A NINETEEN-CHANNEL FILTER BANK SPECTRUM ANALYZER FOR A SPEECH RECOGNITION SYSTEM

I. INTRODUCTION

The goal of this NASA Research Project on Voice Controlled Computing Devices is to produce a computer system which will be capable of adapting to the voice characteristics of an individual. It should then be able to recognize correctly any of a limited set of messages spoken by this individual. The prototype system, which we have constructed is described in NASA Scientific Report #1, (Contract NAS12-138 BBN 11254). In this description, the input system which we use as the base for extracting properties for speech recognition is discussed briefly. This report describes in detail that input system. Essentially the input system is a bank of filters which span that part of the audio frequency range significant in speech. The integrated output of one of these bandpass filters represents the average energy in that portion of the spectrum over some time period. This report not only contains the detailed design of this filter bank, but also presents the criteria which led to the selection of the frequency and time domain characteristics of this spectrum analyzer.

TABLE OF CONTENTS

Job No. 11254

A NINETEEN-CHANNEL FILTER BANK SPECTRUM ANALYZER FOR A SPEECH RECOGNITION SYSTEM

Contract No. NAS 12-138

Scientific Report #2

Kenneth N. Stevens

Gottfried von Bismarck

5 July 1967

Submitted to:

National Aeronautics and Space Administration
Electronics Research Center
575 Technology Square
Cambridge, Massachusetts 02139

Attn: Mr. Daniel Kelleher, X279

A NINETEEN-CHANNEL FILTER BANK SPECTRUM ANALYZER FOR A SPEECH RECOGNITION SYSTEM

I. INTRODUCTION

The goal of this NASA Research Project on Voice Controlled Computing Devices is to produce a computer system which will be capable of adapting to the voice characteristics of an individual. It should then be able to recognize correctly any of a limited set of messages spoken by this individual. The prototype system, which we have constructed is described in NASA Scientific Report #1, (Contract NAS12-138 BBN 11254). In this description, the input system which we use as the base for extracting properties for speech recognition is discussed briefly. This report describes in detail that input system. Essentially the input system is a bank of filters which span that part of the audio frequency range significant in speech. The integrated output of one of these bandpass filters represents the average energy in that portion of the spectrum over some time period. This report not only contains the detailed design of this filter bank, but also presents the criteria which led to the selection of the frequency and time domain characteristics of this spectrum analyzer.

II. DESIGN CONSIDERATIONS

Selection of Center Frequencies and Bandwidths for the Band-Pass-Filters

Several considerations governed the selection of the values that were used for the center frequencies and bandwidths of the filters. First, it was necessary to limit the number of filter to 20 or fewer since this was the number of parallel, multiplexed analog inputs that could be connected to the PDP-1 computer. It was decided to use 19 filters, the 20th channel being reserved for another input signal that might be used in the future, such as the output from a pitch extractor or from some other analog processing device.

In the frequency range up to about 3000 Hz, the filters were arranged to be spaced uniformly and to have equal bandwidths. This is the approximate frequency range encompassed by the first three resonances of the vocal tract (for adult male speakers), and there is evidence that all of these resonances play an important role in the perception and recognition of speech. There is little justification for assigning significantly more importance to one part of this frequency range than to another, and consequently it is reasonable to space the filters uniformly throughout this range.

During normal utterances of male speakers, the fundamental frequency rarely exceeds 180 Hz, and is usually in the range 80-150 Hz. For voiced sounds, it is known that the shape of

the spectrum envelope carries most of the important information, particularly with regard to segmental features, i.e., with regard to the identity of phonetic segments, whereas the fundamental frequency (frequency of vocal-cord vibration) is of less importance, at least in English.

Thus a spectrum analyzer for such sounds should ideally provide a representation of the spectrum envelope, and this representation should be relatively uninfluenced by the fundamental frequency. These considerations led to the selection of 360 Hz as an appropriate filter bandwidth in the frequency range up to 3000 Hz. With such a bandwidth there are always at least two harmonics of the fundamental within a given filter, and probably no more than four such harmonics. Thus, changes in the pattern of filter outputs for a voiced sound due to changes in fundamental frequency should be small.

In order to provide an adequate representation of the spectrum envelope, it was decided to space the filters every 180 Hz in the frequency range up to 3000 Hz. Assuming the filter bank is designed with Lerner filters, [1,2] such an arrangement would permit poles to be shared between the odd-numbered filters and between the even-numbered ones.

If we select a lower-frequency cut-off of 80 Hz for the first filter, then the first 15 filters cover the frequency range from 80-2960 Hz, as shown in Table 1. The frequency range above 3000 Hz is of importance primarily for carrying information about the noise-like components of speech--the frication noise for stop and fricature consonants. For the

Filter No.	Lower Cut-Off	Higher Cut-Off	Center Freq.	Band Width	No. of Poles	Cumulative No. of Poles
	cps	cps	cps	cps		
1	100	440	260	360	4	4
2	260	620	440	360	4	8
3	440	800	620	360	4	10
4	620	980	800	360	4	12
5	800	1160	980	360	4	14
6	980	1340	1160	360	4	16
7	1160	1520	1340	360	4	18
8	1340	1700	1520	360	4	20
9	1520	1880	1700	360	4	22
10	1700	2060	1880	360	4	24
11	1880	2240	2060	360	4	26
12	2060	2420	2240	360	4	28
13	2240	2600	2420	360	4	30
14	2420	2780	2600	360	4	32
15	2600	2960	2780	360	4	34
16	2960	3560	3260	600	5	37
17	3560	4400	3980	840	6	41
18	4400	5480	4940	1080	7	46
19	5480	6560	6020	1080	7	51

TABLE 1. List of Filter Center Frequencies and Bandwidths

sh-like sounds, there is usually a spectral energy peak in the range 2500-3500 Hz; for s-like sounds the major spectral energy concentration may be in the range 3500-4500 Hz; and for f and th there may be useful information up to 6000 Hz or higher. It is apparent, therefore, that fine frequency resolution in this frequency range is not required. Such a conclusion is consistent with what is known regarding the relatively poor frequency discriminating capabilities of the auditory mechanism at high frequencies. Thus the high frequency range from 2960 to 6560 Hz was covered by four fairly broad filters, with no overlap in frequency range between filters.

Since it is important in speech analysis to detect the occurrence of rapid changes in events at different frequencies, to within at least a few msec, each bandpass filter should exhibit approximately the same time delay. Lerner filters[1,2] can be designed to provide the required time-delay characteristics, and, in addition, have several other desirable features, as discussed later in this report.

Selection of the Low-Pass Filter Characteristics

The selection of an averaging time for the low-pass filters following the rectifiers is determined by several factors. First, the averaging time should be sufficiently long that there are not appreciable fluctuations in the outputs of the channels within a period of the fundamental frequency. Since the outputs are sampled periodically, sampling is not synchronous with the fundamental period, and any fluctuation in output would represent an artifact. On the other hand, the averaging time should be sufficiently short that rapid changes in level associated with the onset of stop and nasal consonants can be detected. For such consonants, these changes occur within a few msec, whereas for other classes of sounds, changes occupy an interval of several tens of msec. These considerations led to selection of an averaging time of 10-20 msec. The weighting function (i.e. the impulse response) associated with each low-pass filter should have as short a "tail" as possible so that only events over the 10-20 msec interval are averaged.

III. TRANSMISSION CHARACTERISTICS OF THE SPECTRUM ANALYZER

Input Stage

The functional block diagram of the spectrum analyzer is shown in Fig. 1. The input stage contains an input amplifier, a pre-emphasis network, and two driver amplifiers feeding the bandpass filters. The input and driver amplifiers exhibit flat frequency responses and negligible harmonic distortion over the frequency range of interest (80 Hz to 7 KHz). The pre-emphasis network is a high-pass filter whose frequency response resembles the inverse of the long-term rms speech spectrum of male voices. Since the long-term rms level of the speech signal is relatively low at higher frequencies, the emphasis provided by the high-pass filter equalizes the spectrum of the input signal to the bandpass filter, and thus allows full utilization of the dynamic range of the high frequency filters and detectors. The slope of the high-pass filter (+6 dB/octave, 3 dB up at 740 Hz) is somewhat smaller than the slope of the inverse long-term rms speech spectrum (+6 to +12 dB/octave, 5 dB up at 740 Hz) so that single phones (e.g. the fricatives s and sh) with relatively high energy at the high frequencies do not overload these filter channels.

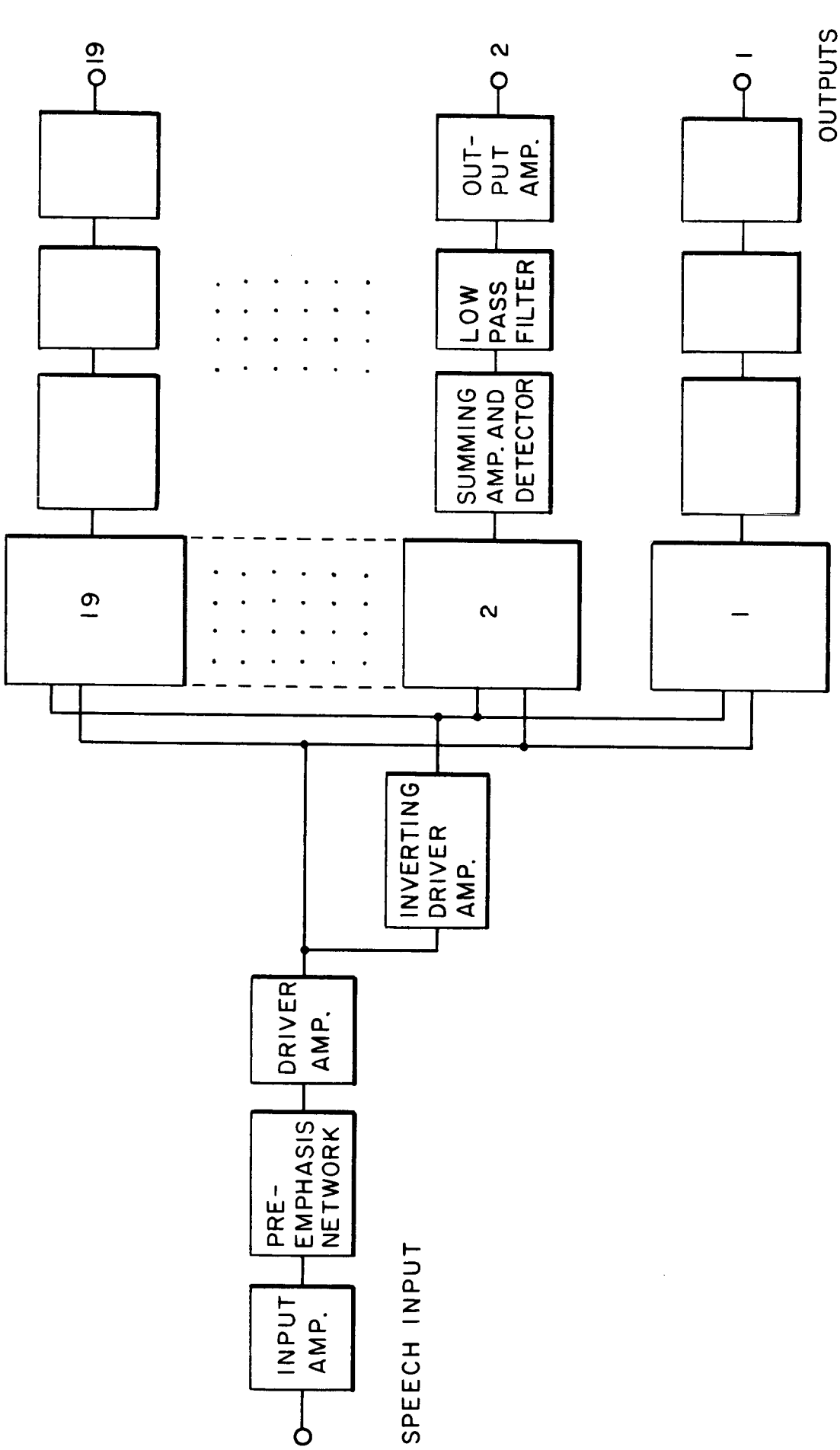

FIG. 1 FUNCTIONAL BLOCK DIAGRAM OF FILTER BANK

Frequency Response of the Bandfilters

Lerner filters [1,2] were chosen to realize the filter specifications described under Section II since these filters exhibit the following characteristics:

1) By sharing two poles between adjacent filters, a bank of n-pole filters can be obtained with only (n-2) additional poles for each filter after the first. In this manner, sharp skirted frequency responses can be obtained with a fewer number of poles than would be necessary for a Bessel, Chebyshev or Butterworth filter with comparable response.

2) The frequency responses of individual filters exhibit nearly perfect symmetry around their center frequencies.

3) The phase response is linear over the passband of each filter.

4) The slope of the phase response, i.e., the delay time of the bandfilter may be kept constant for filters of different center frequencies and bandwidths. Therefore, spectral components of an input signal can appear simultaneously at the outputs of the filter channels.

These characteristics of Lerner filters are illustrated in Fig. 2 by the frequency, phase and inpulse responses, $|H(f)|$, $\theta(f)$ and $h(t)$, of filters no. 5 and 7.

The overall frequency responses of all 19 filters and detectors is presented in Fig. 3. Note that responses of pole-sharing filters have their crossover points close to -3 dB and bandwidths as originally given in Table 1. The center frequencies of the even-numbered filters are nearly equal to the crossover frequencies of the odd-numbered filters.

For attenuations to about -30 dB, the skirts of filters no. 16 through 19 are steeper than the skirts of the remaining filters due to the higher number of poles. For attenuations greater than -30 dB, the high-frequency skirts of filters no. 17, 18, 19 tend to flatten out slightly, most likely due to stray capacitances between circuit connections. This phenomenon was considered irrelevant since high frequency resolution of these filters was not desired (p. 5).

All frequency responses were calculated with the aid of a digital computer and found to agree closely with the measured responses (Fig. 3) except for the slight flattening of the filter skirts at high attenuation levels.

Rectifier and Low-Pass Filter Characteristics

All rectifiers consist of operational amplifier configurations with a linear dynamic range exceeding 60 dB.

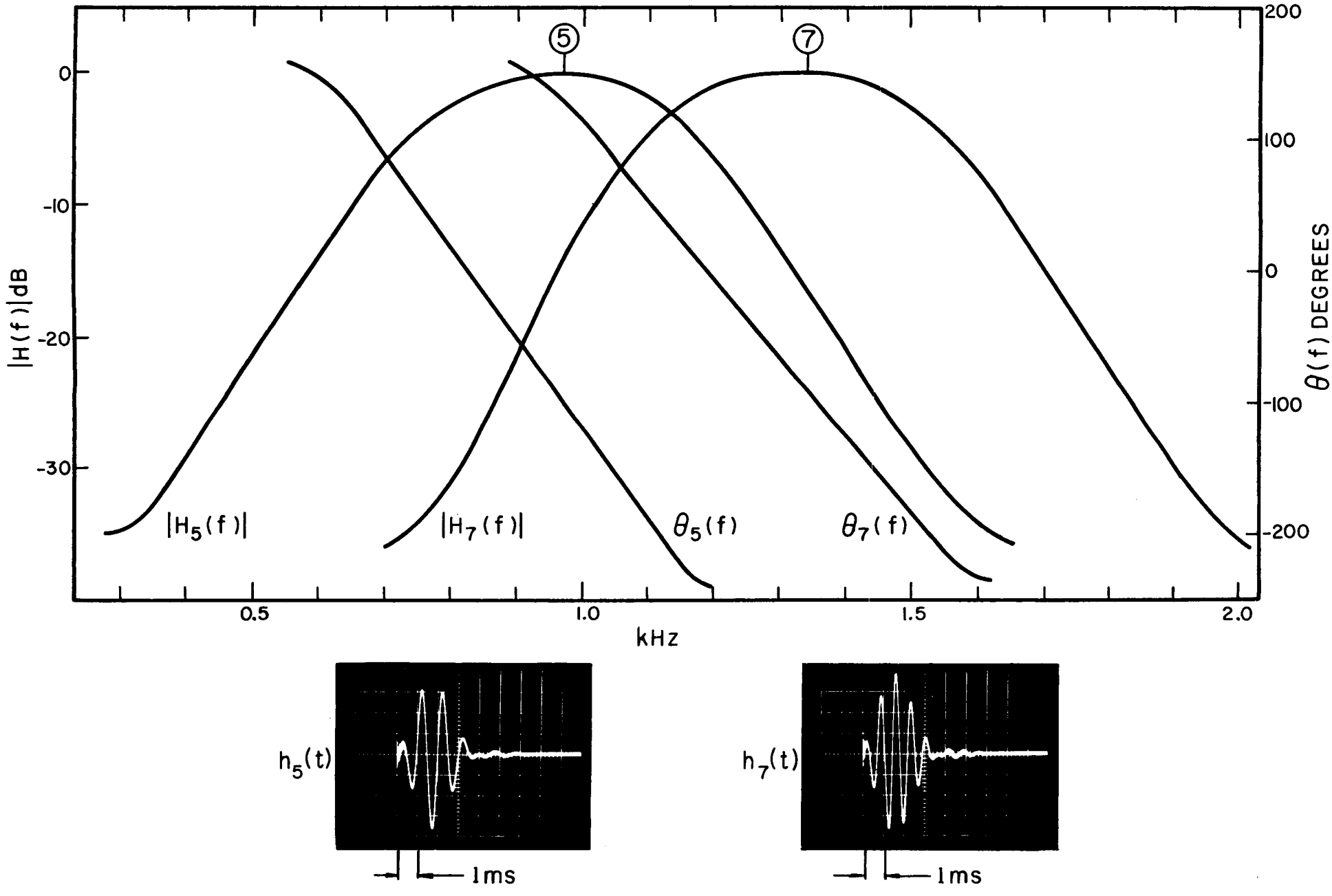

FIG. 2 FREQUENCY, PHASE AND IMPULSE RESPONSES OF FILTERS NO. 5 AND 7

JOB NO. 11254

BOLT BERANEK & NEWMAN INC

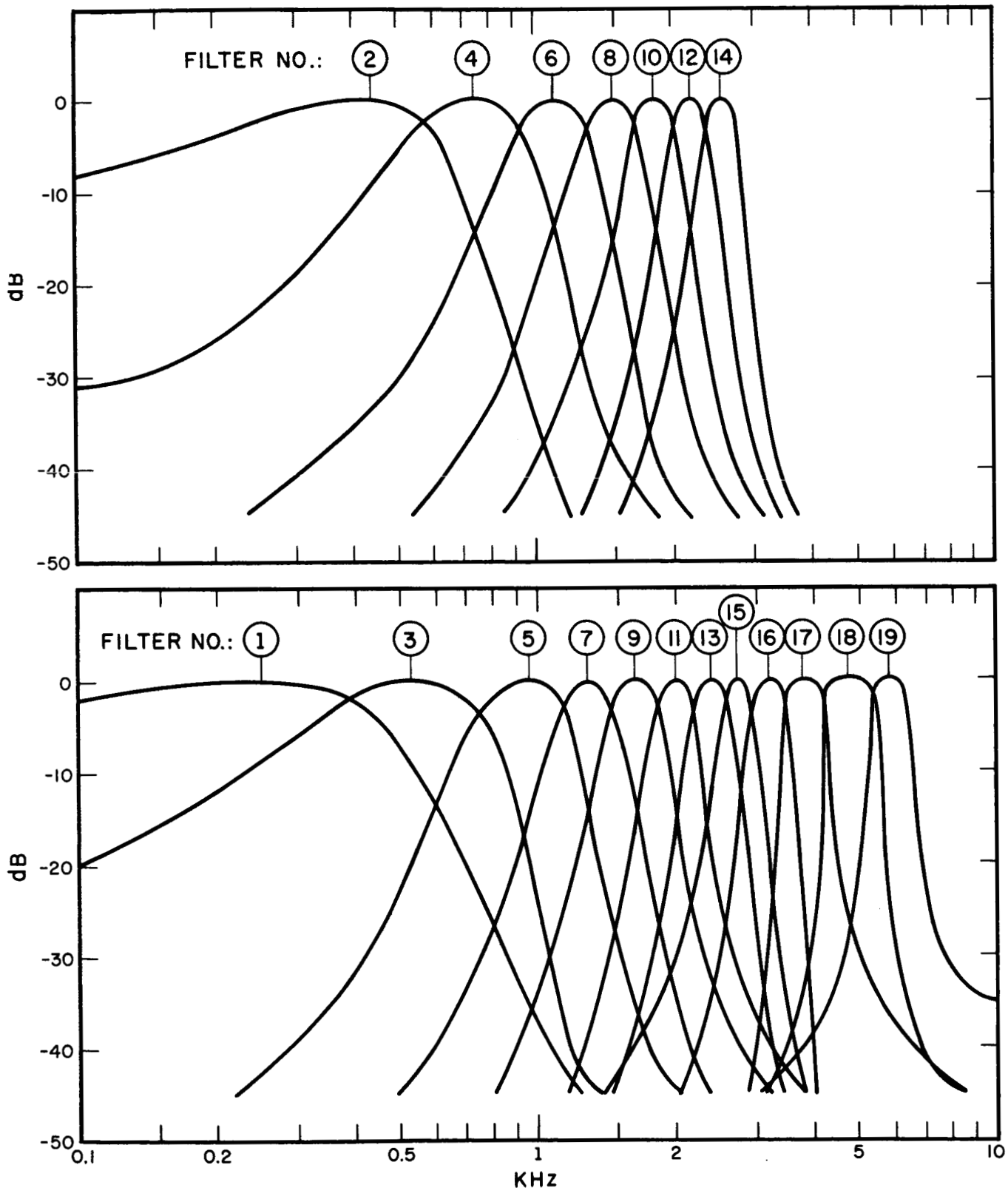

FIG. 3 MEASURED FREQUENCY RESPONSE OF FILTER BANK. RATIO OF DC OUTPUT TO AC INPUT IN dB

A four-pole Bessel filter was chosen to obtain the envelope, i.e. the approximate energy, of the signal from the rectifiers (Fig. 1). A similar design has been found useful in vocoder applications.[3] The Bessel filter provides a non-ringing impulse response, a "tail" that is relatively short, and an averaging time of 10-20 msec. These characteristics are illustrated in Fig. 4 which shows the frequency response and the impulse response of the low-pass filter.

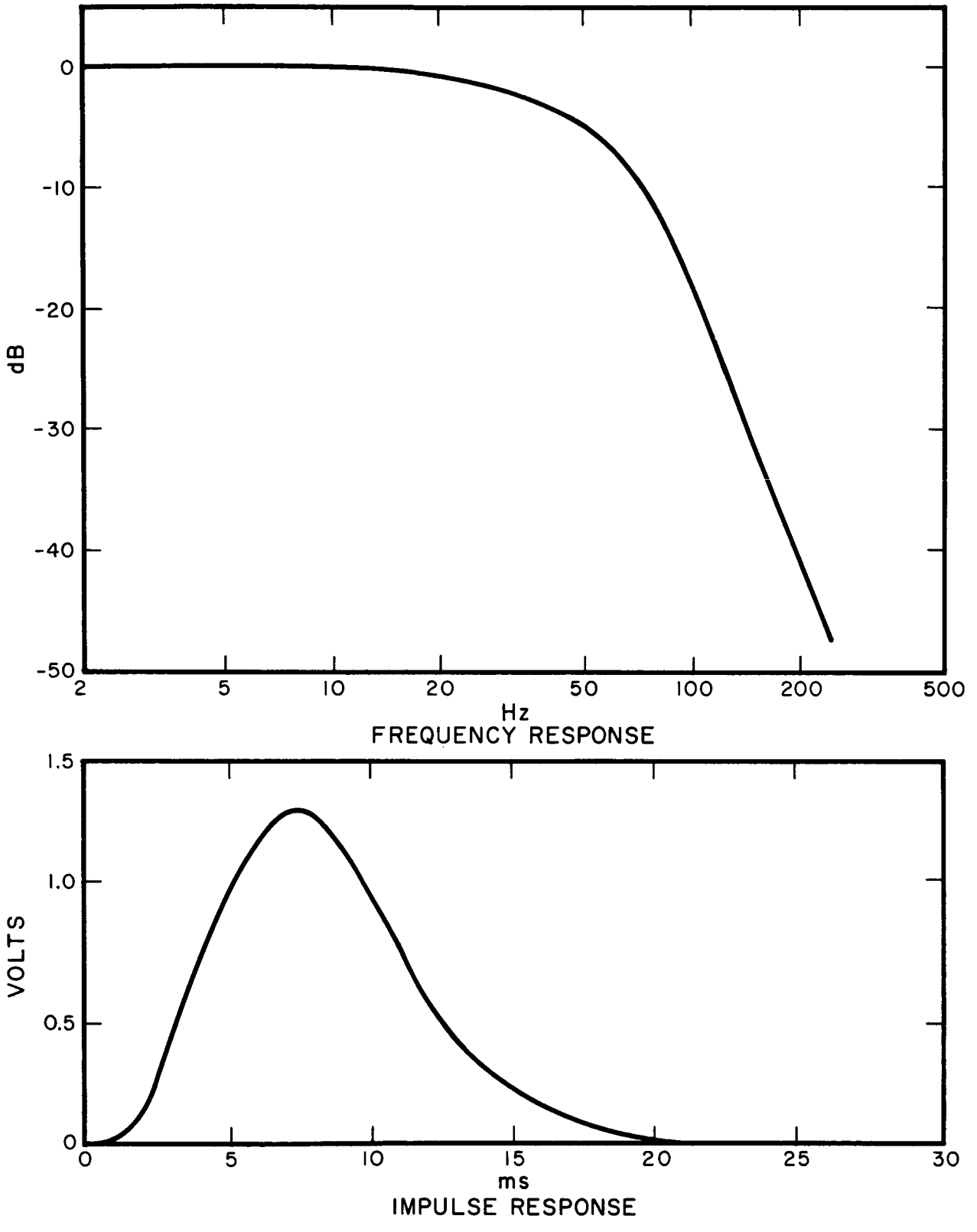

FIG. 4 LOW-PASS FILTER CHARACTERISTICS

IV. CIRCUITS

Input Stage

The circuit diagram of the input stage is shown in Fig. 5. The output from a conventional tape recorder is fed into an input amplifier (Type BBN-DE-200) whose variable feedback resistor provides the common gain adjustment for all filter channels. The gain adjustment procedure is facilitated by a VU-meter connected to the output of the input amplifier. By means of a separate potentiometer this VU-meter may be adjusted such that speech peaks exceeding +3VU will indicate clipping in the filter channels and/or overloading of the A/D converter.

The RL-network in the feedback path of the first driver amplifier provides the high frequency pre-emphasis. By means of a switch this network may be replaced by a resistor for gain calibration of the filter channels (p. 20).

Since the bandpass filter networks require inputs of opposite phase,[1,2] a second inverting driver amplifier is used as shown on the lower right of Fig. 5. A potentiometer in the feedback path provides for adjustable amplitude balance of the two input signals.

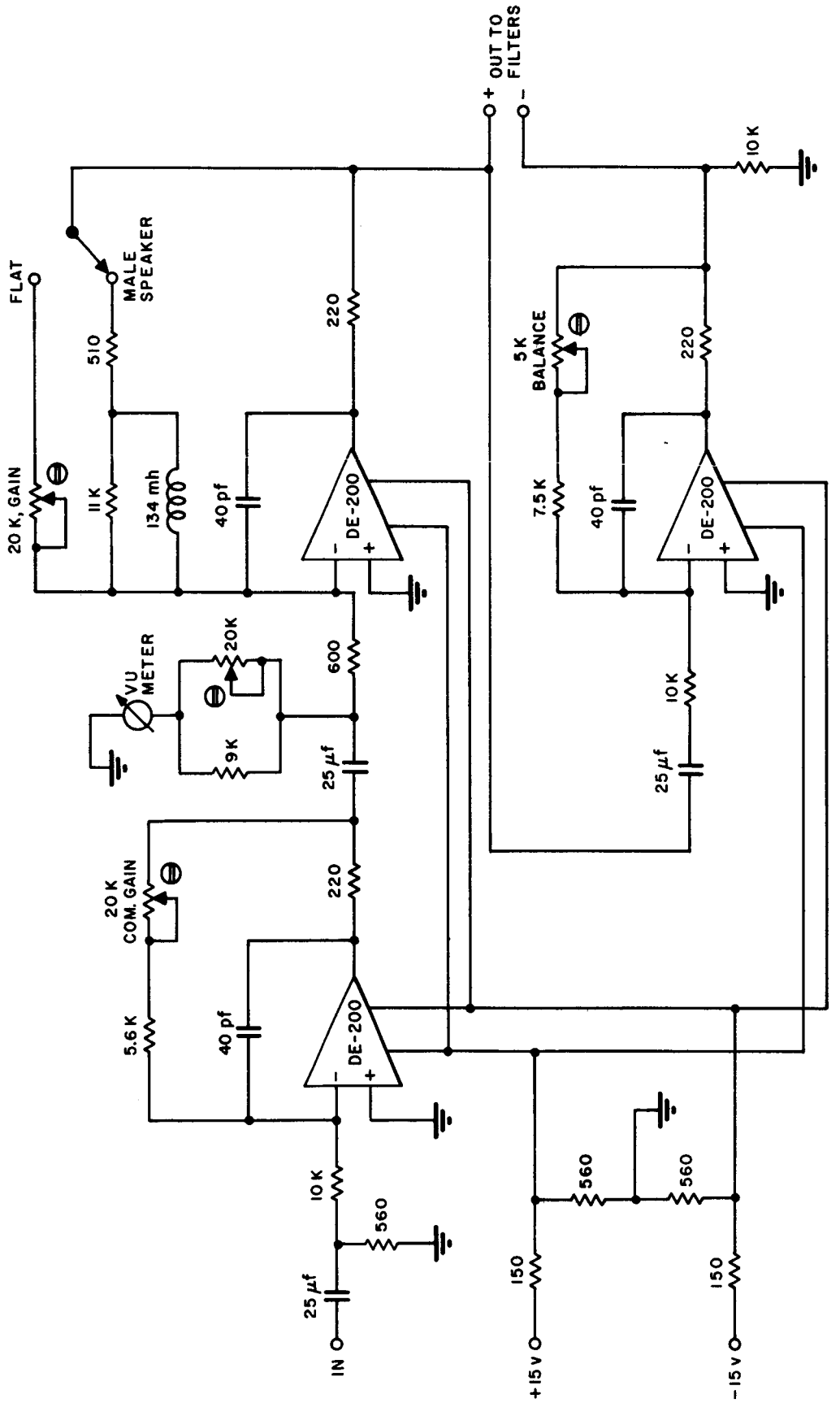

FIG. 5 INPUT STAGE

Band-Pass Filters

The pole pattern yielding the frequency responses of Fig. 3 is shown in Fig. 6. Poles associated with one filter are connected by converging lines. Each filter shares two poles with each neighboring filter.

According to the Lerner synthesis procedure, poles of an individual filter are spaced at equal intervals (240 Hz) in the pass-band and at half intervals (120 Hz) at the band edges. Since the frequency responses of neighboring filters have their 3 dB crossover points at a frequency mid-way between the shared poles, all 4-pole filters (no. 1-15) have bandwidths of 360 Hz. The number of the poles for the remaining filters are selected such that bandwidths as given in Table 1 result.

Each pole of Fig. 6 is realized by a series resonant circuit as shown in Fig. 7 for filters no. 1 through 15. The proper pole residue is achieved by resistor weighting of the currents from each resonant circuit into the summing node of an operational amplifier as described by Drouilhet and Goodman.[1] Identical inductor values were used for all resonant circuits and the capacitors were calculated by $C = 1/4\pi^2 Lf^2$, where f is the frequency of a pole. (Approximately 1% of the C value was realized by a trimmer capacitor to compensate for slight inaccuracies and variations of the filter components.) For a chosen value of L = 0.5 mh, R = 1.66 k was obtained from the relation $L/R' = 1/2\pi \cdot 360$, where $R' = 2/3$ R, the parallel resistance of R and 2R.

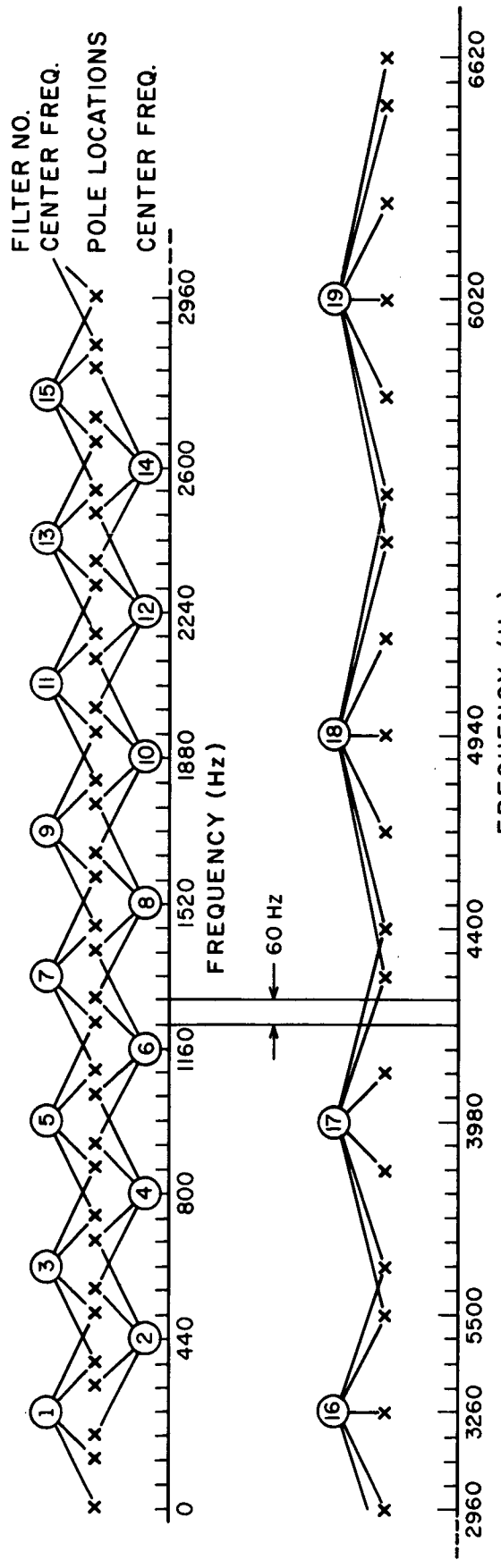

FIG. 6 POLE PATTERN OF BAND-PASS FILTERS

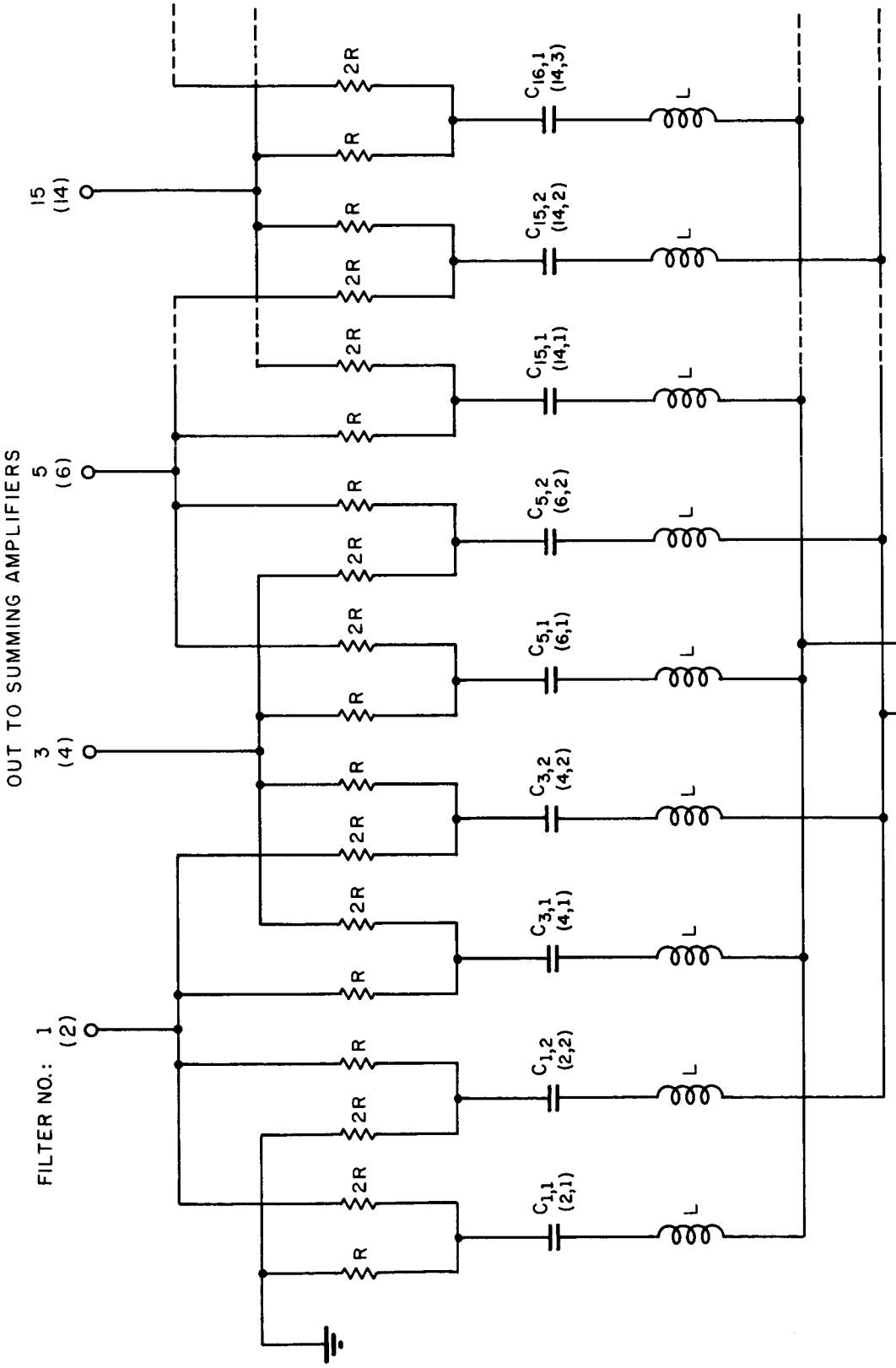

FIG. 7 BAND-PASS FILTERS NO. 1-15

Filters no. 16 though 19 were realized by the same circuit structure as shown in Fig. 7, with poles in the passband and at the band-edges. (Fig. 6) weighted by R and 2R respectively. The highest pole at 6620 Hz was tied over R to ground to simulate the input impedance of an additional neighboring summing amplifier as is also shown for the lowest pole in Fig. 6.

Rectifiers and Low-Pass Filters

The circuitry containing the summing amplifiers, rectifiers, low-pass filters and output amplifiers is shown in Fig. 8. (All operational amplifiers: Union Carbide type H6010). The circuit for filter channel no. 1 is shown in the upper portion of the figure. The summing amplifier is followed by a full wave rectifier circuit. Full wave rectification was necessary for filter channel no. 1 only, in order to reduce the ac-components passed by the low-pass filter of this channel. (At 80 Hz the rms level of the ac-component in the output of filter channel no. 1 is 35 dB below the dc level and decreases with a slope of 22 dB/octave.)

The circuit in the lower portion of Fig. 8 is used for filter channels no. 2-19. The first amplifier (1) sums the current components from the band-filter, (2) half-wave rectifies that signal and (3) provides a signal of adjustable amplitude to drive the low-pass filter. The gain of these amplifiers is adjusted such that all dc channel outputs are equal for white noise input to the spectrum analyzer (pre-emphasis network not used).

The output of the low-pass filters is fed into output amplifiers which provide low impedance sources to drive the A/D converter.

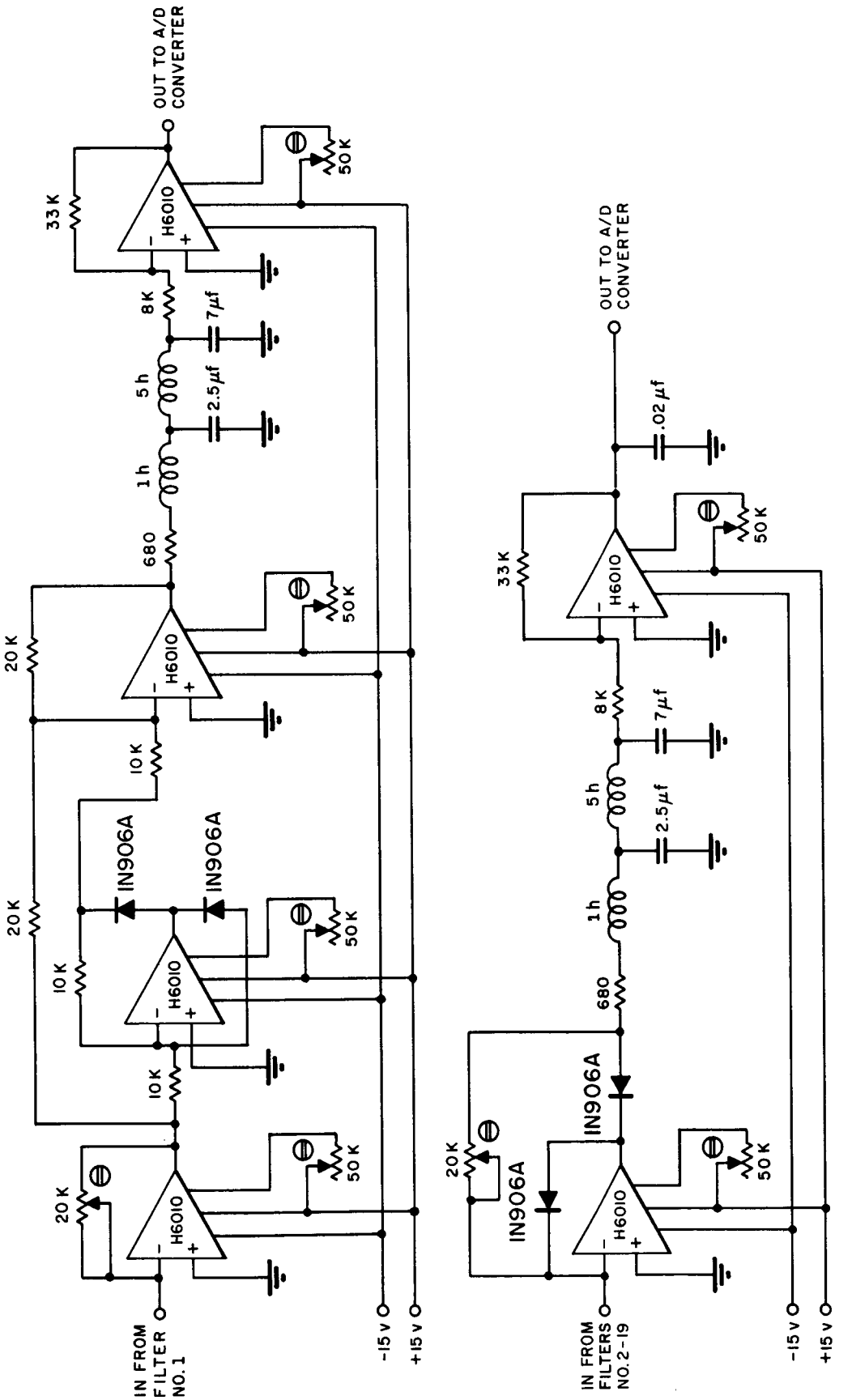

FIG. 8 SUMMING AMPLIFIERS, DETECTORS, LOW-PASS FILTERS AND OUTPUT AMPLIFIER

V. REFERENCES

1. P. R. Drouilhet, Jr., and L. M. Goodman, "Pole-Shared linear-phase band-pass filter bank." Proc. I.E.E.E. 54, 701-703 (1966).

2. R. M. Lerner, "Band-pass filters with linear phase," Proc. I.E.E.E. 52, 249-268 (1964).

3. J. Tierney, B. Gold, V. Sferrino, J. A. Dumanian, and E. Aho, "Channel vocoder with digital pitch extractor." J. Acoust. Soc. Am. 36, 1901-1905 (1964).