XIV. SPEECH ANALYSIS*

Prof. M. Halle
G. W. Hughes
A. R. Adolph

A. SPEECH PITCH-INFORMATION EXTRACTOR

Previous pitch-extraction devices have had three major limitations: restricted frequency range, restricted dynamic-input amplitude range, and inability to function with a nonideal input waveform.

The restricted frequency range is manifest in a device's inability to function with inputs from both male and female voices without manual variation of circuit parameters. In the lowpass filter method this entails shifting the cutoff frequency; in the envelope detection method it involves changing the integrating time constant.

Too narrow a dynamic-input amplitude range results in erroneous pitch indications either in the transient portion of the vowel sound or in the steady-state portion. These errors depend upon whether the range of the device is centered about the required lower level for operating on transients or the average steady-state level.

When a vowel with a relatively undamped first formant and, consequently, a poorly defined periodic envelope structure, is applied to an envelope detector or autocorrelator, an error occurs. Correspondingly, if harmonics of the pitch frequency fall within the filter passband of a lowpass filter pitch extractor, the pitch indication is erroneous.

An improved device should operate without manual adjustment over a pitch-frequency range that includes both male and female voices, should have a dynamic-input range at least as great as the input speech (60-70 db), and should be indifferent to the input waveform, provided that it is periodic.

The problem of frequency range will concern us first. In an earlier Quarterly Progress Report (1) we saw that full-wave rectification followed by lowpass filtering enabled us to utilize a filter cutoff frequency that is high enough (i.e., 500 cps) to be effective for both male and female voices. In practice, lowpass filters are not ideal and have an attenuation that increases at some finite rate above the cutoff frequency. This results in the appearance in the modified waveform of some first-formant energy of reduced amplitude. One possible way of dealing with this is to use a lowpass filter with a higher rate of attenuation above the cutoff frequency. However, as we do this, the phase shift in the neighborhood of the cutoff frequency becomes very great, adversely affecting input signals whose pitch frequency is high.

This problem can be solved by using a time-domain filter with a response that increases exponentially as a function of time, as illustrated in Fig. XIV-1. The effect of using such a filter is to reduce the secondary peaks exponentially. The peaks that

*This work was supported in part by the National Science Foundation.

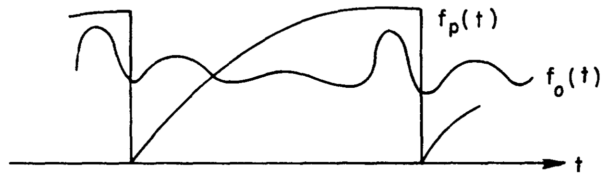


Fig. XIV-1. Exponential time-domain filter response superimposed on speech waveform.

occur immediately after the major peak are more attenuated than those that occur near the end of the pitch period. Ideally, the exponent of the filter waveform should be slightly smaller than the rate of envelope decay. It might appear that introducing time-domain filtering would defeat the objective of attaining a wide pitch-frequency range, since the time-domain filter will be effective at only one frequency, i.e., the inverse of the filter period. This would be true if a fixed filter period were used, but by making the length of the filter period variable and equal to the pitch period, this limitation can be overcome.

By using time-domain filtering, as outlined above, we have also met the requirement of indifference to the input waveform. A waveform with secondary peaks, even after the initial rectification-filtering operation, will be acceptable as an input because of the reduction of these peaks in the time-domain filtering process.

The problem of a wide dynamic range can be viewed in two ways. The part of the device which has to decide when to indicate a pitch period (the "counter") can be built with the required dynamic range; or the signal amplitude at which the counter operates can be modified periodically. The usual monostable multivibrator counters have a dynamic range that is well below the required 60-70 db range. Therefore, the second method – sampling the amplitude of the counter input periodically and maintaining a constant level that falls within the counter's requirements – would appear to be the more fruitful approach. This can be done by varying the amplitude of the time-filter window function proportionally with the amplitude of the input waveform as sampled at the pitch period, as will be described in the following section. Theoretically, this should result in an extremely large over-all dynamic range for the device.

1. Principles of Operation

Figure XIV-2 is a block diagram of the complete pitch extractor and the speech pitch-information extractor (SPIE). The sequence of events leading up to the SPIE input is: (a) The speech signal is amplified from its source level (-55 dbm for microphone, 0 dbm for tape) to a peak-to-peak level of approximately 10 volts and is lowpass-filtered to remove frequencies higher than 1500 cps. (b) At this level the signal is rectified and lowpass-filtered (500 cps). (c) The rectified and filtered signal, which is at a 5-volt peak-to-peak level, is amplified to a 150-volt peak-to-peak level and is applied to the SPIE input.

The counter in the SPIE is a Schmitt trigger circuit. This particular monostable configuration has the property of an adjustable on- and off-triggering level, the difference between the two levels being variable and dependent upon the circuit parameters. Thus an incoming signal to the Schmitt circuit will produce a pulse the width of which is a function of the difference between the triggering levels. The lower limit on the input-signal amplitude is dependent on this difference. It must be at least as great in order to produce a pulse.

In the preceding section we stated that in order to maintain a constant peak amplitude at the counter input it is necessary to sample the input-signal amplitude at the time the pitch pulse occurs. The sampling interval should be as narrow as possible in order to represent truly the signal amplitude at the sampling instant; therefore the differences between the on- and off-levels of the Schmitt trigger must be made as small as is consistent with sharp on-off transitions.

The sample value is applied to an RC network that generates an exponential bias function whose initial step is equal to the amplitude of the input signal at the time the pitch pulse occurs. Figure XIV-3(1a) shows a damped sine-wave input signal. The exponential function derived from this signal appears in Fig. XIV-3(1b). It has been inverted before being added to the input signal, which is the next step and also completes the feedback loop.

The output of a time-domain filter with a quasi-periodic speech-sound input is the

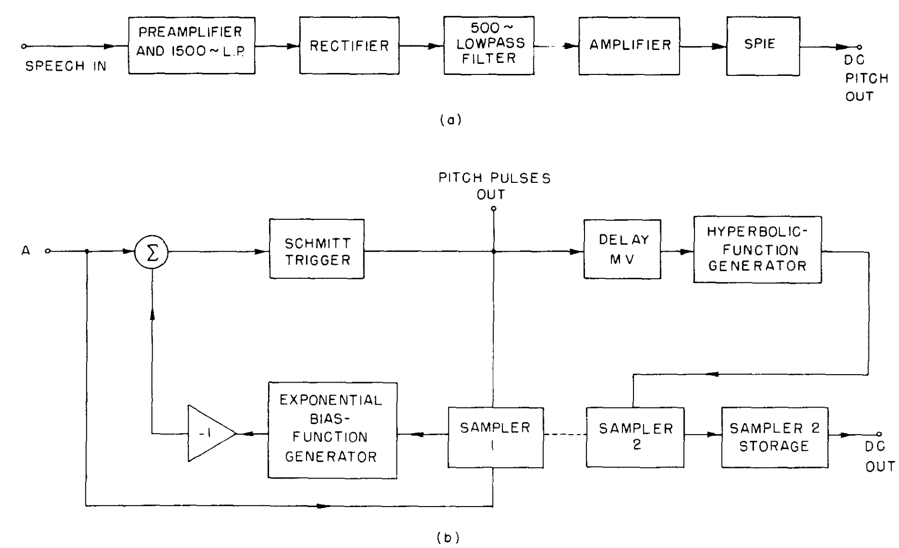
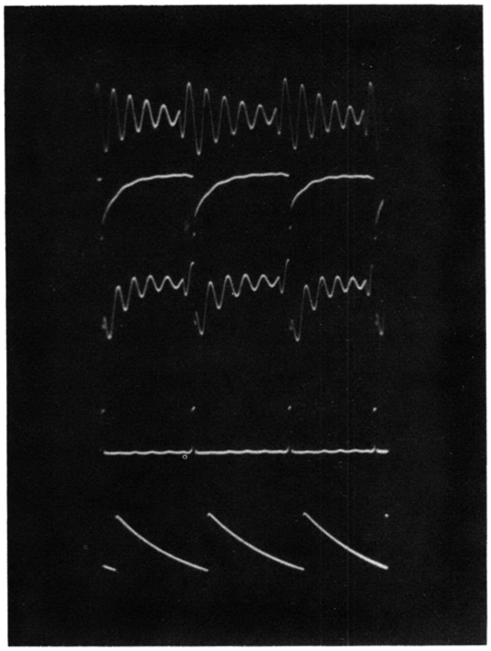
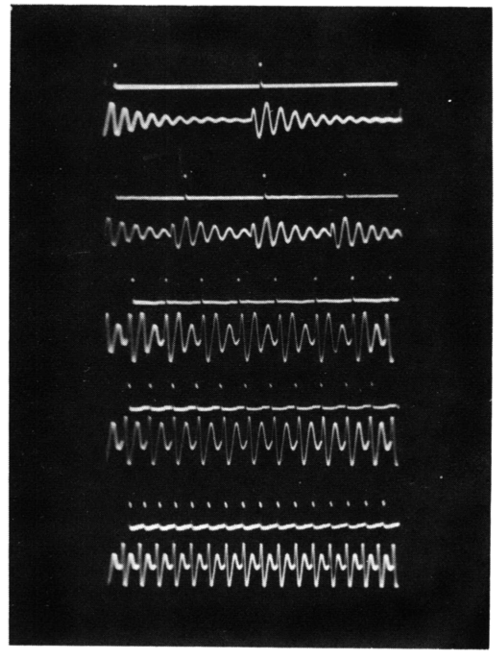


Fig. XIV-2. Block diagrams: (a) pitch-extractor system; (b) SPIE.

a
b
c
d
e



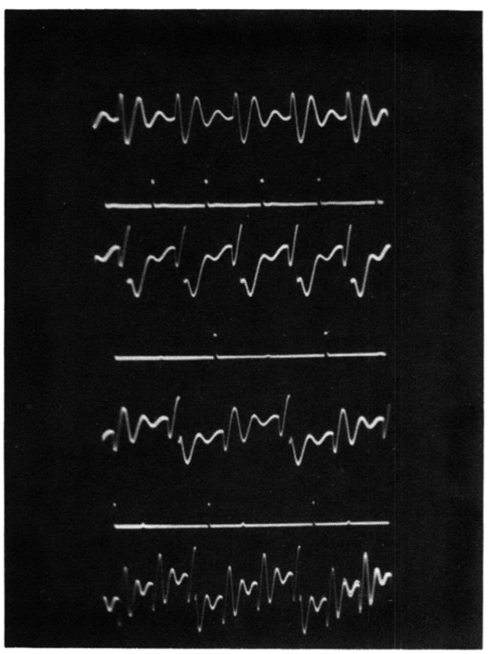
1



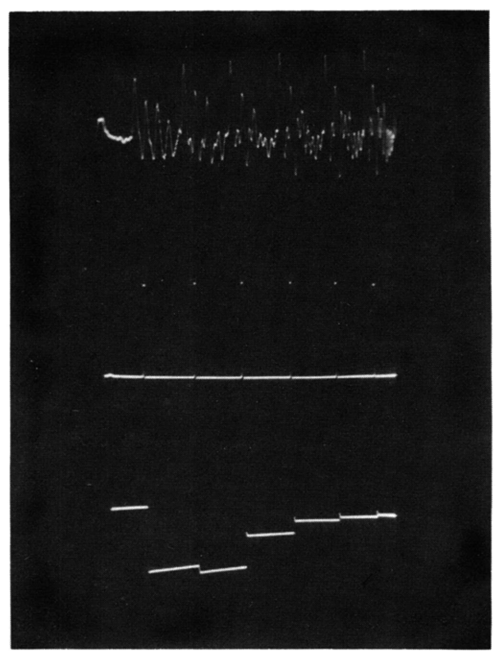
a
b
c
d
e
f
g
h
i
j

2

a
b
c
d
e
f
g



3



a
b
c

4

Fig. XIV-3.

1. a. Damped sine-wave pitch-extractor input.
   b. Exponential bias function.
   c. Sum of (a) and (b).
   d. Pitch pulses.
   e. Converter hyperbolic function.

2. a. Pitch pulses derived from (b).
   b. Damped sine wave with 50-cps pitch frequency.
   c. Pitch pulses from (d).
   d. 100-cps damped sine wave.
   e. Pitch pulses from (f).
   f. 200-cps damped sine wave.
   g. Pitch pulses from (h).
   h. 300-cps damped sine wave.
   i. Pitch pulses from (j).
   j. 400-cps damped sine wave.

3. a. Damped sine wave.
   b. Pitch pulses.
   c. Sum of (a) and derived exponential bias function.
   d. Pulse output for (e).
   e. Exponential bias-function period; twice the input period.
   f. Pulse output for (g).
   g. Exponential bias-function period; three times the input period.

4. a. Speech input: vowel [/a/] as in <u>bar</u>.
   b. Pitch pulses.
   c. Converter output.

convolution integral of the two functions. The output waveform is a periodic function each period of which resembles the product of the time-domain filter window function and a period of the speech wave. If the window function is an exponential that increases from zero to some finite value in a time slightly less than one pitch period, and has a time constant equal to the slope of the speech-wave envelope, the result of the convolution will be a waveform with a single peak in each pitch period.

From the point of view of the counter there is no difference between depressing the intermaxima variations on a negative, exponentially increasing bias, and multiplying them by the window function mentioned in the preceding paragraph. The maximum amplitude peak is always at the same level (the Schmitt triggering level) with the intermaxima variations "nowhere in sight." As far as the counter is concerned, it is of no consequence that the intermaxima variations have been depressed on a negative exponential bias rather than reduced to zero and built up to normal size exponentially. Thus, for all practical cases, the input-signal peak will always occur at the triggering level of the Schmitt circuit, resulting in a narrow output pulse. Figure XIV-3(1c) illustrates the resulting signal when the exponential bias function is subtracted from a damped sine-wave input signal. This is precisely what appears at the counter input. The triggering level is at the point on the waveform at which discontinuity occurs. The resulting pitch pulses are shown in Fig. XIV-3(1d). The pulsewidth is of the order of 1 msec.

Referring to Fig. XIV-3(2) again, we see that after the pitch pulse has been derived, the next step is to convert the digital pitch information to analog form. What is desired is a direct voltage proportional to the pitch period. Actually, this cannot be a continuous variable. The periodic nature of the pitch determination gives a step-like output from the pulse-to-analog converter. In Figs. XIV-3(3) and XIV-3(4) an actual vowel input to the device is shown with corresponding pitch pulse and analog output.

It is desirable that the analog output be proportional to pitch frequency rather than to period. For this reason, a hyperbolic function that represents the relationship between pitch period and frequency is required in the converter. In the pulse-to-analog converter the trailing edge of the pitch pulse triggers a delay multivibrator. The delay of the multivibrator is adjusted to compensate for the deviation from a hyperbolic function exhibited by an exponential near the origin. The value of the hyperbolic function initiated by the preceding pulse is sampled at the next pulse and the sample value is stored until a new pitch pulse occurs. A long interval between pitch pulses (long period) allows the hyperbolic function to drop to a low value, and thus indicates a low pitch frequency; a short interval results in a high sample value corresponding to a high pitch frequency. Figure XIV-3(1d) illustrates the hyperbolic-function approximation generated by a single RC network in the pulse-to-analog converter.

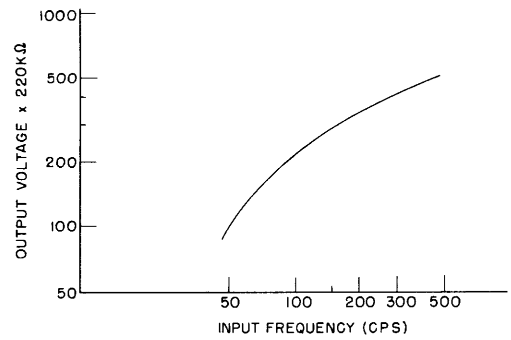


Fig. XIV-4. Converter transfer characteristic.

2. Test Results

Both static and dynamic tests were performed. The input waveforms used for the static tests were undamped sine waves, damped sine waves, and actual vowel sounds segmented from continuous speech and recorded on tape loops. The dynamic tests used direct-speech inputs.

Figure XIV-4 shows a plot of the converter-output voltage as a function of the input frequency. It is evident that a single RC network does not approximate the desired hyperbolic function too well, even with the initial time delay. There is an uneven distribution of frequencies that spreads the low-frequency end of the scale and compresses the high-frequency end. The accuracy of the converter calibration was ensured by correlating the sine-wave input frequency with the pitch period measured on an oscilloscope.

Dynamic-input amplitude-range measurements were made with both damped and undamped sine waves. The minimum input level is defined as the level at which a pitch pulse first appears as the input level is raised from zero. The maximum input level is defined as the level at which an erroneous pitch-pulse output (i.e., more than one pulse in a pitch period) occurs as the input level is raised from zero. The dynamic-input amplitude range is the range of input levels between the minimum and maximum input levels. The dynamic range was measured as a function of frequency and of the bias-function time constant. The results of these measurements for bias-function time constants of 20 msec, 4 msec, and 2 msec are plotted in Fig. XIV-5. The frequencies at which the 20-msec time-domain-filter time constant (switch position 1) gives rise to a dynamic-input range of greater than 50 db are 90 cps (upper) and 75 cps (lower). For a 4-msec time constant (switch position 9) the upper and lower frequency limits corresponding to a 50-db minimum dynamic range are 400 cps and 120 cps. A 2-msec time constant (switch position 10) results in a 50-db minimum dynamic range between frequencies of 1000 cps and 200 cps. The seven other bias-function time constants have resultant dynamic-range curves in the region between the 20-msec and the 4-msec curves.

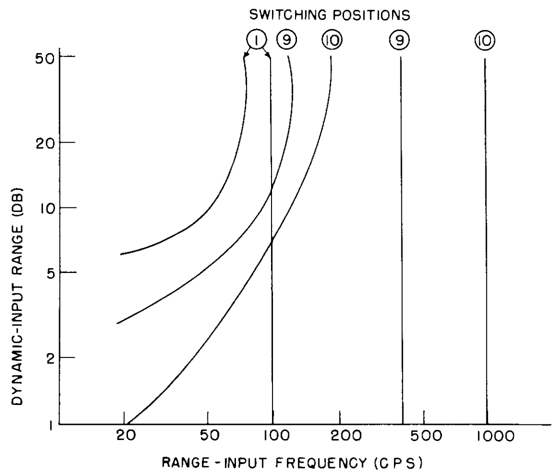


Fig. XIV-5. Dynamic-input, range-input frequency characteristic.

The bandpass nature of the dynamic-range curves seems to be anomalous with respect to the results predicted for an exponentially topped time-domain filter function. On the basis of the theoretical treatment, a gaussian lowpass frequency-domain behavior that would result in a similarly shaped dynamic-range curve was expected. The frequency response on a linear-frequency scale would be well down at a frequency twice the inverse of the filter time constant. This response characteristic corresponds to the part of the dynamic-range curve that is most influenced by the minimum input level, i. e., the upper-limit frequencies mentioned earlier. The gradual nature of the response is not too apparent because of the logarithmic scale. The point at which the dynamic range is down to 1 db is approximately twice the inverse of the time-filter time constant. The fall-off of the dynamic range at low frequencies is most influenced by the maximum input level, which, in turn, is highly dependent upon the segment of the time-filter wave form just before the beginning of a new period. This occurs after an interval of three or four time constants. Re-examining Fig. XIV-5, we see that for the 2-msec time constant, the upper limit occurs at twice the inverse of the time constant, i. e., at 1000 cps, and the lower limit is approximately the inverse of three times the time constant, or at 166 cps.

The dynamic speech-input test was made by means of a continuous motion-picture recording of the two traces of a dual-beam oscilloscope that had the speech wave on one beam and the extracted pitch pulses on the other. Two runs were made with the test sentence "Joe took father's shoe bench out." On the first run the speech was full-wave rectified, 500-cps lowpass-filtered before being applied to the SPIE input. On the second run the speech was half-wave rectified, lowpass-filtered, and fed to the SPIE. Errors were sought only during the voiced portions of the sentence. Table XIV-1 presents the results of the analysis.

Table XIV-1. Error Analysis.

Word	Full-Wave		Half-Wave	
	Errors	in periods	Errors	in periods
Joe	0	36	not filmed	
took	3	11	1	11
father's	3	52	3	73
show	3	22	1	22
bench	3	24	0	30
out	3	31	0	34
Total	15	176	5	170
Errors (per cent)	8.5		3	

The errors made by the full-wave system, and not by the half-wave system, were on the /u/ and /U/ sounds in took, out, shoe, and the /n/ sound in bench. These were errors of omission and were caused by the low fundamental-frequency output (below minimum level) from the full-wave rectifier-filter combination for these sounds. The half-wave system did not err because of inefficiency in removing first-formant energy. Enough of the first-formant energy was present to give fundamental-like periodic waves after lowpass filtering. Both systems erred on the voiced fricative /th/ in father's. This error was also caused by insufficient fundamental energy for operating the SPIE, but in this case there was no first-formant energy to compensate for this lack in the half-wave system.

The larger percentage of errors made by the full-wave system appears to contradict the preliminary conclusion (1). This conclusion was based on a comparison of the efficiency of the two methods of rectification and filtering in separating the fundamental-frequency component from the formants, and not on the effect of variations in the energy content of the fundamental-frequency component. The superior performance of full-wave rectification over half-wave rectification can be realized if a dynamic-input amplitude range is available that will enable the SPIE to respond to all of the possible values of fundamental-frequency energy encountered in speech sounds.

3. Conclusion

The present device does not unqualifiedly meet the criteria stated at the beginning of this discussion; but if we consider the dynamic-input-range and frequency-range requirements together, it almost makes the grade. It will operate without manual adjustment over a frequency range of approximately 100 cps to approximately 350 cps and will meet the dynamic-range requirement of 60 db. Nor is it indifferent to the input waveform without increasing the signal amplitude for certain vowels.

The first-mentioned limitation may be eliminated by using a different time-filter waveform. The second limitation can perhaps be overcome by using an automatic vowel recognizer to make an optimal adjustment of the internal operating conditions of the pitch extractor for particular vowel sounds.

If the performance of previous devices is used as a standard for comparison, the present device represents noticeable improvement on the basis of its dynamic range over any given frequency range and of the percentage of errors made in a continuous speech sample.

A. R. Adolph

References

1. Quarterly Progress Report, Research Laboratory of Electronics, M. I. T., April 15, 1957, p. 127.