June 16, 1959 J. L. FLANAGAN 2,891,111

SPEECH ANALYSIS

Filed April 12, 1957 3 Sheets-Sheet 1

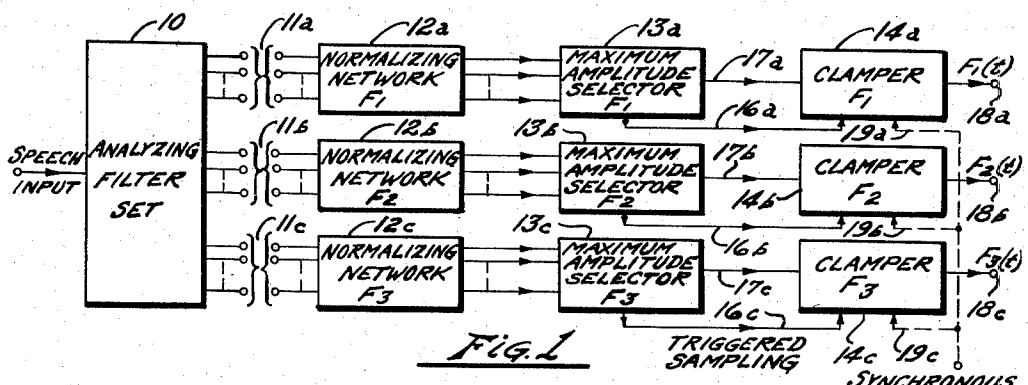


Fig. 1

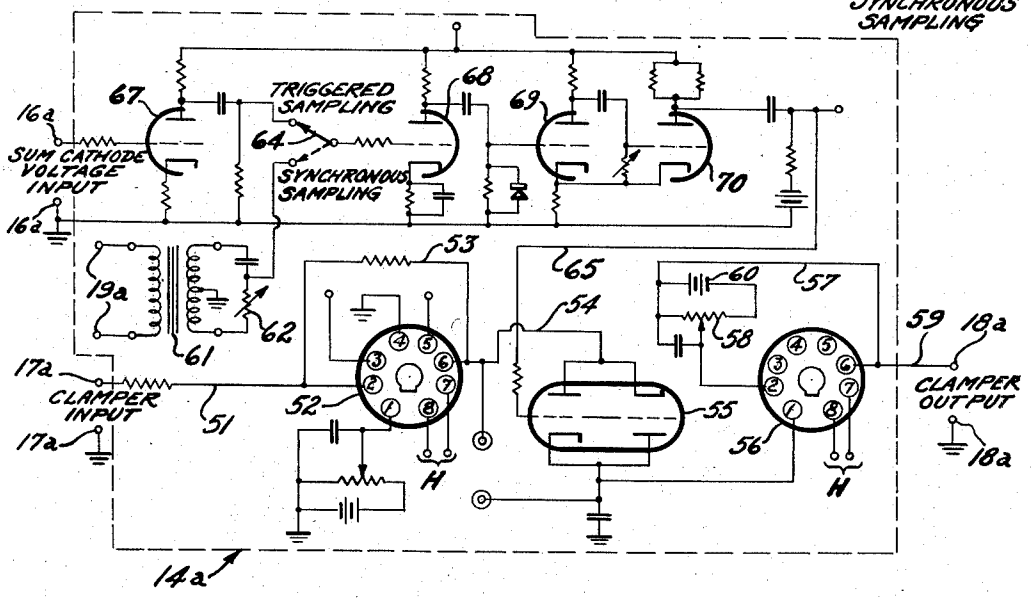


Fig. 3

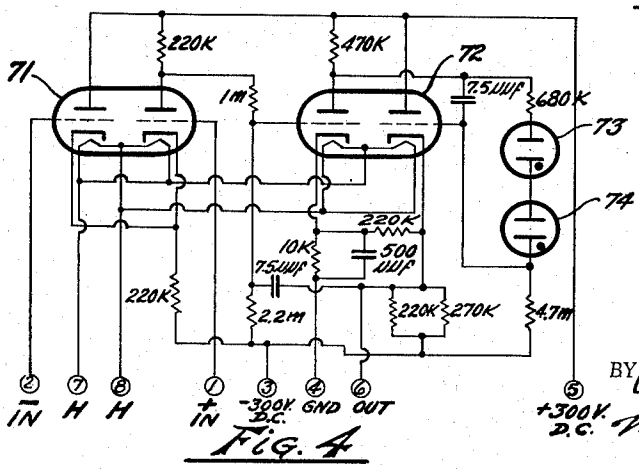


Fig. 4

INVENTOR.
JAMES LOTON FLANAGAN

BY Wade Koontz and Martin J. Finnegan
ATTORNEYS

June 16, 1959 J. L. FLANAGAN 2,891,111

SPEECH ANALYSIS

Filed April 12, 1957 3 Sheets-Sheet 2

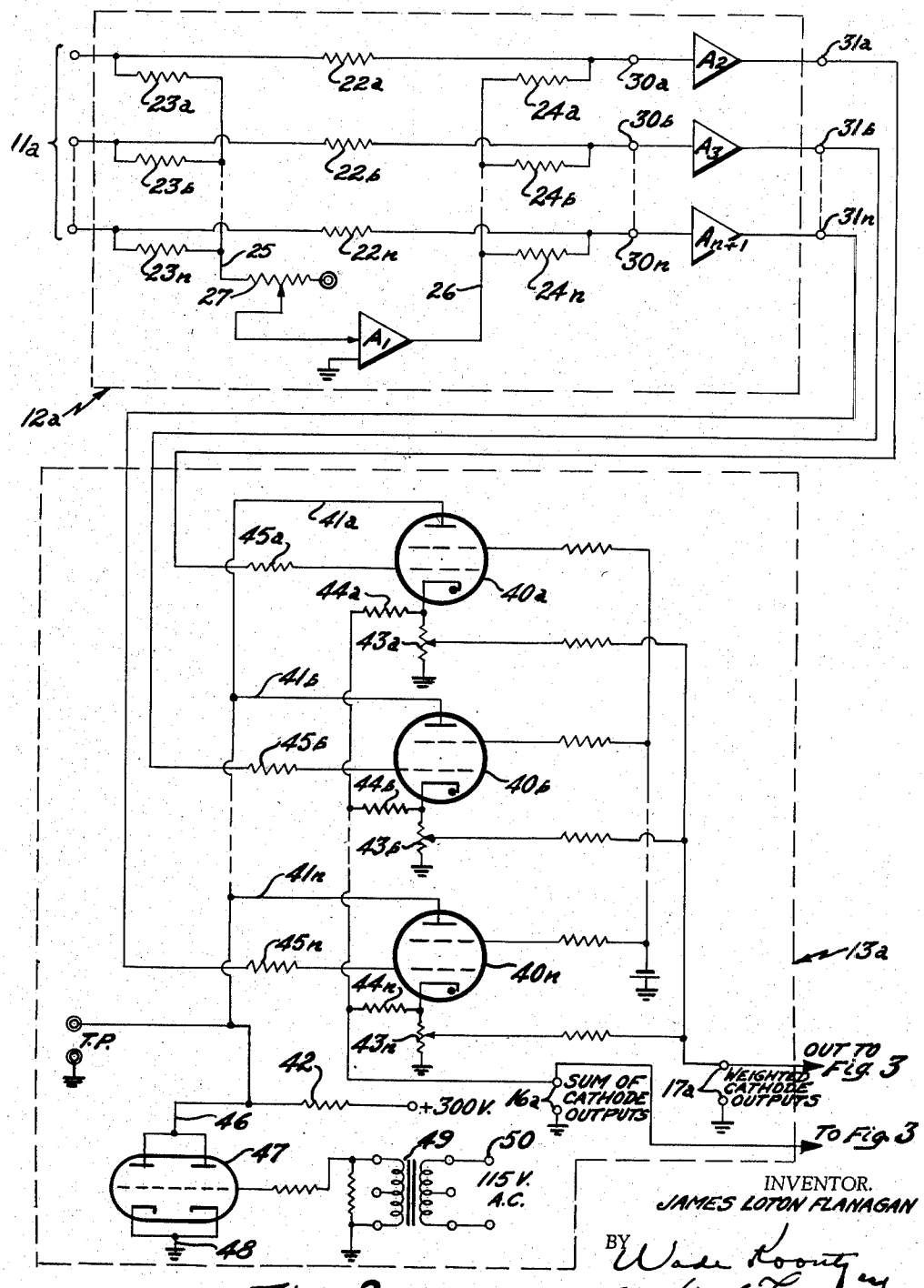


Fig. 2

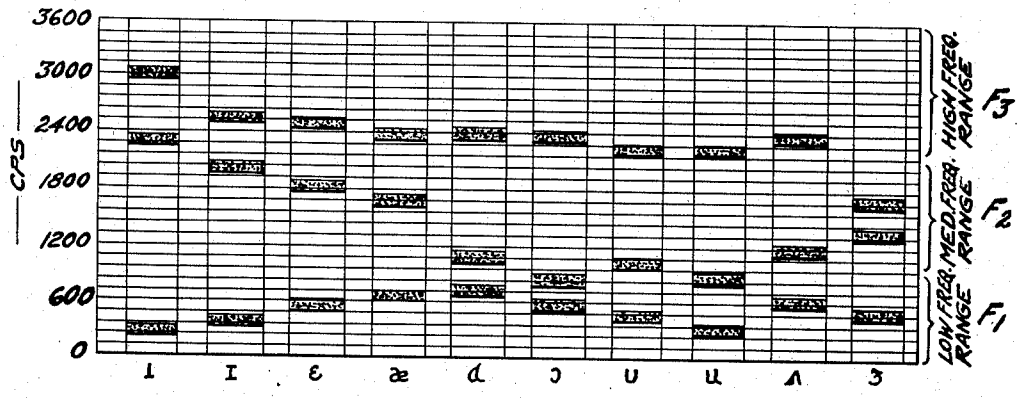
INVENTOR.
JAMES LOTON FLANAGAN

BY Wade Koontz and Martin J. Finnegan
ATTORNEYS

June 16, 1959 J. L. FLANAGAN 2,891,111

SPEECH ANALYSIS

Filed April 12, 1957 3 Sheets-Sheet 3



AVERAGE FORMANT FREQUENCIES (IN CPS) GENERATED BY SPOKEN ENGLISH VOWELS. (ADULT MALE SPEAKERS)

Fig. 5

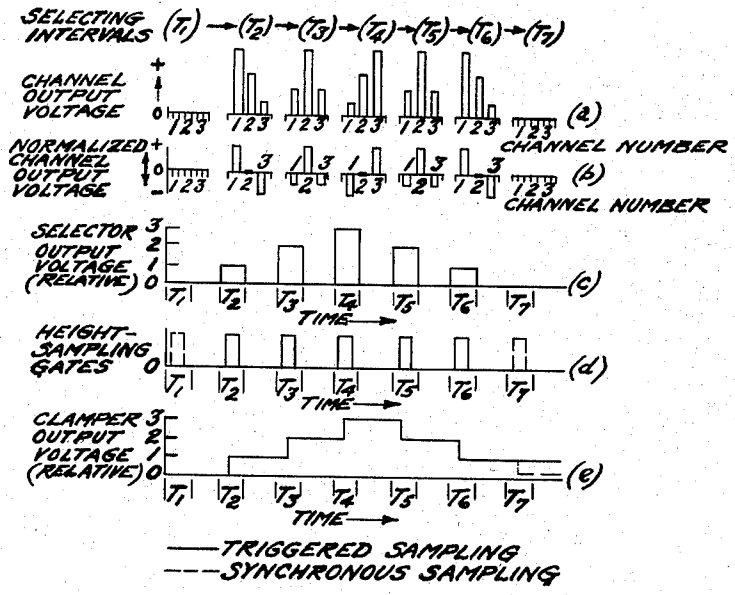


Fig. 6

INVENTOR.
JAMES LOTON FLANAGAN

BY Wade L. Pountney and Martin J. Finnegan
ATTORNEYS

United States Patent Office 2,891,111 Patented June 16, 1959

2,891,111

SPEECH ANALYSIS

James Loton Flanagan, Greenwood, Miss., assignor to the United States of America as represented by the Secretary of the Air Force

Application April 12, 1957, Serial No. 652,634

2 Claims. (Cl. 179—1)

(Granted under Title 35, U.S. Code (1952), sec. 266)

The invention described herein may be manufactured and used by or for the United States Government for governmental purposes without payment to me of any royalty thereon.

This invention relates to speech analysis, and particularly to the transmission of speech over a communication channel highly restricted in transmission bandwidth and capacity.

The invention is pertinent to a speech bandwidth-compression system of the analysis-synthesis type, in which the speech information is coded in terms of signals representing the major vocal resonances (or formants) and the nature of the excitation of the vocal tract, both as functions of time during the production of speech. The major vocal resonances, or formants, are manifested as maxima in the frequency spectrum of speech radiated by a talker. It is basic to the operation of such a speech compression system that the frequencies of these spectral maxima (i. e. the formant frequencies) be determined by automatic analysis as the speech is uttered, and that signals representing these frequencies be transmitted to the speech synthesizer in order that the speech may be reproduced with negligible time delay.

The conventional telephone channel, which is a "waveform" transmission system, requires a transmission bandwidth of the order of 3000 c.p.s. and a signal-to-noise ratio of about 30 db. The speech bandwidth compression system disclosed in my co-pending patent application Ser. No. 551,478, filed December 6, 1955, on the other hand, can function with a much narrower total transmission bandwidth of the order of only 50 c.p.s., and a signal-to-noise ratio of approximately 30 db.

The present application is also concerned with a speech bandwidth compression system which—like the system disclosed in my co-pending application above-identified—requires only a narrow transmission bandwidth. The system herein disclosed also operates to accept continuous speech at its input and to deliver at its output electrical signals, varying slowly with time, whose amplitudes represent the frequencies of the first three major vocal resonances, that is, the first three formant frequencies of the input speech. The input speech signals are directed through a set of contiguous band-pass filters, each with an associated rectifier and smoothing network (as in the previous patent application) to provide a short-time amplitude spectrum of the speech. But whereas the system of the previous patent application employs sampling procedure for examining this spectrum to determine and indicate the frequencies of the first three maxima, the system herein disclosed employs a different examination procedure. The examination procedure of the present invention, which may be defined as spectrum-segmentation, or maximum amplitude selection, takes advantage of the fact that the first three speech formants occupy frequency ranges which, on the average, do not overlap to any great extent.

Utilizing this characteristic of the speech formants to

segregate themselves into reasonably well-defined frequency ranges, the present invention provides a method of speech spectrum examination which comprises, as its first step, the process of grouping the channels of the speech analyzing filter set into three distinct channel families of which the first channel family (0–800 c.p.s.) coincides with a frequency range (hereinafter referred to as range F_1) embracing the first formant of the principal English speech vowels; the second channel family (800–2250 c.p.s.) coincides with a frequency range (F_2) embracing the second formant of these vowels; and the third channel family (2250–3600 c.p.s.) coincides with a frequency range (F_3) embracing the third formant of these vowels. The indicated values of F_1, F_2 and F_3 are for adult male speech.

As its second and third steps, the speech-examination method of the present invention comprises the process of selecting, from each of the three channel families, the individual channel thereof which reflects maximum signal content, and repeating this maximum amplitude-selecting step at a repetition rate of, say, 60 or more times per second. The fourth step is to store or register such maximum amplitude selections at the instant they occur.

The invention also embraces the particular apparatus and circuitry herein disclosed for practicing the speech spectrum-segmentation method of examination above outlined.

Other characteristics of the invention will appear upon reference to the following description of the invention as illustrated in the accompanying drawings wherein:

Fig. 1 is a block diagram of a system employing the speech spectrum segmentation concept underlying the invention;

Fig. 2 is a circuit diagram embodying the normalizing and maximum-selector components of the system of Fig. 1;

Fig. 3 is a circuit diagram embodying the clamper components of the system of Fig. 1;

Fig. 4 is a circuit diagram of a conventional type of amplifier assembly suitable for the circuits of Figs. 2 and 3;

Fig. 5 is a chart showing the average frequencies in c.p.s. and relative intensities in db of the first three formants of the ten indicated English vowels as uttered by male speakers; and

Fig. 6 is a series of wave diagrams illustrating how the system operates.

The arrangement of major components is illustrated in Fig. 1. The task of the system is to accept continuous speech at its input and to yield three output voltages, $F_1(t)$, $F_2(t)$, and $F_3(t)$, whose magnitudes as functions of time, represent the frequencies of the first three major vocal resonances (formants). Continuous speech signals are fed directly (or, if desired through a vowel extracting apparatus such as is depicted in "Fig. 2" of my co-pending application above-identified) into the analyzing filter set indicated at **10** in Fig. 1. This filter set **10** may be composed of 36 contiguous band-pass filters, each with an associated amplifier, rectifier and smoothing network, as illustrated in "Fig. 5" of my co-pending application. The output voltages (negative D.C.) of the individual filter units of the filter set are directed to filter output terminals **11***a*, **11***b*, or **11***c* (Fig. 1), as the case may be, depending upon whether each individually filter signal falls in the F_1, F_2, or F_3 frequency range, corresponding to the first, second, and third vowel formant groups. From terminals **11***a*, **11***b*, and **11***c* the output voltages (or, optionally, the second differenced outputs) are directed into amplitude normalizing networks **12***a*, **12***b*, and **12***c*, respectively.

As shown in Fig. 2, amplitude normalizing network

12*a*—and it will be understood that networks 12*b* and 12*c* are duplicates thereof—includes a series of resistance loads 22 (*a* to *n*), 23 (*a* to *n*), and 24 (*a* to *n*) equal in number to the number (*n*) of input lines connecting terminal group 11*a* to network 12*a*, said resistors functioning (under the control of adjustable resistor 27, and with the assistance of amplifier A_1) to subtract the mean value of all applied voltages from the value of each individual input voltage, and to present to the respective output terminals 30 (*a* to *n*) voltage values equal to one-half the magnitudes of the differences obtained in the several subtraction operations. For example, if e_k is the voltage input to the normalizing circuit from the *k*th filter channel of a channel family have N individual channels, then the normalized *k*th channel voltage is:

$$e'_k = \tfrac{1}{2}\left[e_k - \frac{1}{N}\sum_{n=1}^{N} e_n \right]$$

The normalized set of voltages presented to output terminals 30 advance to amplifiers A_2—A_{n+1} (Fig. 2) where the voltages undergo a gain of the order of 10 to 15, and also undergo polarity inversion. The boosted and inverted voltages are then applied to the appropriate grids of maximum-voltage selecting thyratrons 40*a* to 40*n*, whose plate circuits include a common resistor 42, and are connected by a common conductor 46 to the twin anodes of pulser 47, the latter having its cathodes grounded at 48, and its grids triggered by current from the secondary of a transformer 49 whose primary circuit receives energy from A.C. source 50. With pulsing current being applied to the plate circuit at 60 c.p.s. (assuming transformer 49 to be receiving 60-cycle input from source 50)pulser triodes 47 and pulser output circuit 46 will enable one or another of the thyratrons to fire every 1/60th second, with the activated thyratron being the particular one receiving the maximum positive grid voltage on any given cycle. Each time any thyratron is thus activated it operates to preclude the firing of any other. (If no positive voltages are being delivered to the thyratron grids, at the moment of enabling in the manner just described, then of course there will be no firing of any of the thyratrons.) The cathode outputs of the thyratrons are weighted by means of potentiometers 43*a* to 43*n*, and are summed to provide a single output, by way of resistive summation network 44*a* to 44*n*. The selector output at 17*a*, therefore, is a string of weighted rectangular pulses whose heights correspond to the number (or the frequency) of the channels successively selected as those of maximum output; it being understood that the individual potentiometers 43*a* to 43*n* are set so that they have voltage relationships, each to the others, corresponding to the frequency relationships existing among the individual channels whose frequencies are being monitored by said potentiometers.

The operation of the maximum amplitude selector of Fig. 2 is illustrated in Fig. 6 at (*a*), (*b*), and (*c*). Fig. 6(*a*) represents the output voltages of three arbitrarily chosen filter channels during seven successive selecting time intervals. In the first time interval, no output has appeared. In the second interval, a filter output has appeared and channel No. 1 has the maximum value. In the following intervals, the maximum moves successively from channel one to two; from two to three; from three back to two; and from two back to one. Fig. 6(*b*) shows the normalized values of these channel voltages during the same succession of selecting intervals. Fig. 6(*c*) assumes that the maximum selector is selecting from these three channels and shows its output as a function of time for the same succession of selecting intervals.

The unweighted cathode voltages of the maximum amplitude selector are summed to provide a "cathode sum voltage output" at point 16*a* (Figs. 2 and 3) for triggering the left-hand half 67 of a twin triode (Fig. 3) constituting part of the clamper circuitry. The clamper circuitry also includes a one-shot multivibrator (triodes 69 and 70, Fig. 3) and two amplifier assemblies 52 and 56, each having constituent parts like those of Fig. 4. The control grid of the first stage of amplifier assembly 52 receives the voltage output of the thyratron selection network by way of maximum amplitude selector output terminal 17*a* and conductor 51. The purpose of the clamper circuit is to provide a "staircase" smoothing of the rectangular pulses coming from the selector. To accomplish this, a gating pulse is generated at the proper phase of the enabling-disabling period of the maximum selector by the one-shot multivibrator 69—70, and this pulse is applied by way of conductor 65 to the grid of gating tube 55. Gating pulse 55 "reads" the height of each pulse from the maximum selector, "stores" this value for a brief holding interval, and then delivers it to terminals 18*a* by way of amplifier stage 56 and lead 59. Thus the output at terminals 18*a* represents (in "staircase" fashion) the heights of the successive output pulses from the selector 13A (Fig. 2). This height-reading operation is ilustrated in Fig. 6(*d*). The voltage read by the gating tube is stored and held in the clamper circuit until the next sampling occurs. The clamper output (that is, the "staircase" smoothing of the output pulses from the selector) is shown in Fig. 6(*e*). This output can be smoothed further by a passive low-pass network.

Triggered and synchronous sampling: Two methods for generating the height-sampling gate have been provided. They are termed triggered and synchronous sampling. During triggered sampling (switch 64 in the full-line position, Fig. 3), the height-reading gate is generated only if the thyratron selector is making a selection. The trigger is derived from a summation, without weighting, of the thyratron cathode voltages, and the gate "reads" each time any thyratron fires. During synchronous sampling (switch 64 in the dash-line position, Fig. 3), the height-reading gate is generated in synchronism with the enabling plate voltage of the thyratron set. It reads, therefore, regardless of whether or not a thyratron is firing. If a thyratron is not firing, the gate, of course, reads the value zero, and this appears at the clamper output 18*a*. The dotted portions of the curves in Figs. 6(*d*) and 6(*e*) indicate the result obtained for synchronous sampling.

The method of sampling determines the manner in which the clamper output voltage is extrapolated. With triggered sampling, the clamper holds the last value of voltage read when the thyratrons were selecting and firing. It loses this value relatively slowly, returning to zero or to a neutral voltage with a time constant of approximately a quarter-second. With synchronous sampling, the output voltage goes to zero in the enabling interval immediately following the last selection of the thyratrons. Therefore, if one wishes to extract formant signals which are extrapolated smoothly across consonant and silent intervals, the triggered sampling yields the best results.

The amplifier assemblies indicated in circular form at 52 and 56 in Fig. 3 may be of conventional design as, for example, the design shown in Fig. 4, which is that of a conventional plug-in type of amplifier assembly. The numerals in the small circles spaced about large circles 52 and 56 (Fig. 3) indicate connection of these circuit points to the points similarly designated by circled numerals in Fig. 4. The only difference in the external connections of units 52 and 56 is that unit 52 is connected to function as a polarity inverter (by way of phase inversion loop 53) whereas unit 56 is connected to function as a cathode follower with its output line 59 leading to terminal 18*a*, and with feed-back 57 to adjustable bias control network 58—60. Each unit (52 and 56) includes two twin triodes 71, 72 (Fig. 4), two gas diodes 73, 74 and resistance and capacitance parametric couples

of the values indicated adjacent each unit. Equivalent amplifier circuitry may, of course, be substituted.

Reverting to Fig. 3, resistor **62** is adjustable to set the proper phase relation between the voltages of transformer **61** and transformer **49** (Fig. 2), to assure that the sampling action will be truly synchronous, as described.

What is claimed is:

1. An automatic, electronic speech analyzing apparatus for extracting from input speech narrow bandwidth electrical signals whose varying amplitudes indicate the content of said input speech, said apparatus including means for converting the speech input into electrical signal groups whose frequency ranges differ, each from the others, in accordance with the speech spectral differences distinguishing major speech formants, one from the other, means for selecting from each of the subbands within each formant group the particular signal having maximum amplitude, said signal selecting means including a plurality of voltage-responsive ionization discharge devices, each adjusted to respond to the same voltage input level, and means responsive to the firing of any one of said devices to preclude the firing of any of the associated devices, during any given operating cycle, and means for simultaneously registering the selected signals.

2. Electronic speech analyzing apparatus as defined in claim 1 wherein said maximum signal selecting means includes electronic gating means controlling delivery of the output of said ionization discharge devices to said registering means.

References Cited in the file of this patent

UNITED STATES PATENTS

2,243,527	Dudley	May 27, 1941
2,458,227	Vermeulen	Jan. 4, 1949
2,635,146	Steinberg	Apr. 14, 1953