconvenient, however, for at least two reasons: (i) area functions cannot be measured directly from data obtained by x-ray, photographic or other techniques of articulatory phonetics; (ii) a model of speech generation that takes into account the dynamic aspects of the process presumably can be described better in

terms of the shapes and motions of the structures rather than in terms of cross-section areas of a tube of air. Thus if one is interested, say, in the problem of devising a device or a computer program that generates speech from a phonetic input, one might wish to depict at some point within the computer a description of the motions of the articulatory structures. As part of the computer program one would need, therefore, a specification of the transformations from positions of the articulatory structures to area functions.

The purpose of this investigation is to examine the possibility of writing rules for converting a description of the vocal tract in the midsagittal plane as observed from tracings of x-ray pictures into a specification of the cross-section area of the vocal tract at each point along its length. There are two aspects to this task: the first is the problem of specifying the midsagittal configuration in some standardized form, and the second is the evaluation of the area function from measurements made on this standardized representation.

We shall examine first the specification of the midsagittal configuration. Figure XV-5 shows a midsagittal tracing made from one frame of a cineradiographic film

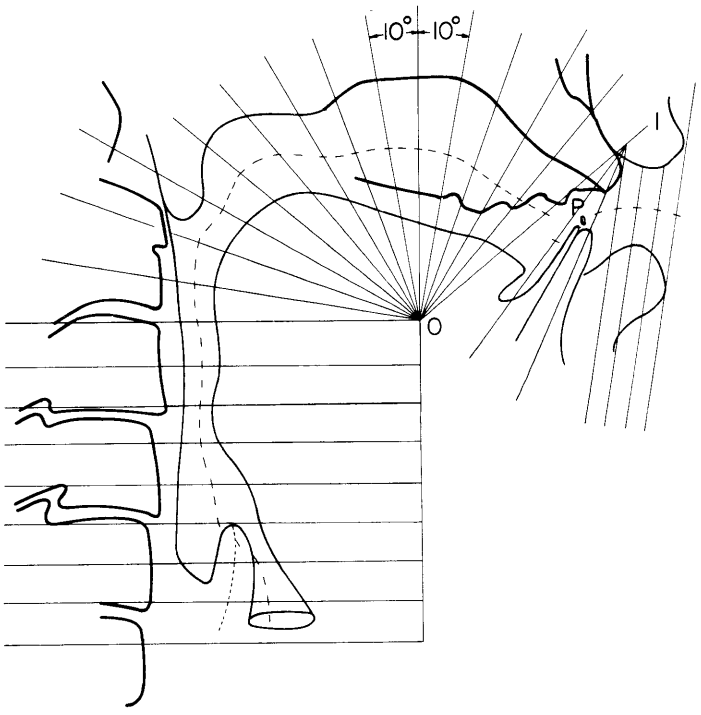


Fig. XV-5. Midsagittal tracing from cineradiographic film taken during the production of the vowel [ɑ]. A grid for the measurement of cross dimensions is shown superimposed on the tracing.

during the production of the vowel [ɑ]. Superimposed on this x-ray tracing is a grid of lines which is defined by a set of operations that can be described as follows. First, a point 0 is defined as the center of a circle that is tangent to the horizontal portion of the hard palate and to the posterior wall of the pharynx. This point appears to be capable of specification in a reasonably unambiguous way, at least for three subjects whose x-ray tracings have been examined. When this point has been located, a series of radial lines intersecting the hard palate, the velum and the upper pharynx are drawn as shown in Fig. XV-5. Then a series of horizontal lines are drawn with approximately 0.3 cm spacing throughout the pharyngeal and laryngeal regions. Only every other line is shown in the oral and pharyngeal regions. Similar procedures are used to define a series of reference lines in the lip region as shown. The entire grid of reference lines bears a fixed relation to the fixed structures such as the hard palate and upper teeth.

When this grid of lines has been superimposed on a given midsagittal tracing of the vocal tract, boundaries are marked. Points midway between these intersection points are then determined, and these are assumed to define the midline of the vocal tract.

Having specified the midsagittal configuration, we now turn to the task of finding the cross-section area in the lateral planes defined by each of these reference lines. The point at which the vocal tract is terminated at the anterior end is established by examining the position of a lead pellet that was attached to a corner of the mouth when the x-ray pictures were taken. It is assumed that the vocal-tract termination is at the reference plane immediately anterior to this pellet, which is labeled P in Fig. XV-5.

For most unrounded configurations for the one talker that was examined in detail in this study the termination is at about the same point as that of the [ɑ] configuration in Fig. XV-5, and thus the lip opening plays no role in the area function determination. When the lips are rounded, however, the termination is at a more anterior plane, and in the lip region the opening is assumed to be circular.

For sections that intersect the hard palate and upper teeth, plaster casts of the upper and lower palates are made in order to find the cross-section shapes. The procedure that was used, together with the assumptions that were made for this region, is shown in Fig. XV-6. Shown at the left-hand side of the figure is a section through the hard palate and the mandible, corresponding to one of the slices indicated in Fig. XV-5. The location of the midline of the tongue is determined from the tracing and is marked on the section. In order to obtain an approximation to the cross-section area, the tongue surface is assumed to be flat, and the shaded area is measured for various values of the cross dimension d, that is, for various positions of the tongue. The inner surface of the cheeks is assumed to be flush with the teeth as indicated by the dashed lines at the left. A typical curve of cross-section area as a function of cross dimension d is shown at the right of Fig. XV-6. The area will, of course, depend slightly on the mandible position, and this position can be considered as a parameter that must be specified.

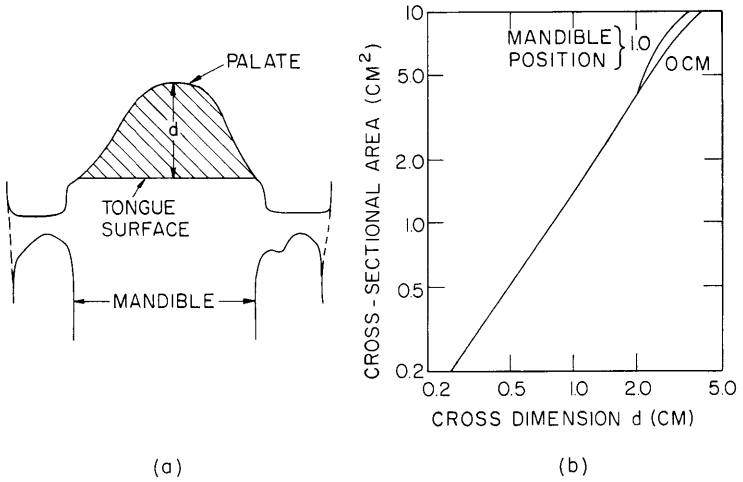


Fig. XV-6. (a) Typical lateral cross section passing through the hard palate. (b) Plot relating cross-section area to cross dimension for the section shown in (a).

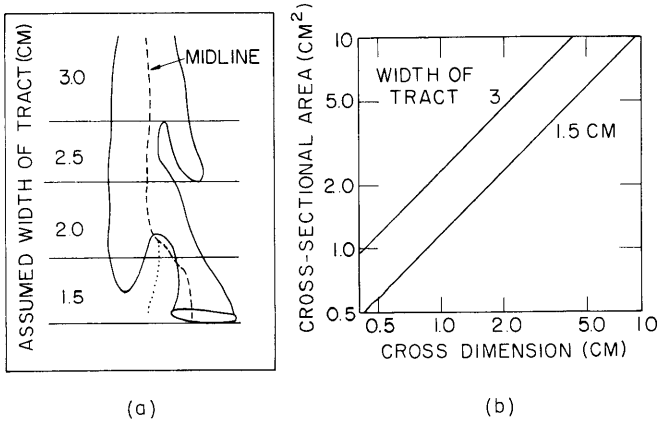


Fig. XV-7. (a) Typical midsagittal tracing in the pharynx and larynx region. (b) Plot relating cross-section area to cross dimension for the upper and lower portions of the tracing in (a).

A different curve is obtained for each reference plane.

For the pharynx and larynx region a somewhat different procedure is used to estimate cross-section areas from cross dimensions, as shown in Fig. XV-7. The width of the pharynx at a given level is assumed to remain the same, independently of tongue position, and the cross section is considered elliptical. It is assumed also that the pharynx width decreases gradually in the manner shown as we proceed from the upper pharynx to the glottis. Evidence for these assumptions comes from study of anatomical atlases, and from discussions with medical specialists. We obtain linear relations based on these assumptions between cross-section area and cross dimension, which depend on the vocal tract width, as shown at the right in Fig. XV-7.

Once the area in each of the specified planes has been determined, it is necessary to make certain corrections. First, since each plane is not necessarily normal to the midline, we have applied a correction by multiplying the computed area by the cosine of the appropriate angle in order to obtain a true area. Second, for regions in which the area change is abrupt – at the teeth for many configurations – corrections for end effects have been used to modify the effective area.

Using the procedures just discussed, we have derived area functions from x-ray tracings corresponding to stressed vowels in a number of different words and for configurations in consonantal regions of several words. In order to test the validity of these derived area functions, the first three formant frequencies were computed for each of the configurations. The computations were carried out by using a digital computer program that obtains solutions to Webster's wave equation. For comparison, the formant frequencies for these samples were also obtained from the sound recordings with an automatic formant-tracking program. A comparison of the formant frequencies obtained by the two methods for seven vowels is shown in Fig. XV-8. Close agreement is found in the first two formants calculated from the tracing (open circles), as compared with those measured from the sound recording (filled circles). The maximum deviation is 40 cps for the first formant, and 100 cps for the second formant. The third formant frequencies also show fairly good agreement except for the vowel [ʊ]. The maximum deviation for the third formant frequency is 150 cps, except for [ʊ], for which the deviation is 460 cps.

Similar calculations have been carried out for samples taken just before closure and just subsequent to release for several stop consonants. Formant frequency comparisons taken from ['hi'gɑ] are shown in Fig. XV-8. The initial portion is indicated by the presubscript i, and the final portion by the postsubscript ɑ. Here the agreement is not quite as good, especially for the first formant in the final portion of the [g]. In this case, however, the calculated formant frequencies are probably more accurate, since it was very difficult to determine the formant frequencies precisely from the acoustic spectra.

In order to obtain an indication of the accuracy needed in the specification of

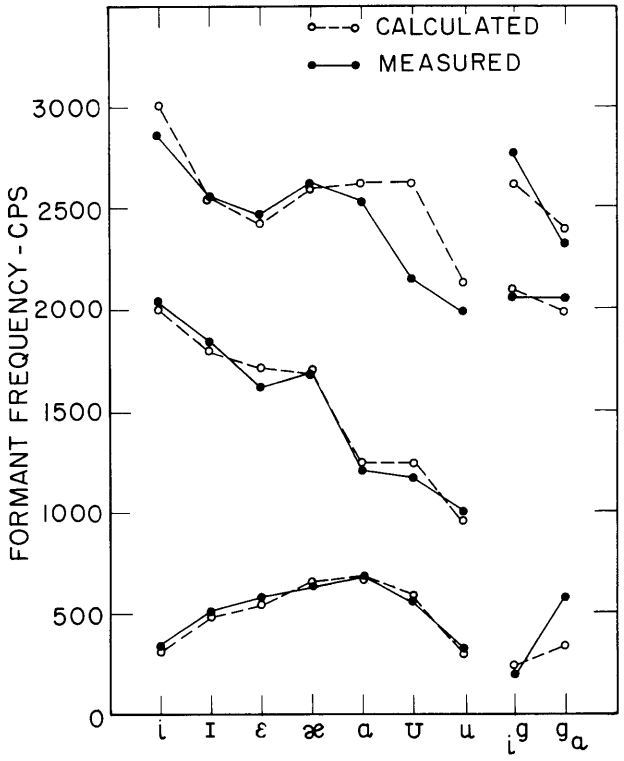


Fig. XV-8. Comparisons of the first three formant frequencies for several vowels and consonants as determined from acoustic spectra (filled circles), and from calculations performed on the x-ray tracings (open circles).

cross-section areas, an area function corresponding to the vowel [ɪ] was subjected to a systematic perturbation. Beginning at the glottis, the area was decreased 22 per cent over a 2-cm length and the formant frequencies were recomputed. The procedure was repeated, perturbing each 2-cm length in sequence. The results are shown in the left half of Fig. XV-9. The shifts in F_1, F_2 and F_3 are plotted against the location of the perturbed portion of the tract. The maximum shift for all three formants is approximately 4 per cent of the unperturbed values. Comparison of these perturbation curves with the volume velocity distribution curves for each formant, shown at the right in Fig. XV-9, indicates that as the perturbation moves along the tract the curves of formant-frequency shift oscillate at twice the frequency of the oscillation of the corresponding velocity distribution. These features can be predicted from considerations that are based on the fact that the average potential energy (proportional to the square of pressure) and the average kinetic energy (proportional to the square of velocity) in the tract must be equal at resonance.

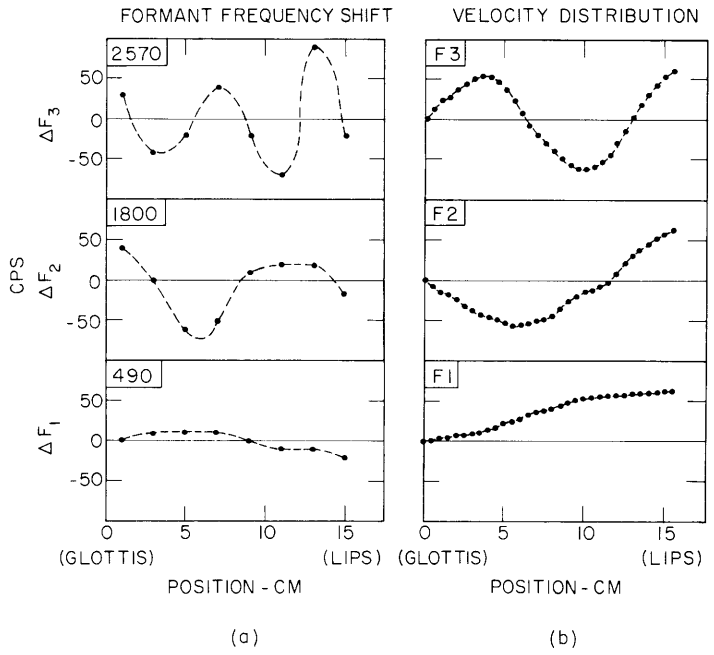


Fig. XV-9. (a) Perturbations in the first three formant frequencies resulting from a 22 per cent decrease in cross-section area over a 2-cm length of the tract centered at locations given by values of the abscissa. (b) Calculated velocity distribution along the unperturbed tract at the frequencies of the first three formants.

Data of the kind just discussed indicate that the first three formant frequencies are not very sensitive to small errors in the determination of the cross-section area function. This finding is encouraging, since it means that many of the small detailed irregularities in the vocal-tract shape may be neglected without seriously affecting the resulting spectrum. Nevertheless, we still have to make a detailed study of the effects of certain assumptions that we have made, such as a flat tongue surface and no cheek cavities.

To summarize, the results that we have obtained for a number of vocal-tract configurations for one talker give us confidence that it is possible to develop simple rules of the kind that we have shown which would specify the transformation from midsagittal dimensions to area functions. Some modifications in the rules will be necessary for certain configurations such as [ℓ], [r] and nasals. Also, an examination of the extent to which the rules for a given talker can be generalized to other talkers without making detailed measurements of the anatomy of each talker has yet to be carried out.

J. M. Heinz, K. N. Stevens