INSTRUMENTATION FOR VOWEL SYNTHESIS.

Jørgen Rischel

I. Preparatory work.

General presentation:

A vowel synthesizer employing a heterodyne system was built by this author in 1965-66. The apparatus grew out of somewhat random experiments as a device for demonstrating synthetic vowels, and it will not form part of our permanent instrumentation for research purposes. However, the experience gained from the experimentation will be utilized in the planning of a more advanced synthesizer. In this report, therefore, I shall present the essential features of the apparatus, and the particular problems that have been considered in connection with the heterodyne technique employed.

The synthesizer is a rather compact unit comprising five formant generating circuits plus a bass-boost network. A voice generator is connected externally.

In making this provisional setup I strived to obtain an optimum of versatility and simplicity of operation in the synthesis of steady-state vowel sounds.

(a)As for versatility it was considered desirable to be able to control the frequency, bandwidth and intensity level of each formant separately. Various arguments can be given in favour of this. In phonetics teaching it is obviously useful that it can be demonstrated how the individual formants (taken as peaks of energy in the spectrum) contribute to make up the phonetic quality of the sound, and it is also useful in connection with auditory tests to provide for the possibility of manipulating the formant levels or of changing the number of formants (e.g. to perform one-formant or two-formant synthesis). In principle this is possible with the present device (although the faithfulness of sound production may not be entirely adequate). Formant frequencies are continuously variable over a calibrated range representative of adult (male) voices. Formant bandwidths are continuously variable over a (roughly) calibrated range of some 50 to 150 c/s (different for the different formants). Formant levels are continuously variable over a calibrated range of 0 to -60 dB and can further be turned down to zero ("-oo"). The correct setting of formant levels presupposes, of course,

a knowledge of the spectrum characteristics of the voice source employed. On the other hand, variations in formant levels can in principle be used to imitate particular voice characteristics. Variations of bandwidths and levels also make it possible to imitate nasalized vowels and some frictionless consonants with a certain degree of realism without the aid of additional circuitry (see below on "F 5").

The above considerations imply that the speech-wave is generated by adding the contributions from the individual formant generators, i.e. the electronic formant filters are connected in parallel, not in a series (cascaded) arrangement.*)

(b) Simplicity of operation is in a sense the opposite of versatility. A synthesizer with parallel connection of formant filters does not have the limitations imposed on human speech by the vocal tract anatomy, and thus a large amount of information on different parameters (16 in our setup) is necessary for correct simulation of a given vowel sound. A cascaded arrangement of formant filters reproduces more directly the transfer characteristics of the human vocal tract and is likely to be superior in high quality speech simulation, cp. the instructive comparison of the two types by Cooper (1). The relative merits of parallel and cascade synthesis of connected speech are still debated, see Flanagan (2).

If the necessary information is present, however, our present synthesizer is easy to operate. Although it was designed for steady-state synthesis only, it was considered essential that each formant be continuously variable over its whole frequency range, and that its true frequency location be directly visible on the scale. This is not generally true of simple vowel synthesizers. A frequency variation is most easily achieved by a stepwise variation (by means of switches) of the capacitors in each resonant circuit, and this is probably the method that is mostly used for the restricted purpose of vowel synthesis. However, it is a drawback of decade switches that the whole formant range cannot be shown in a sweep, and it is a more serious

---

*) It must be added here that the heterodyne technique employed in our experiments dictates the use of a parallel connection of formant filters. It cannot be used with a series arrangement, even though such an arrangement might be preferred for several purposes.

drawback that the scales must be calibrated in capacitance values (microfarads, nanofarads, etc.) rather than cycles per second, so that the frequency location of a formant cannot be determined (if only approximately) without the use of additional measuring apparatus or of detailed nomograms showing the capacitance/frequency relationships.

In our synthesizer the formant frequency variation is at present achieved by means of variable capacitors. The range of variation with these is sufficient only at relatively high frequencies. This has been accounted for by heterodyning the entire signal from the voice source with a sine wave of fixed frequency, thus moving it to a suitable high frequency range, within which the formant filters (resonant LC-circuits) are tuned, and modulating back afterwards to the original low frequency location.

The heterodyne method was adopted not only with the purpose of using variable capacitors but also because it seems potentially applicable to synthesis of connected speech, if adequate means for capacitance variation be found. Experiments with BA 1o2 capacitance diodes (controlled by a negative DC-voltage to give a varying capacitance) inserted in one of the formant circuits of the present device gave a promising result, indicating that formant frequency variation can be accomplished in an extremely simple fashion by applying a DC control voltage directly to the resonant circuits. Admittedly, this simplicity of formant frequency control is obtained at the expense of a fairly elaborate system of modulators and filters in the synthesizer. It should be added also that the variation of formant frequency with DC voltage will not be linear. We are going to investigate into this problem.

An interesting property of heterodyne synthesizers is that a resonant frequency can be changed not only by changing the component values of the resonant circuit but also by changing the frequency of the sine wave from the carrier oscillator. A possibility which is at least theoretically interesting is to use separate heterodyne systems for the several formants. Each formant filter could then be kept at a fixed resonant frequency, and the formant frequency variation would be achieved by changing the frequency of the carrier

belonging to that particular formant generator. However, this method would probably be more complicated than others in current use.

The heterodyne approaches discussed here are essentially different from that described by Lawrence as early as 1953 (3). In his parametric artificial talker a formant-like wave train (i.e. a decaying sine wave recurring at a rate corresponding to the fundamental frequency) is generated at a fixed frequency and distributed to the proper frequency locations of the individual formants by simultaneous heterodyning processes. The synthesizer described by Pohlink et al. (6) is more similar in type. However, they apply the signal directly to the formant filters and only modulate afterwards (F_o being synchronized by the carrier).

Technical outline of the heterodyne synthesizer:

A block diagram of our provisional synthesizer is shown in Fig. 1. The signal is applied to a double-balanced modulator which delivers an output consisting of two sidebands located around a suppressed carrier frequency of 15 kc/s.*) The upper sideband (15 to 2o kc/s) is used for the transmission of the signal. Via a step-down transformer the output from the modulator is applied to the formant filters, which are series resonant LC-circuits of high Q. The formant filters of odd number are fed from one transformer coil, those of even number are fed from another with phase reversal. Thus a combined response without zeroes at the cross-over frequencies (i.e. without antiformants between the formants) is obtained (cf. Weibel (4)). The "standard reference vowel" response of the whole system is shown in Fig. 2. (The comparison of envelopes obtained with parallel synthesis and with series synthesis given in Flanagan (5) presents a radically different picture for parallel synthesis

---

*) This frequency appeared to be a reasonable compromise between the difficulty in obtaining narrow bandwidths of variable resonant circuits at high frequencies, and the difficulty in obtaining the high impedance necessary for the application of variable capacitors or capacitance diodes at low frequencies.

because obviously no phase inversion was used.)

"F5" covers the entire frequency range (in steps) and can be connected in phase with either the odd or the even formants. It can thus be used to simulate nasality or the like, if placed at a low frequency.

The combined output from the formant generating network is high-pass filtered at 15 kc/s (to remove remnants of the lower sideband and possible distortion products) and demodulated in a second double-balanced modulator. The low-frequency output is low-pass filtered at 7 kc/s and mixed with the (phase-correct) output from a low frequency boost channel (consisting of an LRC low-pass filter followed by a variable gain amplifier), which takes the signal directly from the input. (By this procedure the effect of nonlinearity originating from imperfect filtering in connection with the single-sideband modulation is reduced, and correct low-frequency response is obtained.)

Attempts to produce vowels of well-defined phonetic quality with the synthesizer have been reasonably successful considering the fact that we have not had a really adequate signal source at our disposal. This will be provided for in the nearest future.

Although the accuracy of the carrier oscillator is of course crucial, the synthesizer is reasonably stable in operation. Noise is kept at a low level, even though no particular care has been taken in the layout. The high signal-to-noise ratio easily obtainable with this type of synthesizer may be a feature worthy of notice.

When the provisional synthesizer was first constructed, economy had to be a governing factor in a pilot study of this kind, and low-cost components were accomodated wherever possible. Lately, various improvements especially of the modulating system have been incorporated.

2. Plans for future work.

In October 1966 the Institute received a grant from the Danish Technical Research Committee as a financial support of our acquisition of instrumentation for synthesis and filtering of speech. We are at present purchasing various measuring instruments, and in the near future we shall investigate into the principles to be fol-

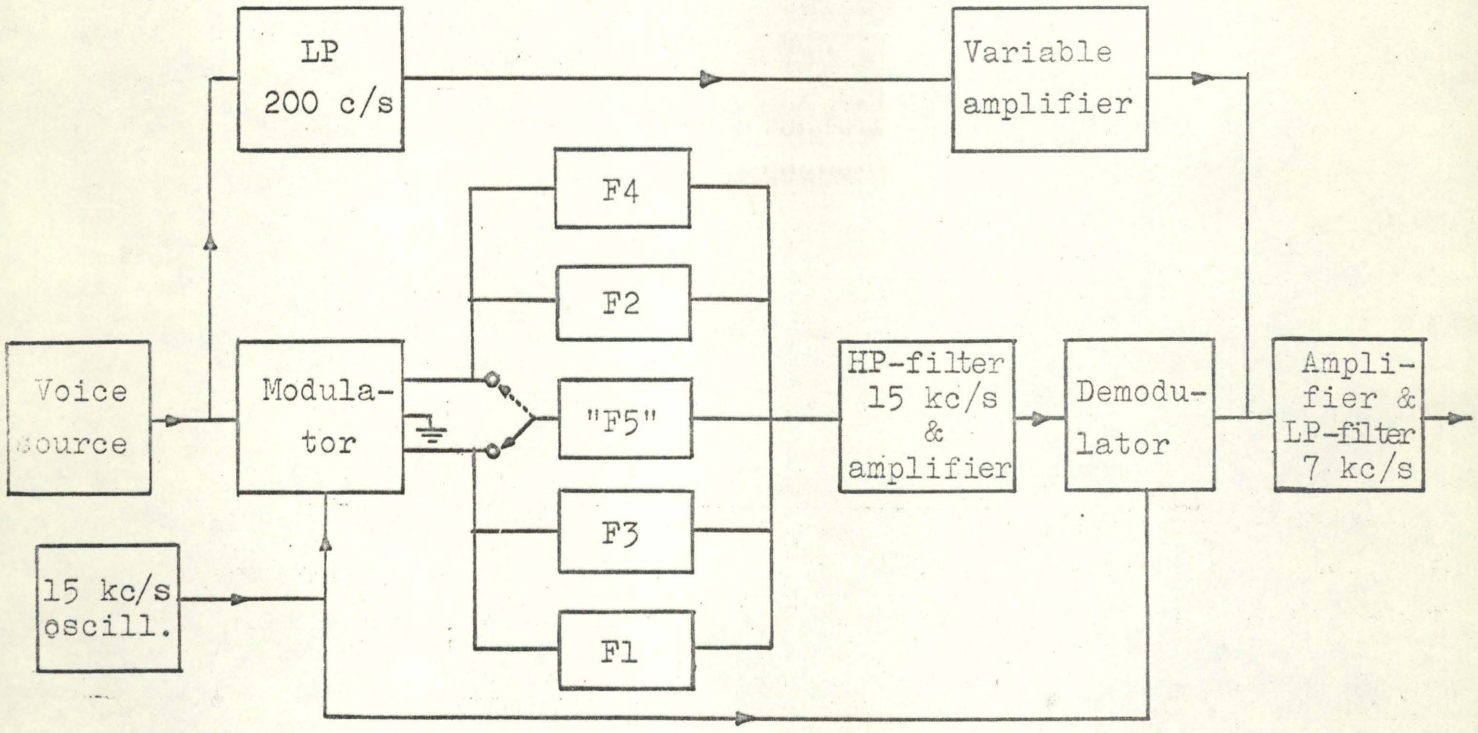

Fig. 1

Block diagram of provisional vowel synthesizer.

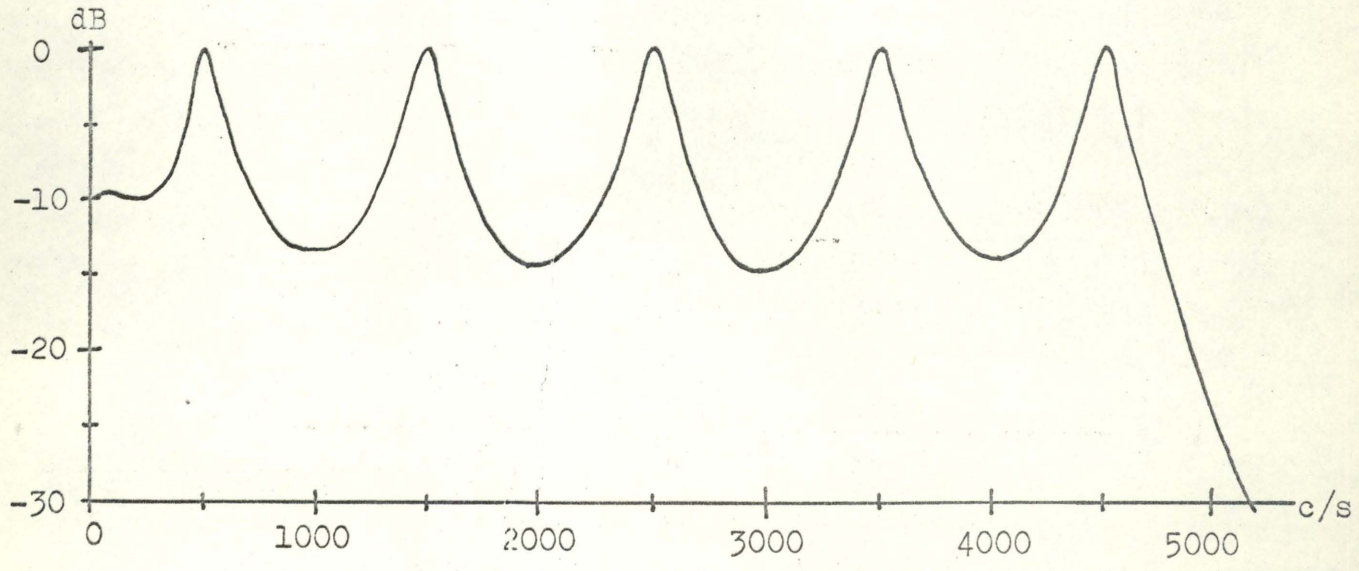

Fig. 2

Response of synthesizer with formant frequencies set for the "neutral vowel" (500 c/s, 1500 c/s, 2500 c/s, etc.), bandwidths = 100 c/s, and levels = 0 dB (the actual choice depends on the voice source spectrum).

lowed in the final choice (design or purchase) of a parametric (formant coded) synthesizer.

References:

(1) F.S. Cooper, "Speech Synthesizers", Proceed. of the Fourth Int. Congr. of Phon. Sc. Helsinki 1961 (1962), pp. 3-13.

(2) J. L. Flanagan, Speech Analysis, Synthesis and Perception (1965).

(3) W. Lawrence, "The Synthesis of Speech from Signals which have a Low Information Rate", Communication Theory ed. W. Jackson (1953), pp. 46o-469.

(4) E.S. Weibel, "Vowel Synthesis by Means of Resonant Circuits", JASA 27 (1955), pp. 858-865.

(5) J.L. Flanagan, "Note on the Design of Terminal-analog Speech Synthesizers", JASA 29 (1957), pp. 3o6-31o (p. 31o, Fig. 7).

(6) Pohlink, Tscheschner, Dumdei, "Demonstrationsmodell zur Sprachsynthese", Zs. f. Ph., Sprachwiss. u. Kommunikationsforsch 18(1965), pp. 4o9-416.