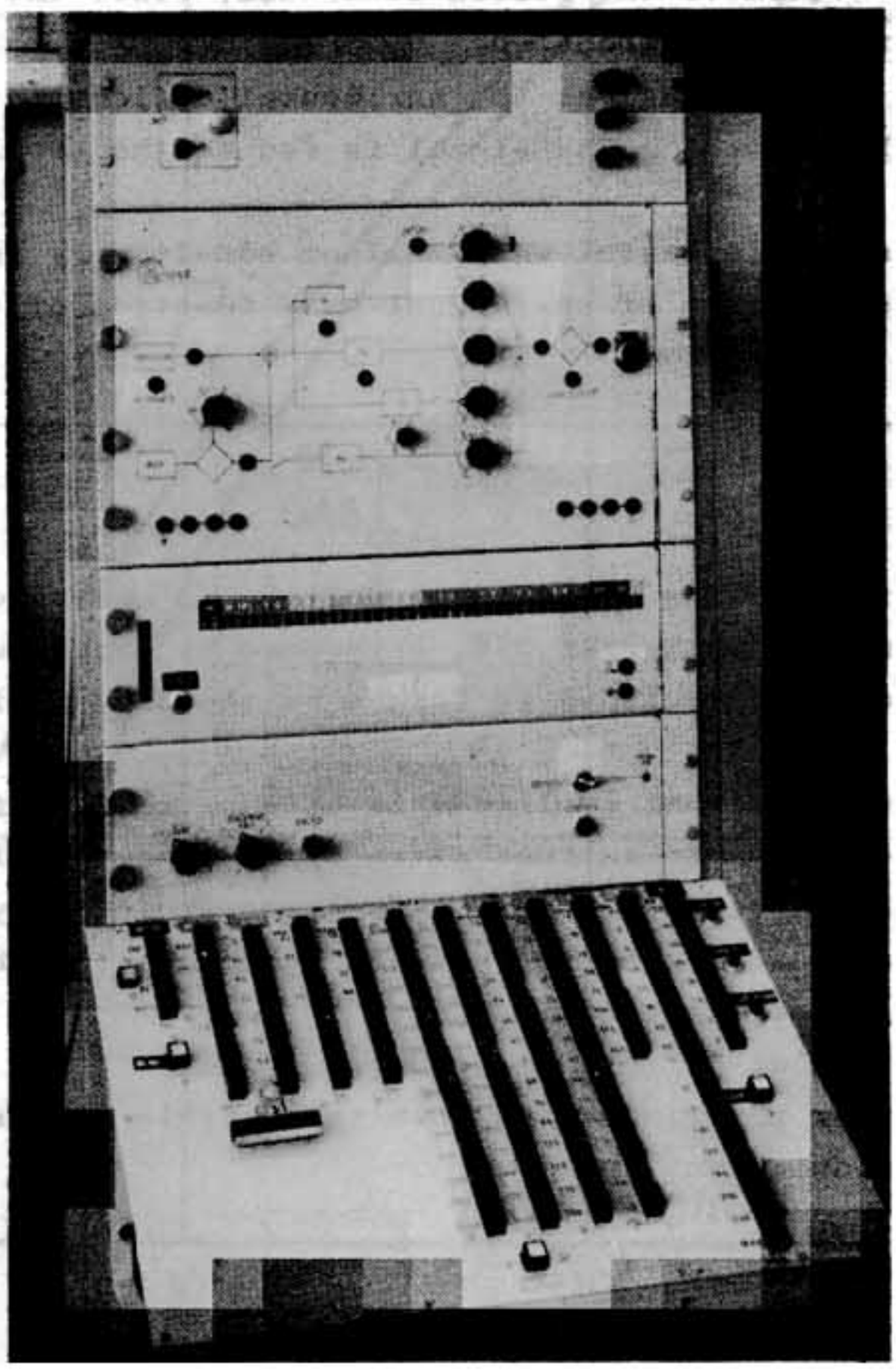
IPOVOX II: A Speech Synthesizer

L.F. Willems



Introduction

With our previous speech synthesizing machine - the IPOVOX I - we were able to produce simple words by placing in succession elementary segments having the size of a phoneme or part of it. The spectral content of the segments, however, was constant. In designing the IPOVOX II emphasis was laid on having formant transitions and on the ease of handling the machine.

General description

A parallel three-formant synthesizer was chosen for the voiced speech segments and a separate channel for unvoiced consonants. If in a segment the formant locations of the first two formants are different from those in the previous segment, then these new formant locations are reached in a smooth way. The time constant of this formant transition is electronically variable and the formant transition is triggered at the beginning of a segment.

In general, all parameters needed for the synthesis are stored in a digital memory, from which analogue circuits are controlled via digital-to-analogue converters or diode decoding networks. This flip-flop memory contains information for 5 segments and each segment occupies 44 bits. Information for the segments can be put into the memory one after another via a read-in desk.

Block diagram

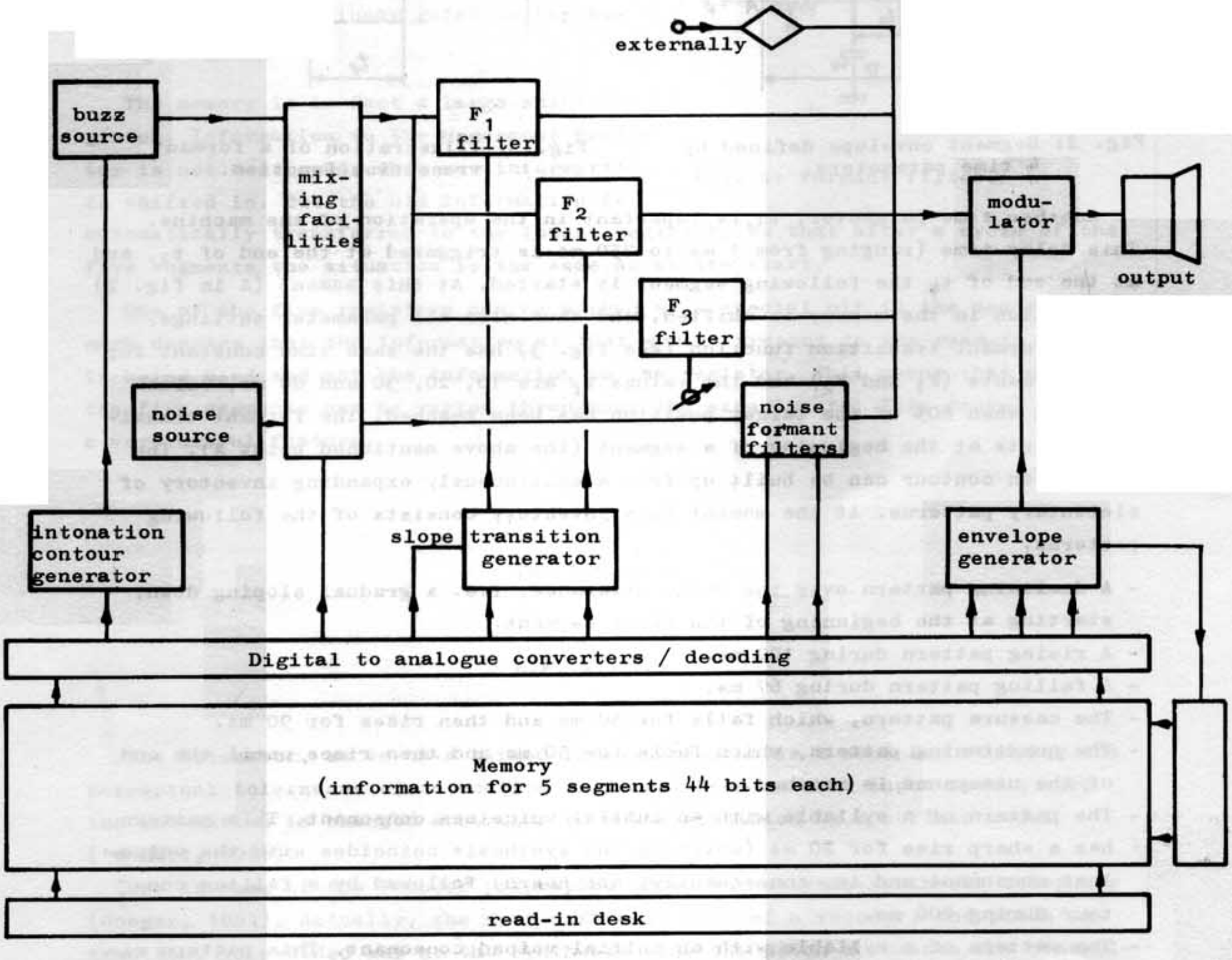


Fig. 2: Block diagram of IPOVOX II

The output waveform of the buzz source is approximately a sawtooth one. The first formant filter has a constant bandwidth of 50 Hz and the resonant frequency can be set at 16 discrete values between 200 and 1,200 Hz. The second formant filter has approximately a constant Q of 8 and the resonant frequency ranges in 16 steps from 700 to 4,200 Hz. The third formant filter can be switched on or off the circuit and the resonant frequency can be set manually between 1,000 and 5,000 Hz with a constant bandwidth of 150 Hz. The noise formant filters have fixed centre frequencies of 2,200, 2,100, 4,500 and 5,500 Hz. They can be switched into the circuit via a diode decoding network and electronic gates.

Several control voltages are required: for the envelope, the formant transitions and the intonation contours. In general, they are produced by ramp generators. The envelope function is described by three parameters (see fig. 2).

t_1 : the rise time can be chosen in 16 steps from 2,5 ms to 150 ms;
t_2 : the duration (as we define it) ranges from 10 ms to 200 ms in 16 steps;
t_3 : the decay time variable in 8 steps between 3 ms and 150 ms.

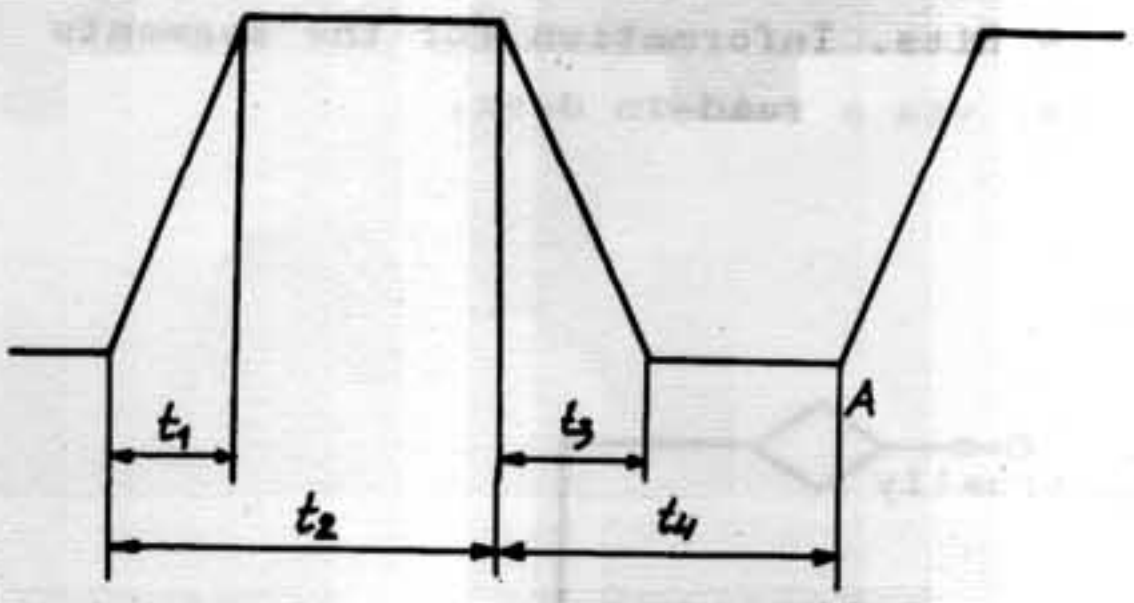


Fig. 2: Segment envelope defined by 4 time parameters.

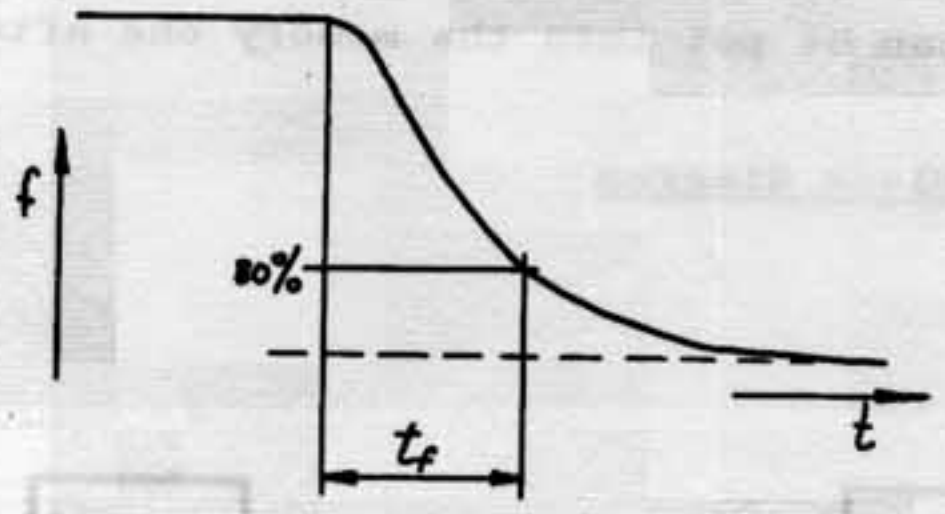


Fig. 3: Illustration of a formant transition function.

Another time parameter, t_4, is important in the operation of the machine. This delay time (ranging from 1 ms to 250 ms is triggered at the end of t_2, and at the end of t_4 the following segment is started. At this moment (A in fig. 2) information in the memory is shifted, and thus also all parameter settings.

The formant transition function (see fig. 3) has the same time constant for both formants (F_1 and F_2) and the values t_f are 10, 20, 30 and 80 ms, values obtained when 80% of the target position has been reached. The formant transition starts at the beginning of a segment (the above mentioned point A). The intonation contour can be built up from a continuously expanding inventory of elementary patterns. At the moment this inventory consists of the following patterns.

- A declining pattern over the whole utterance, i.e. a gradual sloping down, starting at the beginning of the first segment.
- A rising pattern during 100 ms.
- A falling pattern during 67 ms.
- The caesura pattern, which falls for 50 ms and then rises for 90 ms.
- The questioning pattern, which falls for 50 ms and then rises until the end of the utterance is reached.
- The pattern of a syllable with an initial voiceless consonant. This pattern has a sharp rise for 20 ms (which in the synthesis coincides with the voiceless consonant and is, consequently, not heard) followed by a falling contour during 200 ms.
- The pattern of a syllable with an initial voiced consonant. This pattern rises for 100 ms and falls for 150 ms.

All these patterns except the first can be triggered at 8 points to choose in the $t_2 + t_4$ interval. In the block diagram, one block has not yet been referred to, viz. mixing facilities. There are several possibilities as regards the choice of the source and the decision which path in the circuit should be used.

- Periodic source to the formant filters.
- Periodic source to the first formant filter only.
- Noise source to the formant filters (F_1, F_2 and F_3).
- Noise source to the noise formant filters.
- Noise multiplied by the periodic waveform to the formant filters (voiced fricatives).
- No source connected to the synthesizer, but an external signal can be fed into the envelope multiplier.

Circuit details

The resonant frequency of the formant filters is controlled by a voltage This is accomplished by a multiplier and an inductance in the feedback loop For the resonant frequency of this circuit one finds

$$fr = \frac{1}{2\pi}\sqrt{\frac{1 + A}{LC}}$$

By connecting two multipliers in cascade one has an almost linear relationship between fr and A.

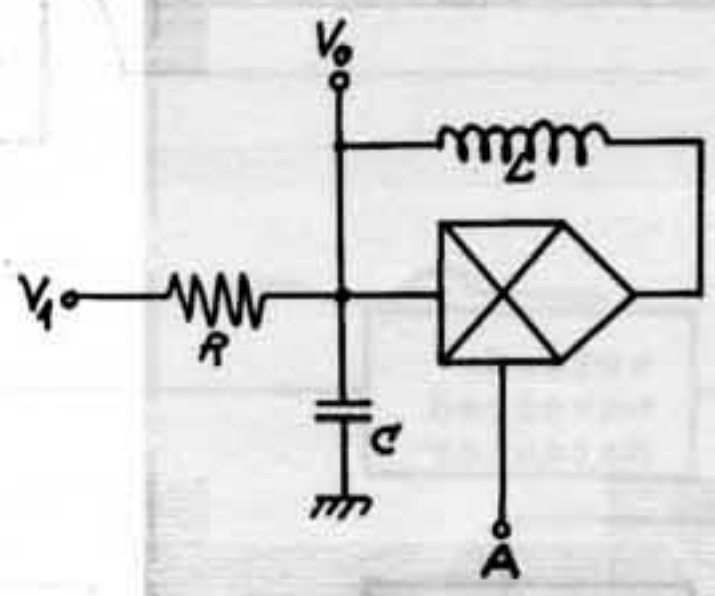


Fig. 4: Formant filter.

The memory is in fact a large shift register. Information in the uppermost register is not destroyed, when new information is shifted in. But the old information is automatically transferred to the lowest register, so that after a cycle of the five segments the situation is the same as at the start.

One of the five registers can be marked by a special bit in the memory. This mark denotes that the information at that moment present in the read-in desk is being used and not the information in the register. This means that one of the five segments can be varied throughout the experiments. This proved to be a very useful feature.