A FORMANT-CODED SPEECH SYNTHESIZER

Jørgen Rischel and Svend-Erik Lystlund

1. Introduction

The speech synthesizer of the Institute of Phonetics has been completed in all essentials according to the design plan outlined some years ago (see the brief mention in Lystlund and Rischel 1968). It may be expedient, therefore, to give a survey of the apparatus, and the reasoning behind its design.

The synthesizer is a formant-coded device controlled by means of a set of time-varying DC-voltages. At present, these parameters are supplied by a function generator.

2. The synthesizer proper

2.1. Phonetic strategy

The phonetically interesting aspects of the synthesizer layout are shown in Fig. 1. It is seen that there is a generator of quasi-periodic pulses which functions as a voice source, and a generator of random noise. The transfer function of the vocal tract is approximated by a system of resonators all coupled in parallel. Each of these resonators (henceforth referred to as formant filters) takes care of a local section of the spectrum, comprising one formant peak.

The repetition rate of the voice pulses (F_o) can be varied over a wider or narrower range according to the compromise between requirements on intonation range and accuracy of resolution which seems preferable in each case.

The contributions from the voice and noise sources are led

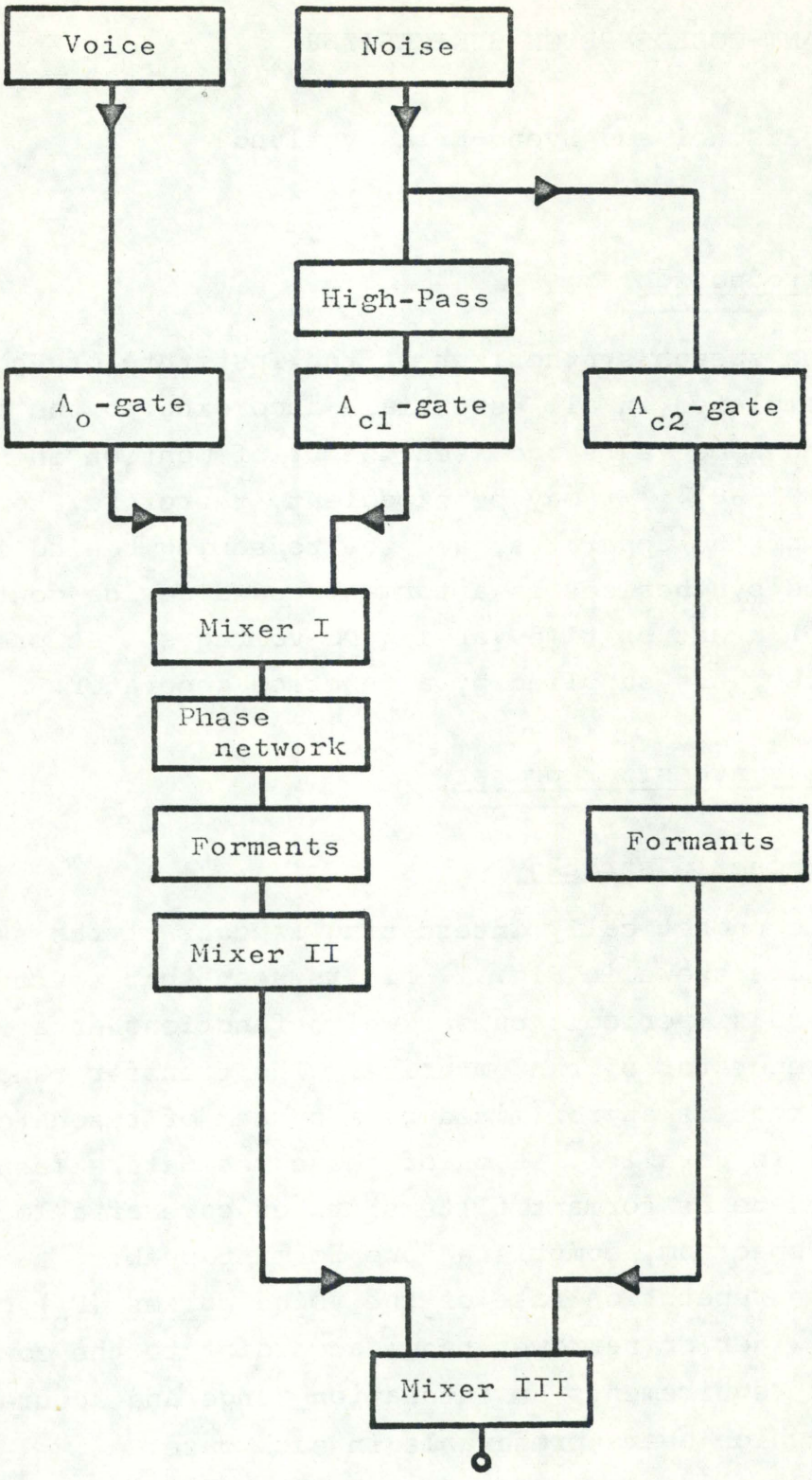

Voice
Noise
High-Pass
Λ_o-gate
Λ_{c1}-gate
Λ_{c2}-gate
Mixer I
Phase network
Formants
Formants
Mixer II
Mixer III

Fig.1

via gates (providing a continuous amplitude control) to each formant filter, and the contributions from these filters are combined by a summation. The individual contribution of each formant filter can be varied continuously (and independently of the other filters) in terms of resonant frequency, bandwidth, and level. The ranges of variation are chosen empirically to be suitable for synthesis of various sound types, but since they can be modified to satisfy special needs there is no point in listing them here.

There is a total of nine formant filters in the system. Five of these are intended to imitate the five lowest resonant modes of the vocal tract, viz. F_1 to F_5. In addition there is a low frequency pole, "F_{sub}" serving to shape the spectral region below F_1, and another pole, "F_{nas}". Both of these can be used to improve the match of natural sounds which have an excess of formant peaks compared to ideal vowels, e.g. vowels with "split" formants or nasal consonants. However, F_{sub} is constantly employed to approximate the low frequency boost of the human voice, which is thus contained conceptually in the transfer function as a low frequency pole.

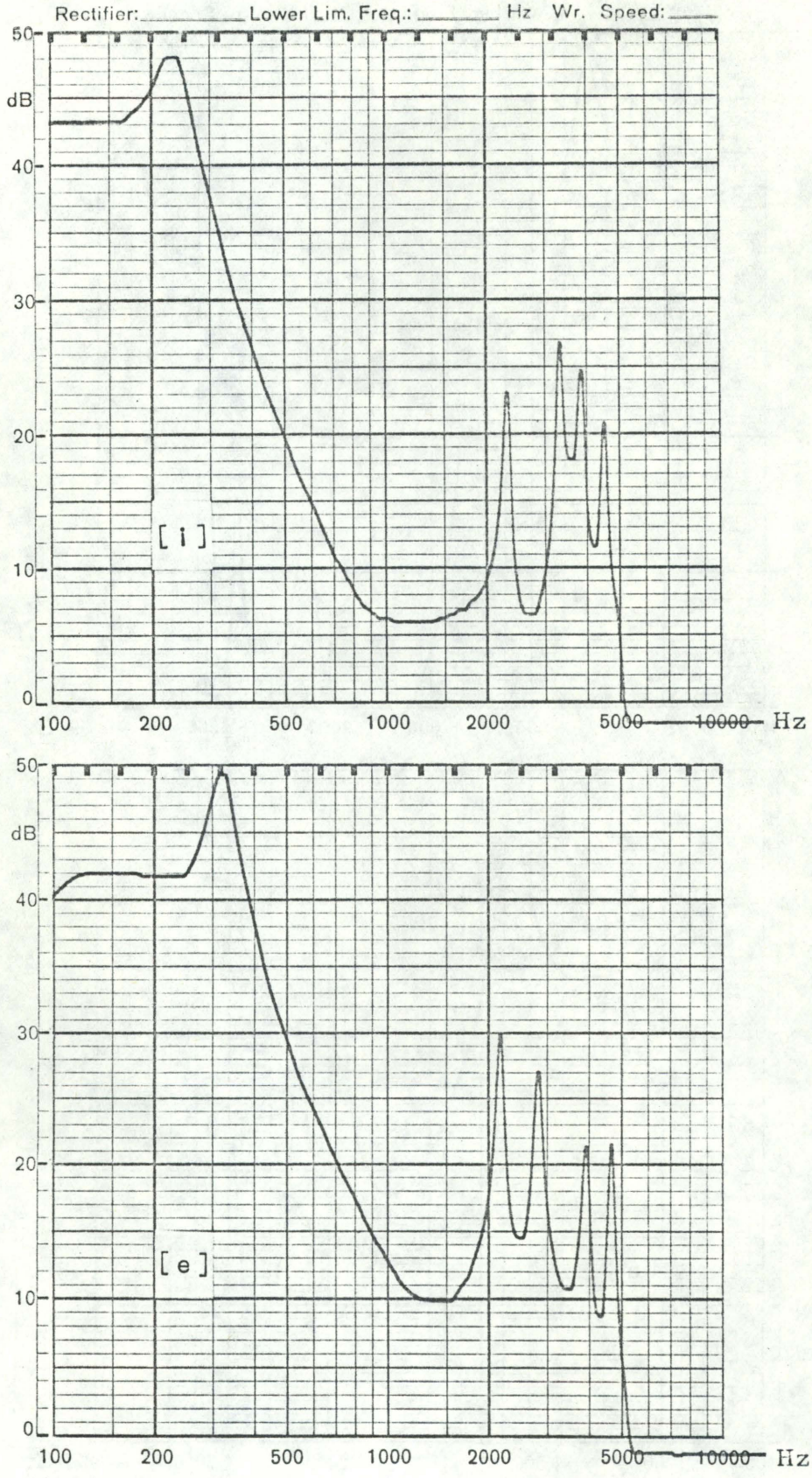
In addition to the formant filters mentioned so far, there is a separate configuration of two high-frequency formants, nominally F_6 and F_7. In contradistinction to the lower formants, which can be supplied with voice pulses (via the A_o-gate) or noise (via the A_{c1}-gate), or both, the two highest formants can be supplied with noise only, which is introduced via a separate gate common to them (A_{c2}). F_6 and F_7 are coupled in such a way that they form a block of energy in the upper part of the spectrum, the upper and lower limits of this region being determined by the frequency locations of the resonant peaks. The spectral contribution of this subsystem falls off very rapidly below the lower resonance, so that it can be used to give a rough approximation of the high-frequency characteristics of

fricatives, whereas the fine structure of the lower frequency region, i.e. roughly below 4500 cps, is represented more faithfully by the lower formants.

The parallel arrangement of formant filters has an important consequence in vowel synthesis: the residue of a formant at the frequencies of other formants is (almost) negligible on a summation basis, i.e. changes in formant frequencies will not automatically provide the changes of formant levels inherent in human speech. Without such level variations the quality of synthetic speech is quite poor and unsuitable for many research purposes. Recent synthesizers intended for high-quality speech are mostly of the series coupled type (i.e. with a cascade arrangement of the formant filters), which automatically provides the level variations desired. If, however, the amplitude of each formant is controlled, and if the circuits are given appropriate frequency and phase characteristics, vowels of high quality can be produced on a parallel coupled synthesizer, as we have experienced over the last years (also see Fig. 2). For various reasons we have found it worth while to try out the potentialities of such a device for phonetic research purposes.

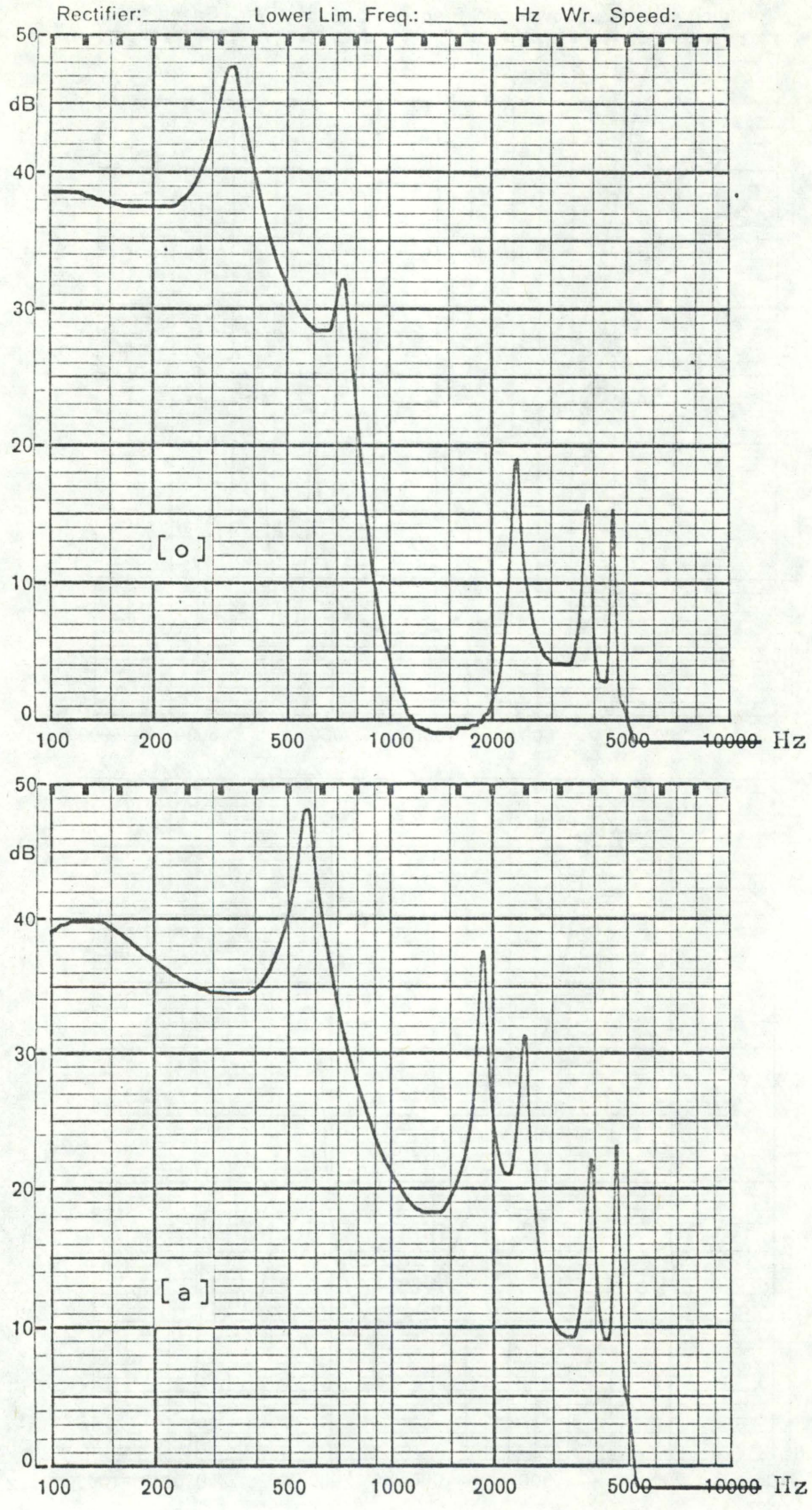
The predictability of formant levels from formant frequencies should clearly be made use of in a synthesizer designed to produce high quality speech in an economic way. From this point of view the series coupled synthesizer is preferable since the parameters required to operate it contain less redundancy (cf. Fant 1959 p. 47). It is worth noting, however, that this argument applies only with essential reservations. Nasal coupling, leakage through the glottis, and local constrictions in the oral cavity may introduce anti-resonances and additional resonances which considerably influence the relative prominence of different parts of the sound spectrum. This effect can, at least to some extent, be approximated in a parallel connected synthesizer simply by adjusting the formant levels, whereas the

Brüel & Kjær



Spectral envelopes of synthetic vowels

Fig. 2a



Spectral envelopes of synthetic vowels

Fig. 2b

series coupled synthesizer is less versatile in this respect (the approaches available with the latter, e.g. a change of formant bandwidths or a modification of the voice source spectrum, are of course equally available with the parallel connected device). Hence, for research on the fine structure of human speech sounds, and its relevance to perception, the parallel coupled synthesizer may be a useful tool.

If we turn now to consonants, a transfer function involving poles only is disputable for liquids and entirely inadequate for fricatives and stops. A series coupled synthesizer must, therefore, be composed of several parallel branches, e.g. one for vowellike sounds and [h], one for nasals, and one for friction noise, as in the Swedish Ove II and Ove III. In the parallel coupled synthesizer one single system can, in principle, be used for the different sound types since formants can emerge or vanish according to the settings of their individual level controls. By strongly attenuating some of the lower formants, for example, one can approximate the transmission characteristics of voiceless fricatives. This means that there is no discontinuity in the formant specification of the spectral properties of adjacent sounds. Thus F_2 of a vowel may (with appropriate bending, and with addition of noise or replacement of voice by noise) continue as F_2 of an adjacent fricative without any switch of parameter, F_3 of a vowel may (with appropriate weakening) continue as F_3 or F_4 of an adjacent nasal, etc. Formants may be suppressed or emerge from nothing, but the formants which are continuous in the sonagram will be controlled consistently by the same set of parameters.

This means that the synthesizer can be operated rather directly from sonagrams by simple spectrum imitation (note that formants which are invisible on sonagrams will be largely negligible in parallel synthesis, since their frequency location does not influence the general shape of the spectrum).

This may be a useful feature in connection with experiments where the perceptual importance of spectral details are at issue.

As a systematic approach to speech synthesis we do, however, employ pre-calculation of sound spectra on the basis of the theory of speech production.

Synthesis by means of a parallel arrangement of formant filters poses a major problem with regard to voiced fricatives. A sound like [z] requires a vowellike formant filtering of the contribution from the voice source (with a strong predominance of F_1, the higher formants being weakened because of the low frequency of F_1), whereas the contribution from the noise source should be characterized by an attenuation of lower formants. This effect of having the voice and noise sources located at different places in the human speech organs cannot be imitated in a straightforward manner. There are two ways to remedy the situation. One is to shape the spectrum of the noise source before it is processed by the formant filters. Another possibility is to apply noise to higher formants only. In our present setup we take both of these measures. The noise applied to the lower formants is filtered in such a way that it is strongly attenuated at very low frequencies, i.e. a low F_1 has practically no noise component in it. Moreover, the highest formants are supplied with noise via a separate gate (the strategy may have to be improved further on this point). Finally, the low frequency boost of the voice source is enhanced by the low-frequency pole (F_{sub}), which is located in a region where the noise is maximally attenuated.

2.2. Electronic design

The synthesizer is based on the heterodyne principle presented in Rischel (1967). The pulses from the voice source are processed by a system of filters and modulators which shifts the

useful frequency band from 0-5 kc to 8-13 kc (carrier frequency 8 kc), and back again. The formant filters are inserted into this system at a point where the signal occurs with frequency transposition (as the upper sideband from modulator I). The noise signal is applied directly to the system of formant filters, i.e. frequency transposed only once, whereas the voice signal must be frequency transposed twice in order to ensure a preservation of the harmonic relationship among its components.

Since the formant filters are tuned to much higher frequencies by this technique than they would otherwise be, the relative variation in formant frequencies is drastically reduced. This means that the formant levels vary very little with frequency variation, because the Q, and hence the bandwidth, remains practically constant. This is a necessary prerequisite for a calibration of the scales of the control system (see 3. below). Moverover, we have found it possible and expedient to accomplish the frequency variation by varying the bias voltage applied to capacitance diodes in resonant circuits. This is a very simple method (though it requires a resistance network shaping the control voltage in order to get a reasonable frequency scale). With the high-quality components (particularly inductances) now available, it has been no problem to obtain sufficiently high Q's, i.e. formant bandwidths of the order of 50 cps or less.

The design appears from Fig. 3, which is an elaborated version of the block diagram of Fig. 1 (with inclusion of the heterodyning system). A single formant circuit is shown in Fig. 4.

The formant circuits are preceded by phase inverters (in actual practice the phase correction is more complicated than suggested in the block diagram), and moreover, the circuitry includes extra lowpass filters. The reason for this increased complexity is that a parallel arrangement of formant filters

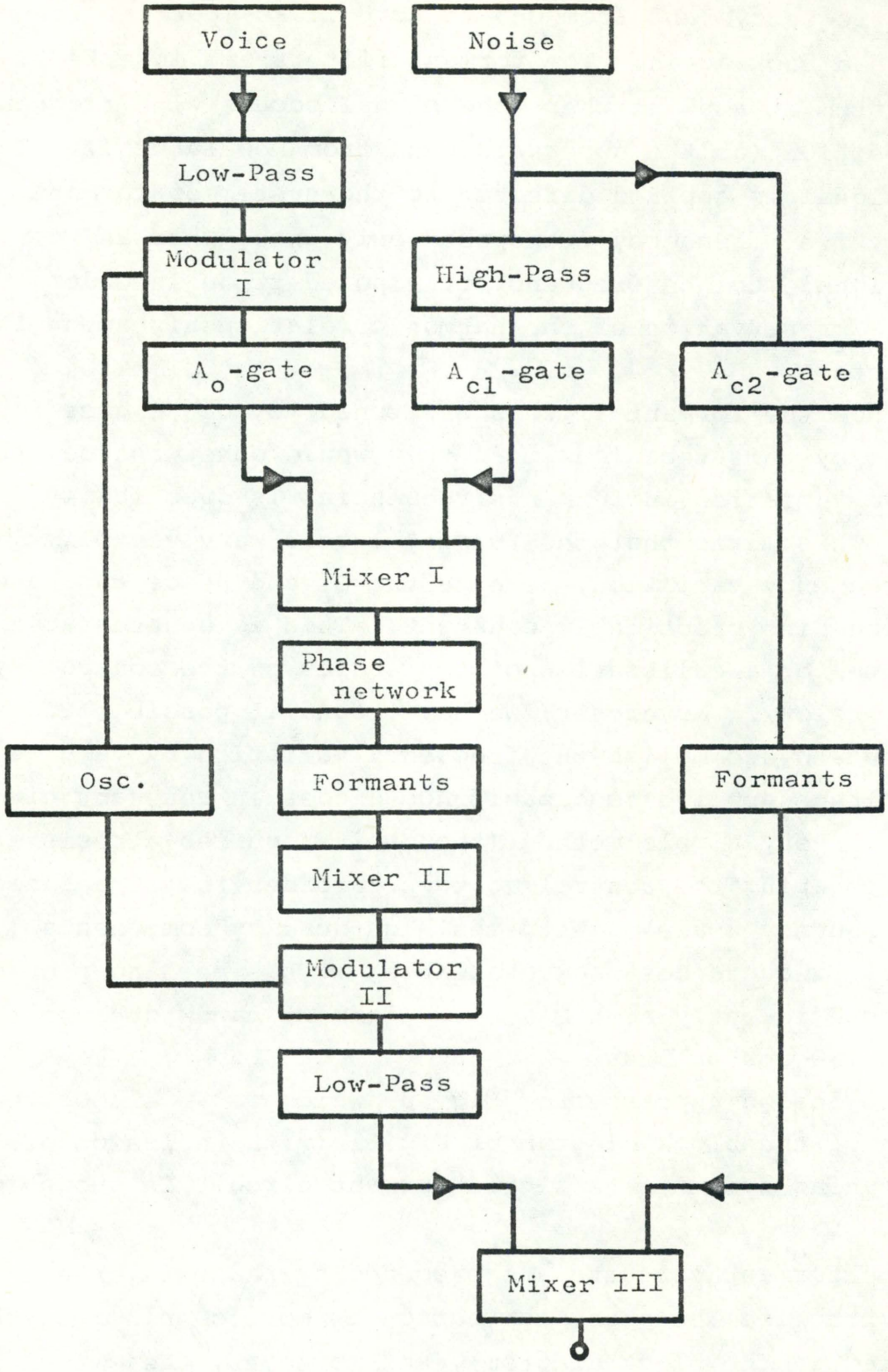
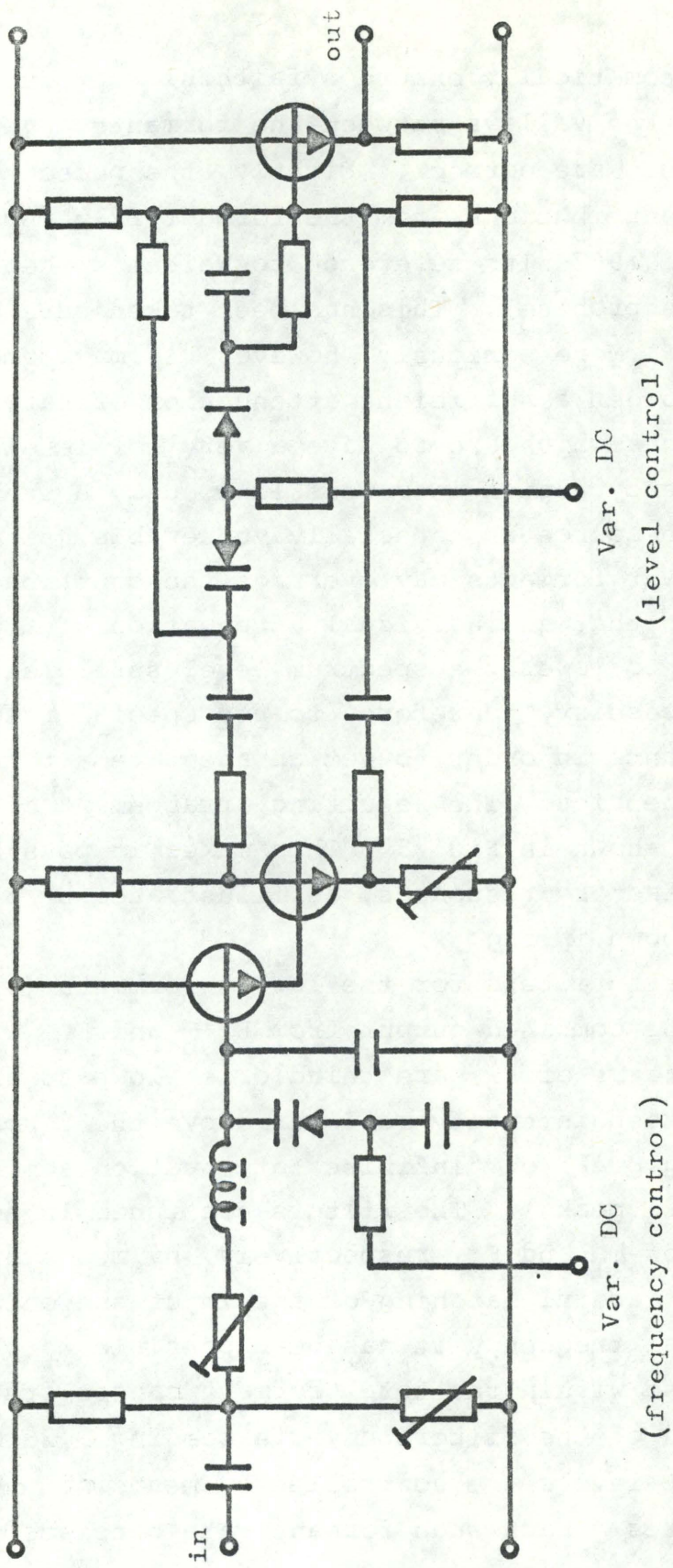

Voice
Noise
Low-Pass
Modulator I
High-Pass
A_o-gate
A_{c1}-gate
A_{c2}-gate
Mixer I
Phase network
Osc.
Formants
Formants
Mixer II
Modulator II
Low-Pass
Mixer III

Fig.3



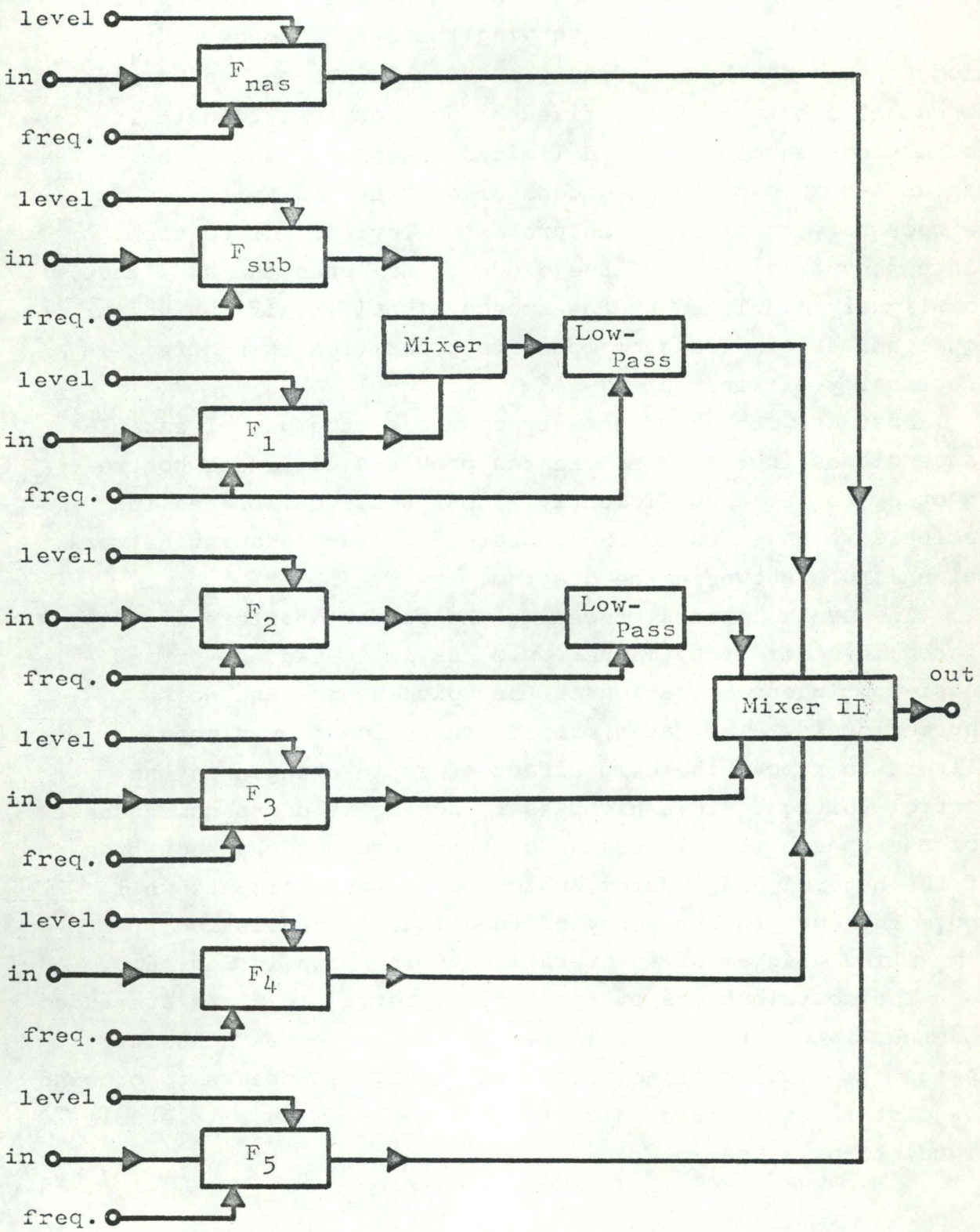
Formant circuit

Fig.4

does not automatically ensure a faithful reproduction of the spectrum in the valleys between the formants, even if the formant levels are correct. Firstly, the phase relationships among the contributions from the formant filters must be correct (see Rischel 1967 with reference to Weibel's theoretical treatment of this problem). This has been taken care of in our synthesizer. More seriously, however, it may sometimes be difficult to get a sufficient attenuation of harmonics in frequency regions which are to have a very low level, e.g. the upper frequency region in vowels like [u]. A synthesizer using a flat voice source is especially vulnerable, since the residues from the lower formants may override the contributions of higher formants, so that an individual attenuation of these latter does not suffice to lower the spectrum level sufficiently. We have found it necessary, therefore, to use special filters for the lowest formants in order to get an adequate attenuation at higher frequencies. The resulting arrangement of the formant circuitry is shown in Fig. 5. This makes it possible to get quite satisfactory responses, as illustrated by the spectrum envelopes shown in Fig. 2.

The extra filters for the lowest formants (one of which processes the combined output from F_{sub} and F_1, whereas the other takes care of F_2) are third order lowpass filters which provide a steep intensity roll-off above the formant peaks (there being peaks of "infinite" attenuation some 1 kc above the F_1 and F_2 peaks). The filters are tuned in accordance with the tuning of F_1 and F_2, respectively, by means of capacitance diodes. By careful matching of the input and output impedances at an optimal frequency it has been possible to obtain a very good response within the total formant ranges (the ripple in the passband of the filters never exceeding 0,25 dB).

Formant levels are controlled by means of continuously variable gates. Each such formant gate consists of a bridge



Arrangement of formants in heterodyne system

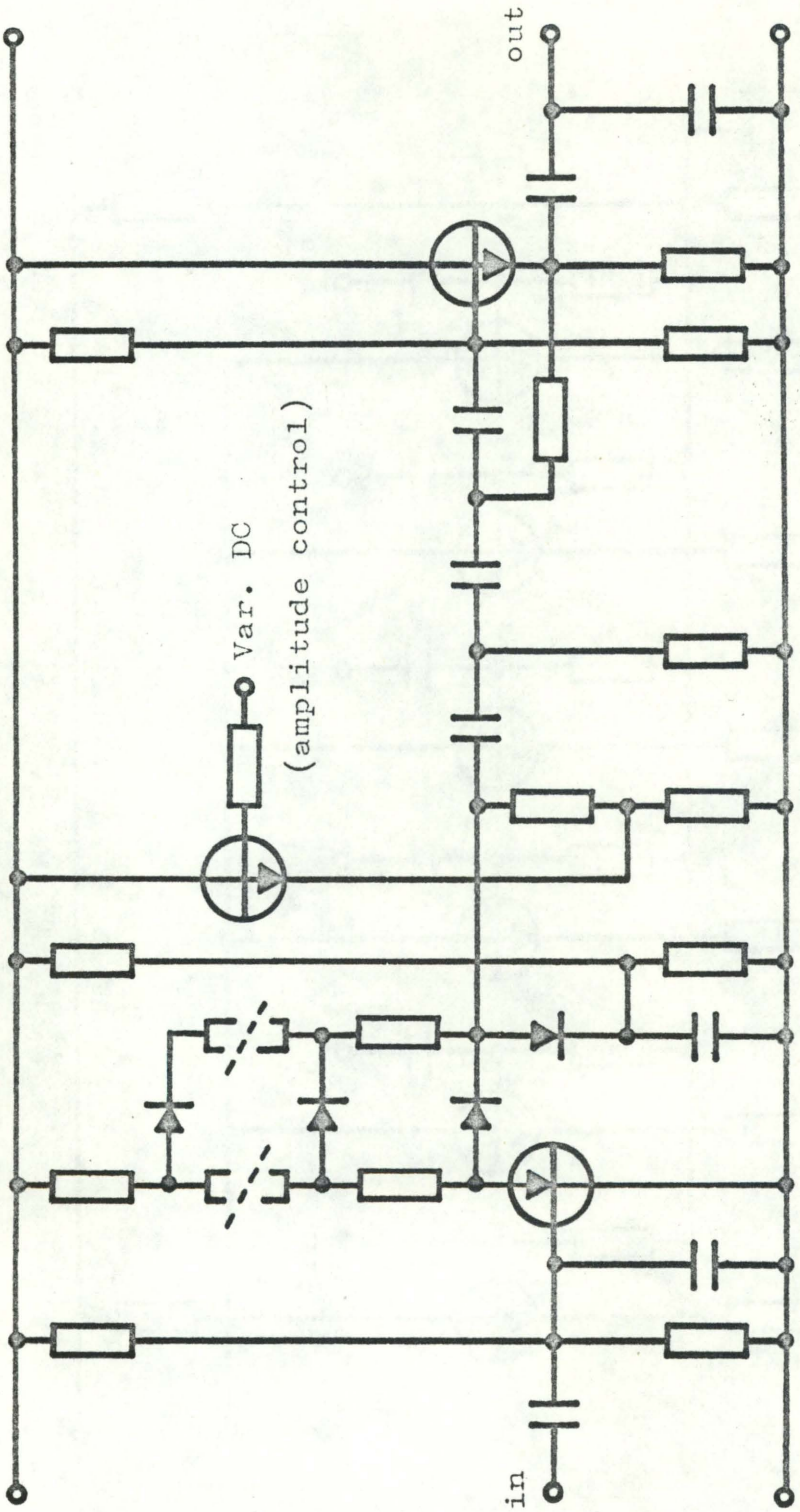
Fig.5

circuit (see Fig. 4) one branch of which is represented by a fixed capacitance whereas the other contains a configuration of capacitance diodes. By varying the bias voltage to the diodes the contribution from the variable branch can be made to cancel out that of the fixed branch, or to attenuate it to varying degrees within a limited frequency band. The variable branch contains diodes of opposite polarity. This is necessary in order to ensure a low level of distortion, since the capacitance of the diodes, and hence the amplitude, is slightly modulated by the speech signal itself. With an equal number of diodes of opposite polarities this effect is practically eliminated.

Just as with the frequency controls, the use of capacitance diodes in the gates creates problems of linear control. In order to get a sufficiently linear scale calibrated in decibels we have had to incorporate a voltage shaping network, which is not shown in the diagram.

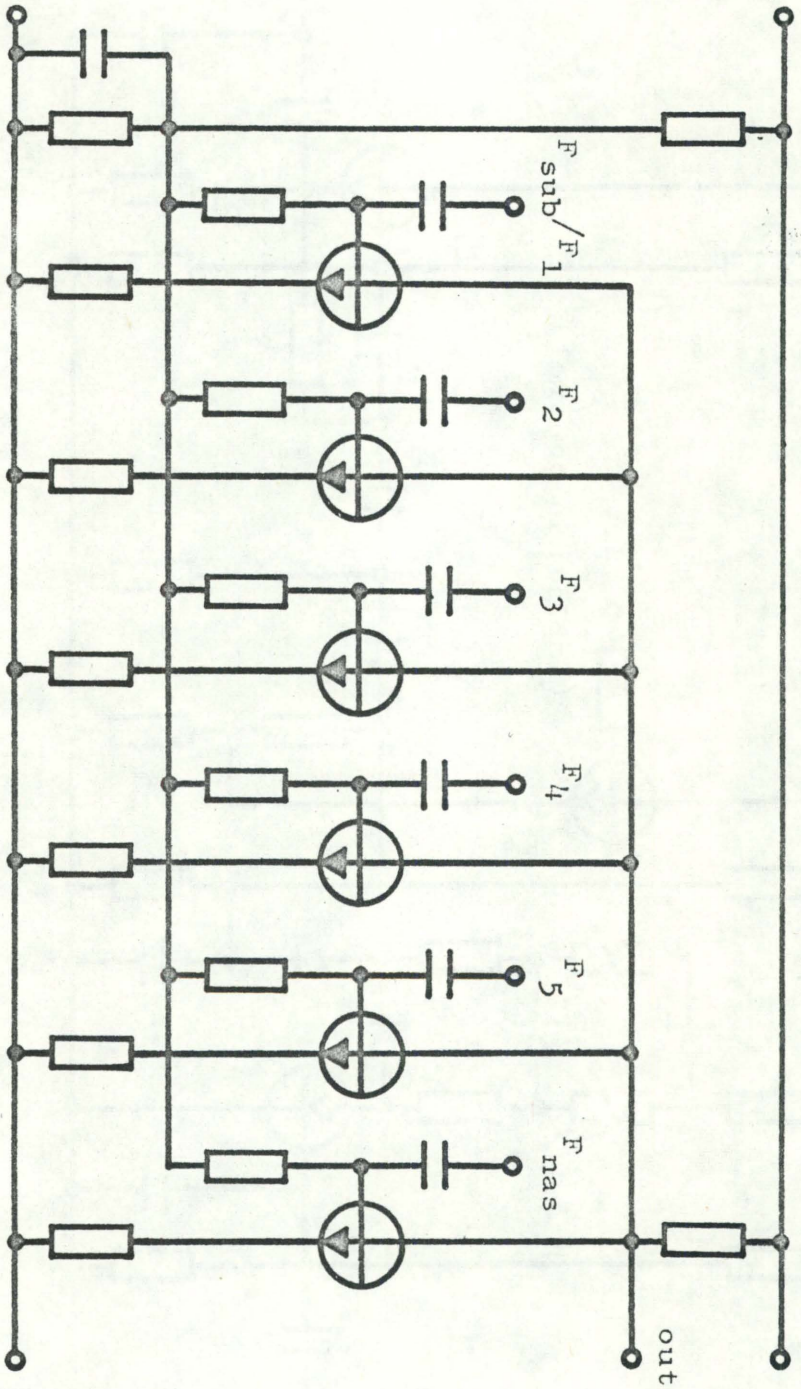
The overall gates preceding the formant filters are designed in a different manner, viz. as diode ladders giving a quasi-continuous variation of the voice source and noise source levels. The gates proper are followed by highpass filters to remove the bump effect of rapid changes in the control voltage. This highpass filtering is of no consequence for the speech signal because of the frequency transposition of the heterodyning system, which places the formants well above the cutoff frequency of the filters in question. - A simplified diagram of an overall gate is shown in Fig. 6.

The contributions of the various formant filters are added together in mixers, three in all in the present synthesizer design. The mixer principle shown in Fig. 7 appears to combine low distortion (even at high signal levels) with a tolerable signal-to-noise ratio.



Amplitude gate

Fig.6



Mixer circuit

Fig.7

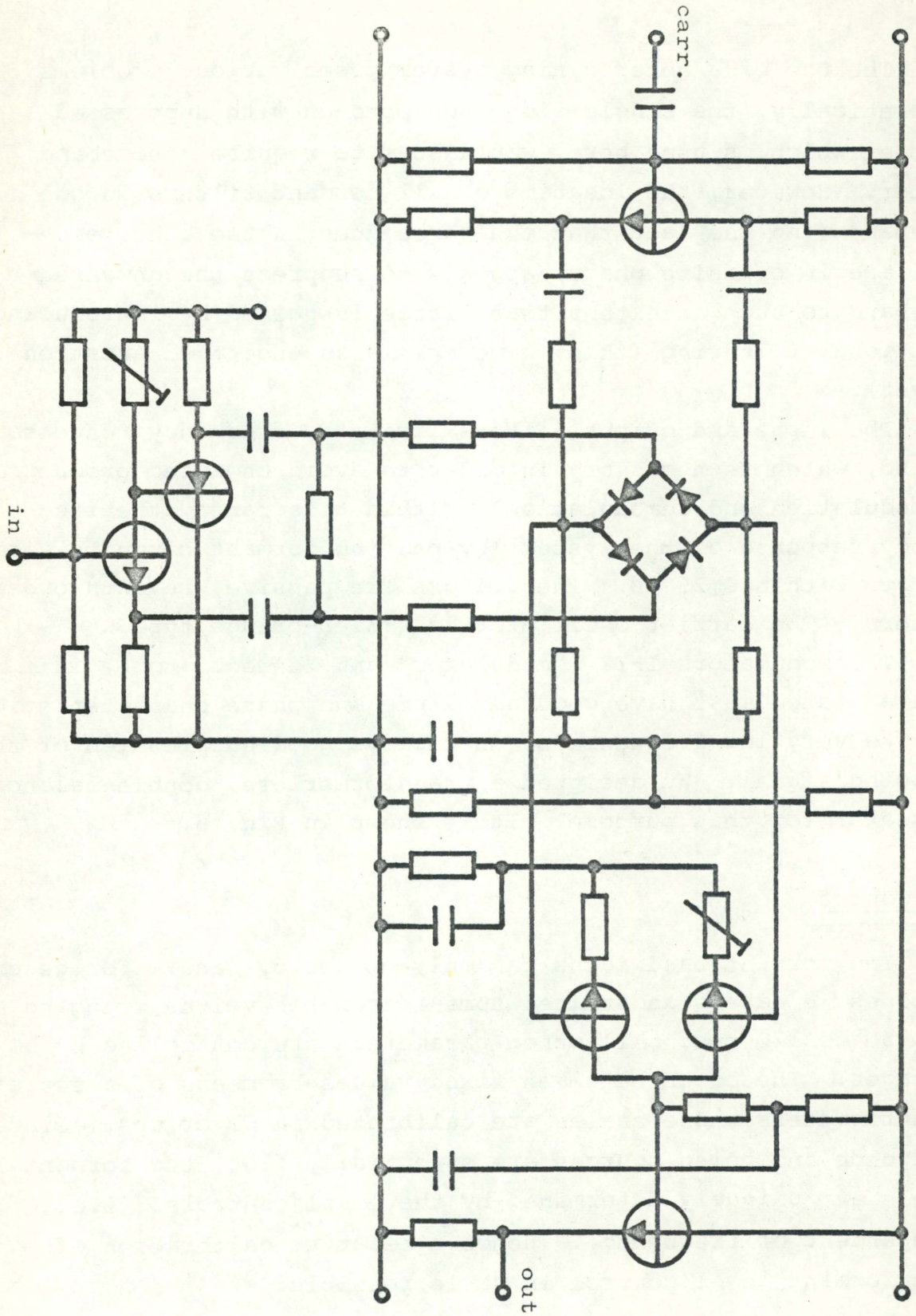
The use of a heterodyning system poses various problems. Theoretically, the single sideband approach with suppressed carrier which is used here, would seem to require a sideband filter. However, the location of all formant filters in one sideband (and the fact that their residues in the other sideband are in opposite phase) appears to suppress the unwanted sideband to such an extent that little is gained by introducing additional filtering (this, however, is an unsettled question as yet).

The input and output filters limit the frequency range to 0-5 kc, which is necessary in order to avoid unwanted products of modulation and demodulation. Within this range, the frequency response of the system (by-passing formant circuits) is flat to within 0,25 dB. The filters are passive, seventh order filters. The carrier oscillator is a Wien-bridge type.

The modulators I-II ("modulator" and "demodulator") in such a system must have good amplitude and phase characteristics down to very low frequencies, as well as good suppression of the incoming signal. We designed a transformerless, double-balanced modulator for this purpose. It is shown in Fig. 8.

3. Control

The formant bandwidths can only be set by hand. For general purposes we have standardized some reasonable values, ranging from 50 to 120 cps. All other parameters are controlled by DC voltages. These can be given fixed values by means of a set of potentiometers whose scales are calibrated in dB or cps. Since the voice and noise sources are essentially flat, the formant levels are uniquely determined by the level controls (i.e. independent of frequency). Hence a relative calibration of each formant level control in dB is possible.



Modulator circuit

Fig.8

There are at present 20 parameters which can be controlled in this way, viz. voice source frequency (F_o) and gate (A_o), noise source gates (A_{c1}, A_{c2}), resonant frequencies of nine formants, and levels of seven formants (F_6 and F_7 sharing the A_{c2} control at present). Although the possibility of controlling all of these parameters independently was introduced in order to make the synthesizer versatile, it is clear that they need not all be varied in order to synthesize some particular stretch of speech. Parameters that are not varied within a stimulus can be given a fixed value, while the others are controlled externally. Moreover, each formant can be switched off separately, and for measuring purposes the voice and noise sources can likewise be disconnected from the system.

A dynamic control for the synthesis of connected speech is obtained by means of a function generator constructed according to the principles outlined in Rischel (1969). Since the said paper gives a rather detailed account of the strategy, a brief presentation may suffice here. In its present shape the function generator supplies 16 time-varying voltages (parameters). Each of these varies in a piecewise linear fashion (though with a slight smoothing). The parameter values are specified in successive steps, 20 in total. For each such step there is a column of potentiometers, one for each parameter. Some parameters are used constantly for specific purposes (e.g. A_o, F_1, etc.); their scales are calibrated in cps or dB, so that the parameter values can be set directly. Others are left open for varying use. For each of the 20 successive steps the duration can be varied in the range 5-100 ms. There is another temporal parameter, viz. transition time, which specifies the duration of a linear transition from the parameter value of the preceding step to the value of the step under consideration. If the transition time coincides with the duration of the step we get a constant rate of change (or invariance if the two

steps have similar settings). If, however, the transition time is made shorter, we get a ramp followed by a steady-state portion, the latter increasing in duration as the former is shortened. In this way the duration of transitions can be varied within a synthesized item without affecting its overall duration. (Similarly, the durations of the steps can be increased without affecting the transitions). If two formants have different transition times (e.g. if F_2 moves after F_1 has fulfilled its transition), a faithful reproduction of the transitional stage requires two steps, the one transition being completed in one step, whereas the other takes two.

The stability of the voltage levels produced by the function generator is good. For precision work the settings are controlled by means of external measuring apparatus, the function generator being "locked" to keep the values of each step as long as desired.

The main limitation of the function generator is that there is a limit to the amount of speech that can be synthesized in one sweep. The theoretical maximum is 2 seconds, the practical limit being determined by the precision with which the user wants to reproduce short-time temporal variations. The synthesizer is intended primarily for synthesis of short stimuli which are to be varied systematically with respect to one or several parameters, i.e. vowels, syllables, single words or very short phrases.

In future there will be another option, viz. to control the synthesizer via a computer, which will make it possible to synthesize longer stretches of connected speech.

4. Operation and applications

As mentioned above, the synthesizer can be set according to spectrographic evidence. A useful shortcut can be obtained, however, by having a set of standard sound types with pre-

calculated data. Such a set of numerical data on vowels is being generated by Mr. Peter Holtse.

The synthesizer has been in use for some time for the synthesis of speech-like stimuli of various kinds, as required by researchers inside or outside the Institute of Phonetics. It is presently being used particularly for research on the discrimination of vowels. There has been a set of pilot experiments concerned with voice source characteristics, including dynamic control of the waveshape of the voice source, and shaping of the low frequency region by means of filters. Such experiments will be taken up again in the near future.

Acknowledgements

The speech synthesizer was financed by grants from the State Research Foundation for the Technical Sciences (Statens teknisk-videnskabelige Fond).

References

Fant, Gunnar 1959: Acoustic Analysis and Synthesis of Speech with Applications to Swedish, Ericsson Technics No. 1 (Stockholm).

Lystlund, Svend-Erik and Jørgen Rischel 1968: "Speech synthesizer", ARIPUC 2/1967, p. 34.

Rischel, Jørgen 1967: "Instrumentation for vowel synthesis", ARIPUC 1/1966, p. 15-21.

Rischel, Jørgen 1969: "Constructional work on a function generator for speech synthesis", ARIPUC 3/1968, p. 17-32.