CONSTRUCTIONAL WORK ON A FUNCTION GENERATOR FOR SPEECH SYNTHESIS

Jørgen Rischel

1. General status of speech synthesizer project.

As mentioned briefly in last year's Annual Report (6), a formant-coded speech synthesizer is being constructed at the Institute. The apparatus consists of two parts: a function generator supplying the control voltages for synthesis of connected speech and a sound producing system comprising voice source, noise source, formant filters, and variable gain amplifiers associated with the formant filters.

As for voice and noise sources, our present devices exhibit no special features.

The formant filters for vowel synthesis have been designed along the same lines as the circuits employed in a provisional vowel synthesizer built by this author some years ago (5). The filters are high-Q LC series circuits in which part of the capacitance is represented by capacitance diodes (type BA 163), and the resonant frequencies can be varied simply by varying the bias voltage applied to the diodes. A sufficient range of frequency variation for the individual formants is obtained by letting the circuits resonate at rather high frequencies, the frequency spectrum of the voice source pulses being shifted correspondingly by a heterodyning process. This, of course, involves the use of a modulating and a demodulating system (common to all formant circuits). In order to obtain good amplitude and phase linearity down to low frequencies we have designed a transformerless double-balanced modulator, which seems well suited for the purpose.

The new version of this whole system is not yet ready for use, so it would be entirely premature to give any circuit details. (Our "phonetic" experience is limited to the old model.)

Similar circuits (but including antiformant filters) have been planned for nasals and stops/fricatives.

The use of frequency transposition in the formant filter system necessitates a parallel connection of the filters (with phase relationships taken care of). This means that the level of each formant can and must be controlled independently by the experimenter. Our newly constructed formant circuits include voltage controlled attenuators exhibiting a linear relationship between control voltage and output level over a 30 dB range. The freedom of control thus offered is crucial for some kinds of experiments, but there may certainly be other cases in which the experimenter wishes formant levels to be derived automatically from the pattern of formant frequencies. This can perhaps be done by means of nonlinear circuits, possibly a "potential field"(7). However, since the formant filter system and the function generator are functionally independent units it would also be possible to use alternative versions of each (e.g. a series connexion of filters can be added eventually) and to combine these freely according to the immediate purpose. This may be a simpler solution.

In the past year a major part of our efforts within the synthesis project have been devoted to the development of an expedient function generator, which can store an arbitrary configuration of synchronously varying control parameters and, on request, reproduce a corresponding set of time-dependent voltages. The part of the generator that functions by now can deliver eight independent voltage functions which can be used to specify a sound sequence lasting not more than 2.0 seconds (this maximum duration is possible only with a rather coarse resolution in time). The number of parameters is going to be increased (this is simply a matter of adding strings of potentiometers in parallel with those already present), whereas we have no immediate intention of increasing the capacity in the time dimension very much.

Various considerations have influenced our choice of type of function generator, such as: a desire to avoid demanding mechanical constructions, demands for accuracy and good reproducibility, and a wish to be able to synthesize sound sequences at different information rates ("fine surgery" and "synthesis by rule" being the extremes).

The function generators most widely used with analog synthesizers are in a sense simple transducers: they convert an exact graphical representation of each parameter (mostly in the form

of a painted trace on a sheet) into an electrical control signal, and ideally, the traces painted by the experimenter should be exact replica of the formant movements etc. to be synthesized. Our function generator belong to another category of devices which represent the configuration of parameters as a sequence of states each lasting for a preset amount of time. In this case the experimenter must somehow specify the parameter values once per temporal segment (the duration of each of these may, of course, be quite short if a good definition of transient phenomena is the research objective). The output voltages will thus be varying stepwise. In the following the function generator (taken in a restricted sense) producing these step functions will be referred to as the "staircase voltage generator".

The most obvious problem in using a staircase voltage generator for analog synthesis control is that the discontinuities (jumps from one voltage to another) must somehow be smoothed out in order not to produce transient bursts in the sound producing system. This smoothing out is in fact a very crucial feature of the system generating the control voltages, since the choice of smoothing function may influence the specification of the input information to the function generator considerably. The problems which have presented themselves to us will be dealt with in some detail in a later part of this paper.

Our staircase voltage generator was designed independently of other similar devices in current or earlier use (a leading principle being to avoid mechanically functioning parts), but it must be stated that the type as such is not at all new (1). Each parameter is represented by a series of potentiometers (one per time segment) whose scales can be calibrated directly in cps or decibels depending on the parameter. The staircase voltage is produced by setting each potentiometer to the appropriate voltage and scanning the whole series by means of a multiplexing system. This approach is supposed to be favourable in the case of experiments in which relatively short sequences are to be synthesized and altered systematically in respect of one or several parameters (typically perception tests). Changing the setting of a knob is likely to be a more agreeable job than deleting a trace and painting a different trace. The difference is most

obvious the moment the duration of a single portion of the sound sequence is to be changed. With our function generator this can be accomplished within wide limits by means of a switch, "duration" being a separate parameter of the system.

Electronically, the function generator has been designed in such a way that a fairly high degree of stability is obtained. The multiplexing system is a sequence of gates each of which triggers the next after having been in the "ON" state, the durations of the individual states being determined by adjustable CR circuits. Each section includes a Schmitt trigger ensuring well defined switching times and "ON" voltages (the rise time of the gates is considerably better than 50 microseconds, which is sufficiently short for the purpose). The actual multiplexing is obtained by alternative biasing of diodes as scetched in Fig. 1.

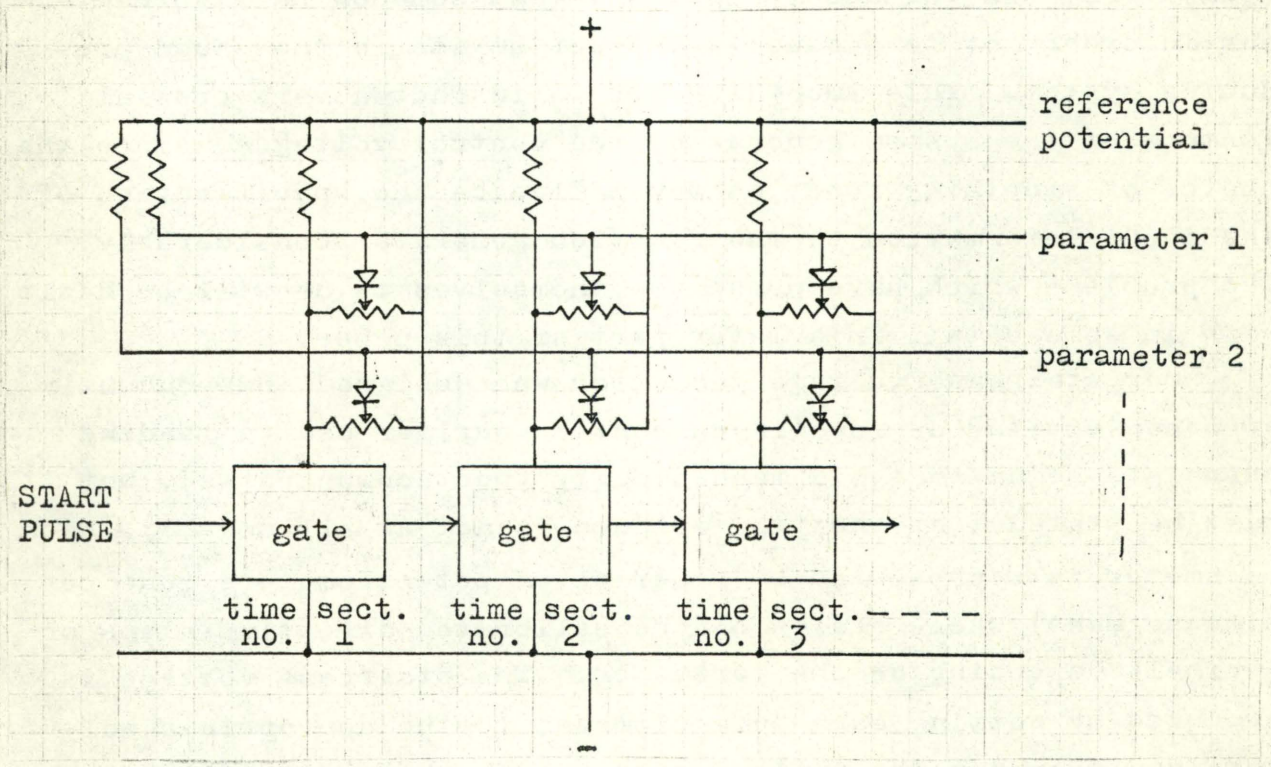


Fig. 1.

It may be argued that a function generator of this kind copies some of the properties of a digital generator (either in the form of a flexowriter or in the form of a small computer) without sharing other of its virtues, such as complete tempera-

ture stability,*) extreme versatility and storage capacity, and appropriateness for other purposes than synthesis. It should be noted, however, that the analog function generator is in several respects preferable to a generator working from a punched tape (flexowriter). It shares with the on line operated computer such features as immediate feedback and easy manipulation of stored data. Conceptually, the manual setting of parameters is an extremely simple matter, so the system seems attractive also for teaching purposes. And even if the system may eventually become obsolete with a complete change of emphasis from analog to digital control methods even in work on a small scale, we feel confident that experience with it will furnish a very useful background for later computer-oriented work.

A direct comparison between analog and digital control of the same synthesizer would probably not show any considerable difference in the time it takes to synthesize a short sound sequence by the two methods, provided that the segment durations and the smoothing function (see below) are optimal.

2. Quantizing control signals in time.

When planning our function generator we considered different possibilities of performing a quantization in time, the only prerequisite being that for practical reasons all parameters must be specified synchronously (even though the specification often consists in a repetition of the preceding value for one or several parameters). Although only one of these was deemed really acceptable they will all be described briefly in order to make the point clear:

1) One possibility is to specify parameter values a fixed number of times per second. Fant (3a) has obtained good quality of synthetic speech with a sampling rate of some 40 per second.

---

*) As for temperature-dependent drift the superiority of digital systems is most obvious if the formant circuits are controlled digitally as in OVE III of the Speech Transmission Laboratory in Stockholm (4). If the digital control signals must be converted into analog signals before they can be applied to the circuits some of the calibration problems must return. Nevertheless, this may be a compromise worthy of consideration.

This approach is of interest in analysis-synthesis techniques, but for our purposes is must be immediately discarded because the number of potentiometer settings will be prohibitively high.

2) A more favourable solution (which forms part of our present "strategy") is to make the duration of each segment independently variable. This means that the "specification rate" can be made quite high in the transient parts of the speech-wave but considerably lower in parts where only slow changes occur. However, since the smoothing function must be chosen in accordance with the highest specification rate used the smoothing will be imperfect in parts with a low specification rate. Since the aberrations from the true parameter slopes resulting from imperfect smoothing can only be tolerated to a certain degree, this problem sets a limit to the gain in economy that can be obtained by varying the specification rate (with the exception of obvious cases like silent intervals).

3) A considerable gain in economy is obtained if the transition rate of the parameters can be reflected by corresponding values of the smoothing function. Ideally this would mean that sequences like ba, wa, ua could be represented by essentially the same stored configuration of parameter values, the differences in transition rates being brought about by varying the time constant of the smoothing circuit. - If only a single CV combination is being synthesized, such switching poses no problem, but in longer sequences the transition rate may change from one part of a syllable to the next. Thus a full exploitation of this method requires that "transient time" be included as a parameter along with "segment duration". The strategy then approaches the concept of "synthesis by rule", since a sequence of two sounds will be represented by two sets of target values (valid for the particular combination), two segment durations, and a transient time (being the time elapsing from the theoretical beginning of the second segment until its target value is reached) which is conditioned by the type of sound combination, but specified independently by the experimenter.

4) Finally it is possible to design the smoothing circuit in such a way that a greater or lesser "overshoot" in the voltage step function speeds up the transition from one target value to the next. It is thus possible to vary the transient

times of the individual parameters quite independently. The drawback of this method is that the possibility of calibrating the potentiometer scales is lost, since an increment in the setting of a potentiometer may function both to speed up a transition and to increase the target value of the parameter at this point. - This loss of a simple relation between potentiometer setting and parameter value is a reality if ordinary integrators are used as (the first stage of) smoothing circuits (cp. below).

Solution "3" above (i.e. parametric control of transient time) has seemed to us the most attractive solution, although it does not remove all inconveniences stemming from the quantization procedure. Technically, the design of electronically adjustable smoothing circuits has turned out to pose very difficult problems, and we have obtained acceptable results only recently. There are two sources of difficulties: one is that the exact shape of the step response of the smoothing circuit becomes quite crucial when slowly varying phenomena have to be generated from step functions, the other is that there must be no interference at all between the switching voltage controlling the transient time and the voltage function to be smoothed out. It is no overstatement that the circuits we are experimenting with are unduly complicated. Nevertheless we consider the implications of the approach so interesting that we are going to build such smoothing circuits for at least some of the parameters.

Although circuit details will not be dealt with at all in this report, the principal questions raised in connexion with the design can be approached from the more general point of view of "synthesis strategy". The major points which we have had to consider will be presented below.

3. Transition shaping circuits.

The simplest way to convert a voltage step into a raising or falling slope is to use an RC circuit. The slopes generated by such a circuit are, however, exponential. Spectrograms of real speech reveal that the formant transitions in consonant-vowel sequences can often be matched very well by means of such exponential slopes, but in vowel-consonant sequences the agreement will often be rather poor. The differences will be clear

from the stylized curves shown in Fig. 2:

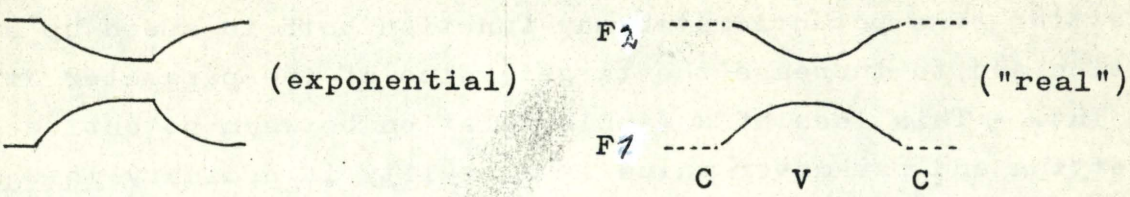


Fig. 2.

A particular inconvenience arising with the use of RC circuits is that theoretically the target values are never reached. The transient time must be defined as the time it takes for the slope to reach 63 per cent of its final value (or rather 63 per cent of the difference between the previous and present target values), but the control voltage occurring at this or any other point in time will never be equal to the calibrated value read off the function generator. With very long time constants this may be a problem.

An oblique slope can also be obtained from a voltage step function by using a lowpass filter with good phase characteristics. This is in fact a very attractive solution, but if very different transient times are to be used in the same piece of synthesized speech difficulties may arise, both because of the switching problems and because of time lag phenomena. The inconvenience caused by the variation of delay time with filter cutoff frequency may be serious if an exact control of the temporal relationships among the events is required.

Conceptually, the simplest solution is to produce a linear slope going obliquely from the point where one time segment ends and reaching the target value of the next segment sooner or later depending on the transient time chosen. Such a linear slope is not an ideal approximation to a formant transition, for example, but it is a simple and well-defined function to work with. If an extreme degree of fidelity is required, it is perfectly possible to generate a formant transition by combining two or several such linear slopes to give the desired curvature, but there is probably seldom any reason to do so. - The discontinuities present at the points where the slope starts and ends can be

removed by simple means if this is considered essential. In our experimental set-up the slope-shaping circuit is terminated by a buffer amplifier, which by the addition of a few components can be changed into a simple (active) lowpass filter. If a sufficiently high cutoff frequency is chosen for this filter, it will not have any appreciable influence on the shape of the control voltage passing through it except that disturbing discontinuities are removed. The oscillogram in Fig. 3 below is shown without this additional filtering in order to exhibit the linear slope more clearly, whereas the oscillograms shown later in this paper are representative of the voltages after filtering.

Theoretically the simplest way to produce such oblique slopes is by linear interpolation between values specified at discrete points in time. This approach has been employed in the computer control of OVE III in Stockholm.

We have not investigated into the possibilities of constructing an expedient "interpolator" by analog methods. Our considerations were based on the assumption that the input function must be a step function and that we wish to change this step function into a ramp function with the same minimum and maximum values and with a duration of the oblique slope which can be controlled externally.

According to this method every constant-voltage segment of the input function (the voltage delivered by the staircase voltage generator) is replaced by a succession of an oblique and a horizontal portion.

Phonetically it is implied that a transition is always counted with the <u>following</u> segment. If, for example, the bottom curve of Fig. 3 (next page) is taken as a crude approximation to the first formant movement in a sequence like <u>dad</u>, it is seen that this CVC sequence can be synthesized by means of three segments. Each of these comprises a transition plus a following steady portion (the latter may, of course, be absent if the transition is made equal in duration to the whole segment) but it is specified by the experimenter in terms of three simultaneous parameter values: a target value for the formant in question, a duration of the transition to this target value, and a duration of the total segment.

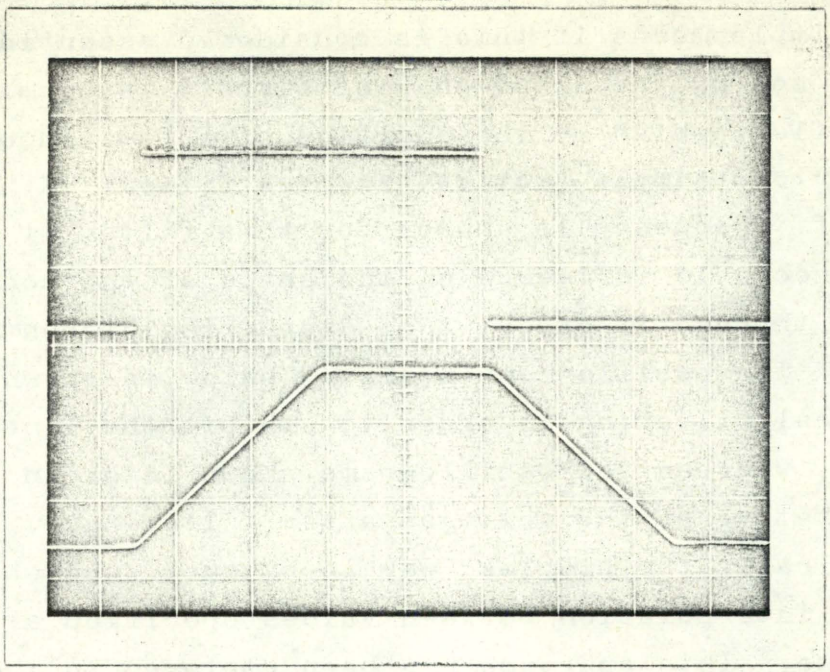


Fig. 3.

Trapezoidal voltage functions like the one shown in Fig. 3 can, of course, be produced by integrating a succession of positive and negative square pulses, cp. Fig. 4, which shows the input to and output from an operational amplifier coupled as an integrator:

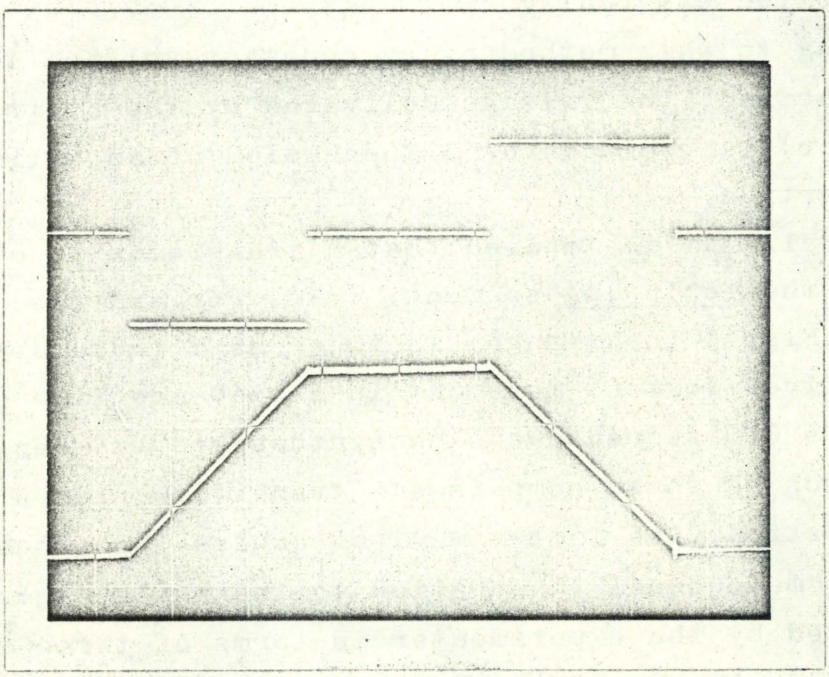


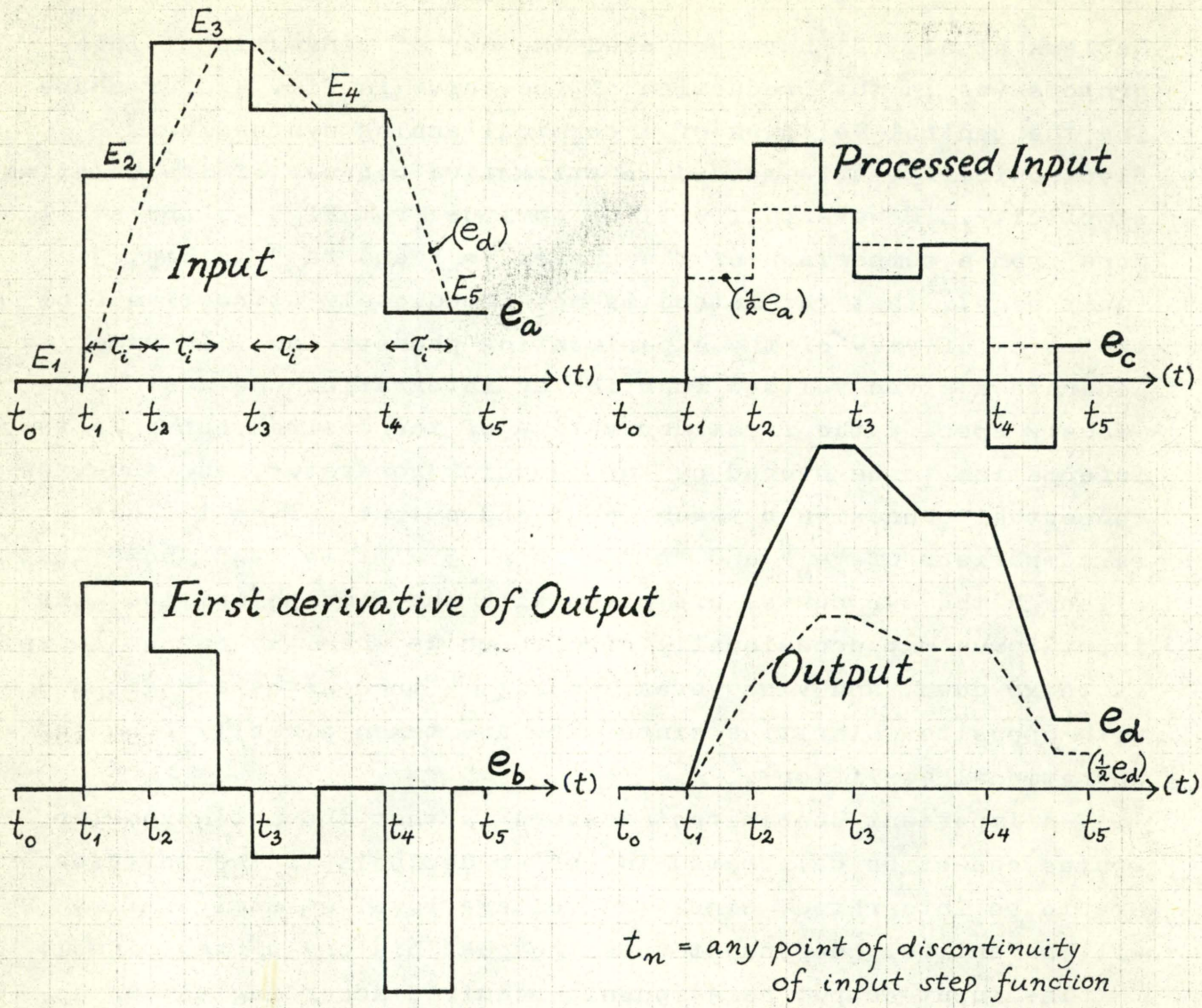
Fig. 4.

DeClerk et al. (2) have proposed the use of conventional integrators (as in the production of the curve in Fig. 4) for shaping the control voltages of a terminal analog synthesizer. Electronically, this method is attractive because of its relative simplicity. However, it will be obvious from Fig. 4, and still more from a comparison of the curves "e_b" and "e_d" in Fig. 5 (next page), that the method is not immediately attractive from the point of view of the experimenting phonetician. In order to generate a given voltage function by integration one must obviously specify the first derivative of the desired curve as the information to be stored by the function generator, and thus the conceptual connexion between input and output curves is lost (a comparison of "e_b" and "e_d" of Fig. 5 at time "t_2" shows that although the two curves are given with the same polarities, the input curve may occasionally have to go up when the output curve is to go down, and vice versa; in Fig. 4 the curves are given with opposite polarities since they are taken directly from the operational amplifier).

An important concomitant feature is that the potentiometer scales cannot be calibrated in cps or decibels if the voltages are to be integrated, since the voltage level at a given time will be entirely dependent upon the past history of the signal.

The input-output relationship required according to our approach is as exemplified by "e_a" and "e_d" of Fig. 5 (for clarity the latter has been projected on the former and shown by a dashed line), except that here the transient time "τ_i" is assumed constant in order to simplify the exposition. The points "t_1", "t_2", etc. indicate the limits of each segment as specified in the input curve. The curve has (arbitrarily) been drawn in such a way that one of the segments (between "t_1" and "t_2") is of duration τ_i (whereas the other segments last longer). In this way the rising transition has been given a certain curvature.

The conversion of e_a into e_d is not a simple matter electronically even if the transient time is kept constant. However, it is possible to modify the input voltage function in such a way that the required output is obtained by integration. The principle underlying our experimental circuit can be stated by reference to the "Processed Input" curve "e_c" of Fig. 5.



t_n = any point of discontinuity of input step function

τ_i = preset transient time of integrating circuit

$$e_a = E_n \qquad t_{n-1} < t < t_n$$

$$e_b \begin{cases} = E_n - E_{n-1} & t_{n-1} < t < (t_{n-1} + \tau_i) \\ = 0 & (t_{n-1} + \tau_i) < t < t_n \end{cases}$$

$$e_c \begin{cases} = E_n - \frac{1}{2} E_{n-1} & t_{n-1} < t < (t_{n-1} + \tau_i) \\ = \frac{1}{2} E_n & (t_{n-1} + \tau_i) < t < t_n \end{cases} \quad \therefore \quad \boxed{e_c = \tfrac{1}{2}(e_a + e_b)}$$

$$e_d \begin{cases} = E_{n-1} + (E_n - E_{n-1}) \cdot \dfrac{t - t_{n-1}}{\tau_i} & t_{n-1} < t < (t_{n-1} + \tau_i) \\ = E_n & (t_{n-1} + \tau_i) < t < t_n \end{cases} \quad \therefore \quad \boxed{e_d = \int e_b \, dt = 2 \cdot \int (e_c - \tfrac{1}{2} e_a) \, dt}$$

Fig. 5.

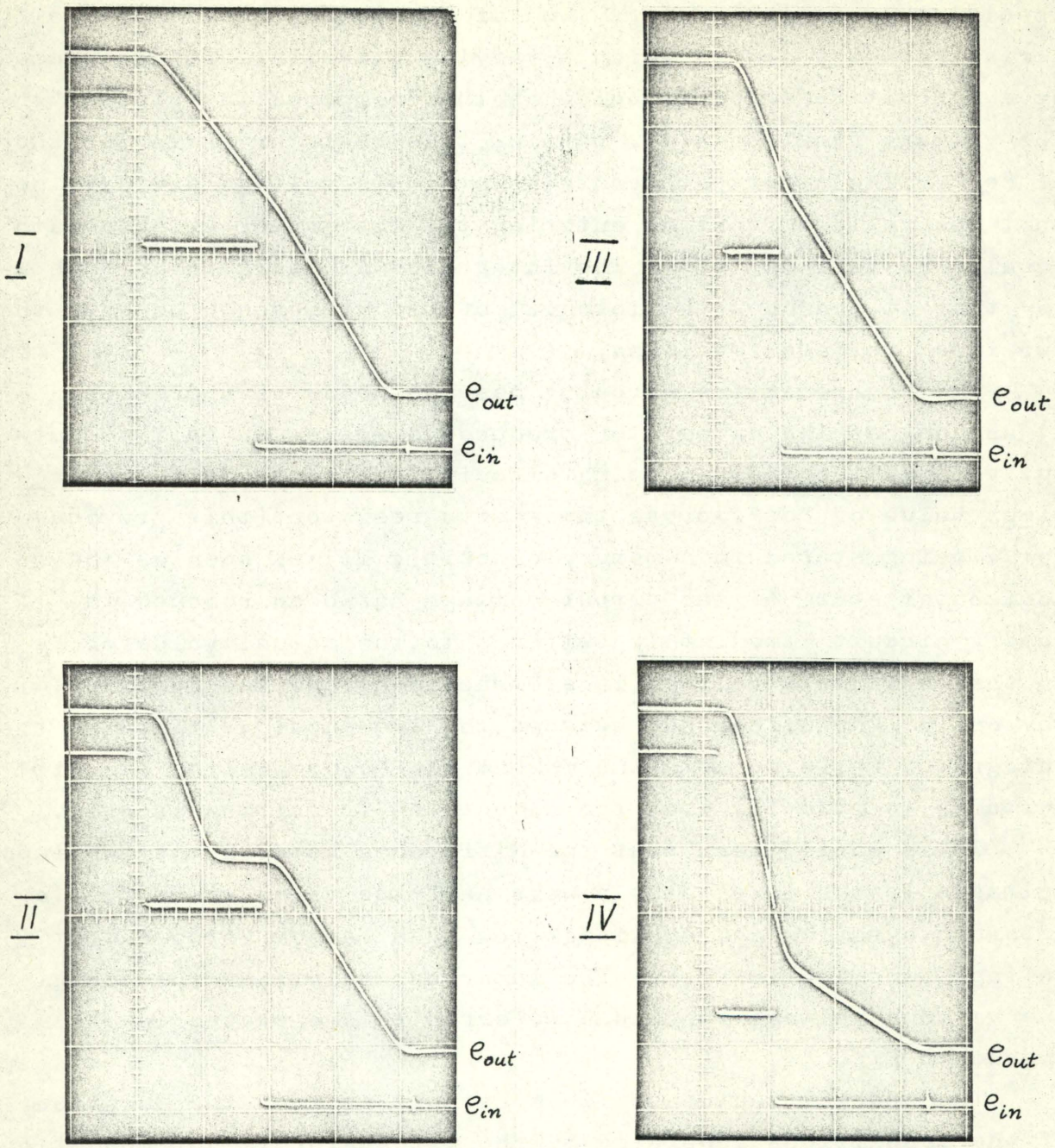
Logic of smoothing circuit.

The input voltage "e_a" is led by two different paths to a special type of integrator. One of the paths contains a small series resistor across which a varying potential drop is caused by a circuit referred to below as the "compensating circuit". The voltage function after passing this resistor is in the shape of "e_c". The other path contains a simple voltage divider, so that the voltage function entering the integrator by this path equals one half of "e_a". The integrator is designed in such a way that it produces the integral of the difference between the two input voltage functions.

The "compensating circuit" has two modes of operation: I) as long as the integrator produces a rising or falling output voltage the "compensating circuit" subtracts half the previous value of "e_a" across the series resistor (this previous value being stored in a memory circuit); II) as soon as the horizontal state of the output voltage has been reached the memory circuit immediately switches to the actual value of "e_a" so that half this voltage is subtracted across the resistor. Obviously, the difference between the two input voltages to the integrator falls to zero the moment the "compensating circuit" switches to mode II, i.e. the moment the transition is over.

It is easily seen that the difference between the two input voltages to the integrator equals half the value of "e_b", i.e. after integration and amplification by a factor 2 the output voltage function is "e_d". The numerical relationships among the various voltage functions referred to are stated in the legend to Fig. 5.

The special feature of this circuit is that the duration of the transitions generated by the integrator is not conditioned by the durations of the input voltage steps (except that these should not be shorter than τ_i). In the integrator a constant comparison between input and output voltages takes place, and the moment the output voltage equals half the instantaneous value of "e_a" the integrator is stopped and the output voltage remains constant until "e_a" assumes a new value. Thus the output voltage is clamped to the input voltage after every transition, which is an important safeguard against long-time DC drift.



I: τ_i of the first step equals the duration of the step - the slope produced is continuous though inflected.
II: τ_i of the first step has been shortened (everything else being equal) - two separate slopes are produced.
III: the duration of the first step has been reduced to $\tau_d = \tau_i$ - the two slopes become contiguous again.
IV: the voltage at the point of inflection has been lowered - the shape of the slope is changed.

Fig. 6.

Oscilloscope display of parameter synthesis with 3 variables.

(The constant offset between input and output is caused by a fixed filter inserted in order to round off the edges.)

Electronically, the basic unit of the integrator is a high gain transistor whose emitter is connected via a suitable resistor to the input terminal where "e_c" occurs, whereas the base is connected directly to the mid point of the voltage divider across which "e_a" occurs. The output is taken across a capacitor connected between collector and ground. Unfortunately this integrator can function only with one polarity of the input step function (with steps of the opposite polarity the output just follows the input though with some distortion). In order to obtain the same behaviour toward voltage steps of both polarities (cp. Fig. 3 above) it is necessary to have two complementary integrators and to combine their outputs in a mixing circuit which at any instant selects the most slowly changing voltage. As shown by Fig. 3 satisfactory results can be obtained in this way but certainly at a price.

The transient time is determined by the values of R and C in the integrator. By means of electronic switches it is possible to change the value of R from one event to the next. In the case of vowel formants a moderate number of different values of τ_i over the range 10 - 100 ms will certainly suffice.

Fig. 6 (preceding page) displays the three "dimensions" in which a control parameter can be varied according to the suggestions given above. It remains to be tested whether this three-dimensionality is a sufficiently attractive feature from the point of view of the experimenter to warrant its existence.

Acknowledgements:

The project is supported by grants from the Danish Research Committee for the Technical Sciences (Danmarks teknisk-videnskabelige Forskningsråd).

I am indebted to the staff of the Speech Transmission Laboratory of the Royal Technical High-School, Stockholm, for a stimulating discussion of some of the problems treated in this paper.

References:

(1) John M. Borst, "The Use of Spectrograms for Speech Analysis and Synthesis", Journal of the Audio Engineering Society 4 (1956), pp. 14-23, see pp. 19 ff.

(2) J. L. DeClerk, D. L. Phyfe, and R. A. Fisch, "Formant Synthesizer Electronically Controlled", Proceed. of the Speech Comm. Seminar, Stockholm 1962 (1963), Paper F4.

(3) J. B. Dennis, "Computer Control of an Analog Vocal Tract", ibid., Paper F3.

(3a) G. Fant and J. Mártony, "Quantization of Formant-Coded Synthetic Speech", STL-QPSR 2/1961 (Stockholm), pp. 16-18.

(4) Johan C. W. A. Liljencrants, "The OVE III Speech Synthesizer", IEEE Trans. on Audio and Electroacoustics AU-16 (1968), pp. 137-140.

(5) Jørgen Rischel, "Instrumentation for Vowel Synthesis", ARIPUC 1/1966 (1967), pp. 15-21.

(6) Jørgen Rischel and Svend-Erik Lystlund, "Speech Synthesizer", ARIPUC 2/1967 (1968), p. 34.

(7) W. Tscheschner, "Ueber ein Steuersystem zur Sprachsynthese", 5e congrès international d'acoustique (1965), rapport A53.