New Methods of Pitch Extraction

MAN MOHAN SONDHI

Abstract—Three new methods will be described for the extraction of the fundamental pitch from a speech signal. These are: 1) spectrum flattening followed by a minimum phase correction to synchronize harmonics, 2) spectrum flattening followed by autocorrelation, and 3) nonlinear distortion followed by autocorrelation.

The last two methods will be shown to be exceptionally rugged, in that they can tolerate a considerable amount of high-pass filtering and additive noise with little degradation in performance.

Manuscript received September 28, 1967. This paper was presented at the 1967 Conference on Speech Communication and Processing, Cambridge, Mass.

The author is with Bell Telephone Laboratories Inc., Murray Hill, N. J.

I. Introduction

A LARGE proportion of present-day vocoders are based upon the analysis of a speech signal into an excitation signal and a vocal tract transfer function. Both the excitation and the transfer function are then described in terms of a small number of slowly varying parameters, from which an estimate of the original speech wave is synthesized. There is need for improvement in our descriptions of both the transfer function as well as the excitation. However, the remarkably small degradation of speech quality in a voice-excited vocoder[1] indicates that the greater need is for an improved parametric representation of the excitation.

Traditionally, the excitation is regarded as consisting of intervals that are either voiced (v) or unvoiced (uv). Such a v/uv dichotomy is clearly an oversimplification, as indicated, for instance, by the existence of voiced fricatives. However, it is generally accepted that many improvements in our methods of deriving the excitation signal are possible, even without the embellishment of partial voicing.

We will use the term pitch extractor for a device or algorithm which makes a v/uv decision, and, during periods of voicing, gives an indication of the pitch period (or the fundamental frequency). In this paper, we describe three out of a number of pitch extractors considered by the author during the past few years. Our only apology for adding to the already large list of methods for pitch extraction described in the literature[2] is that the methods described here are based upon promising new ideas not previously reported.

The basic notion common to all three pitch extractors described here is the following. If the harmonics of the fundamental frequency could be made equal in amplitude and put into phase synchronism with each other, the resulting waveform would be a train of highly peaked pulses, and the interval between these pulses would correspond to the current pitch period. During unvoiced intervals, no such pulse train would obtain and, thus, a v/uv decision could be based upon the presence or absence of the pulse train.

We turn now to a method for equalizing amplitudes of the harmonics. This will form the basic element for the first two of our pitch extractors.

II. The Spectrum Flattener

Due to the variability of the fundamental frequency and of the formants, no fixed equalizer can adequately equalize the amplitudes of the harmonics. What is needed is a system that dynamically adapts to the varying spectrum. A block diagram of one such scheme is shown in Fig. 1. The speech signal is filtered through a bank of n band-pass filters (approximately 100-Hz wide) $F_1, \cdots, F_n$, which span the bandwidth of the

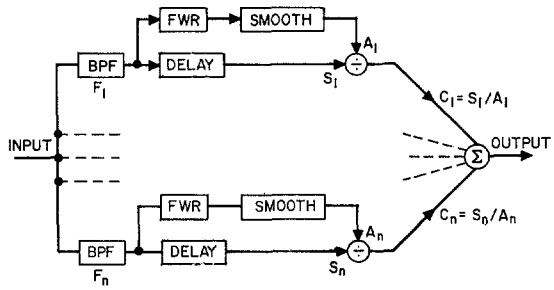

Fig. 1. Schematic of spectrum flattener.

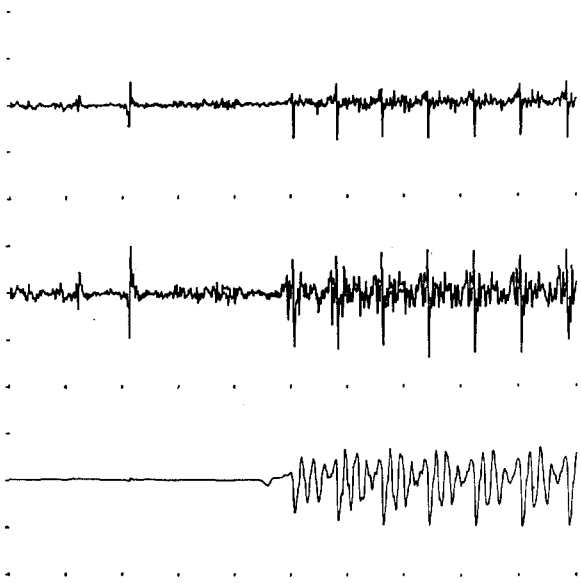

Fig. 2. Example of spectrum flattening (middle trace), spectrum flattening and minimum phase compensation (upper trace). The lowest trace is the original speech. The total duration is 100 ms.

signal. The output of the filter F_i is fed to a full-wave rectifier and to a delay line. The full-wave rectifier output is smoothed to give a short-time estimate A_i of the amplitude of the output of F_i. The signal S_i is the output of F_i delayed by an amount D to compensate for the delay of the smoothing filter. The signal $C_i = S_i/A_i$ is, thus, the output of F_i normalized to unit amplitude. The processing of the channels 1 through n is identical. The signals C_i are summed to give the desired flat spectrum signal. An example of such a spectrum flattening is shown in Fig. 2, where the lowermost trace is the original speech input, and the middle trace is the spectrum flattened version. The voiced intervals, as expected, show up as trains of rather peaked pulses. No such peaking occurs in unvoiced intervals.

The basic objective of spectrum flattening can be achieved in ways other than the one shown in Fig. 1. Thus, for example, the Hilbert envelopes of the signals S_i can be used as estimates of amplitude A_i; or the signals S_i can be infinitely clipped and refiltered through filters identical to the F_i.

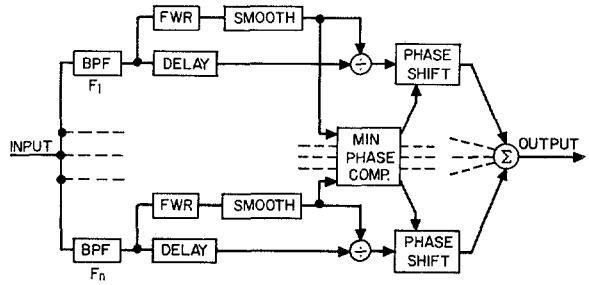

Fig. 3. Schematic of spectrum flattener with minimum phase compensation.

III. Pitch Extraction by Minimum Phase Compensation of Spectrum Flattened Speech

As remarked in Section I, the fundamental periodicity would be rendered much more detectable if the equal-amplitude harmonics could be phase synchronized. The following method suggests itself. The amplitude estimates A_i $(i = 1, \cdots, n)$ provide n equally spaced samples of the short-time spectrum of the speech signal. There exists a unique minimum phase network which has this amplitude response, and whose phase response can be readily computed.[3] Approximate synchronization can be achieved by phase shifting each of the signals C_i by an amount equal and opposite to the minimum phase computed for the ith channel. This is illustrated in Fig. 3. The phase shifting may be implemented by adding the proper proportions of C_i and a signal in phase quadrature to C_i. (The signal $dC_i/d(w_{0i}t)$ is a good approximation to the quadrature signal, where w_{0i} is the center frequency of the ith filter.)

The uppermost trace in Fig. 2 is an example of speech processed this way. The pitch markers turned out to be much sharper than we had anticipated. We must hasten to add, however, that the method is not always as successful. It is clear that the minimum phase assumption is not always valid, e.g., for nasalized vowels and whenever the glottal pulse shapes are not well approximated by the impulse response of a minimum phase network. Therefore, this method does not appear to be well suited for use as a part of a vocoder system. Nevertheless, it is an interesting approach, and could be a useful laboratory tool for providing synchronous pitch markers.

IV. Pitch Extraction by Autocorrelation of Spectrum Flattened Speech

Another method that we have successfully used to synchronize the harmonics is autocorrelation of the output of a spectrum flattener. The idea of autocorrelation is not new, and has been used both for pitch extraction as well as various other applications where an indication of periodicity is desired. However, the preprocessing by the spectrum flattener makes this type of pitch ex-

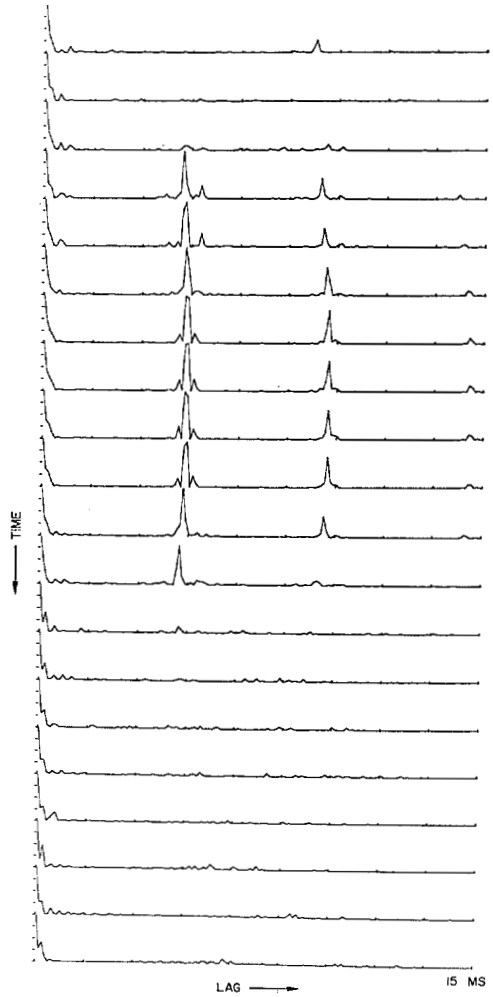

Fig. 4. Sample of correlation functions. Each trace is the correlation function of a 30-ms segment of spectrum flattened speech with a 15-ms overlap with the preceding and succeeding segments.

tractor highly reliable. To our knowledge, the only other type of preprocessing tried with autocorrelation pitch extractors is infinite peak-clipping.[4] This results in a binary signal which considerably simplifies the implementation of the correlator. However, as is well known, infinite peak clipping does not remove the formant structure of the speech, and the peaks in the correlation function due to formants are difficult to distinguish from those due to the fundamental periodicity. The spectrum flattened signal is, on the other hand, virtually devoid of formant structure, and its correlation function has no discernible peaks due to formants.

The correlation function is computed as follows. A 30-ms segment of the spectrally flattened speech signal is isolated and multiplied by a Hamming window.[5] The autocorrelation of this truncated segment is computed for lags up to 15 ms and normalized to unity for zero lag. Another 30-ms segment of the signal is selected with a 15-ms overlap with the previous segment, and the process continued. A sample of successive traces obtained in this manner is shown in Fig. 4. (To make a better visual display, the correlation functions were half-wave rectified and squared before plotting.[1])

The peaks in the correlation function corresponding to the pitch period are readily picked up. However, to ensure that peaks are not missed during rapid vocal tract transitions and during the trailing portions of voiced intervals, a decision algorithm is used. This algorithm and the method used for generating slowly varying parameters to describe the v/uv decision and the pitch will be discussed after a description of the third pitch extractor.

V. Pitch Extraction by Center Clipping and Autocorrelation

As shown in Section IV, autocorrelation of a speech signal, after removal of the formant structure, is a powerful method of pitch extraction. We have successfully tried another way of removing formant structure, namely, center clipping. The center clipped speech signal is obtained in the manner illustrated in Fig. 5. In every 5-ms interval, the maximum absolute value a_0 of the signal is found, and all portions of the signal between $\pm ka_0$ are removed. Typically, k is chosen to be about 0.3. Autocorrelation functions for this center clipped signal are then computed as in Section IV. A sample of traces of correlation functions obtained from center-clipped speech is shown in Fig. 6.

Center clipping of speech was first used by Licklider and Pollack[6] in an experiment in which they showed that, whereas speech that has been infinitely peak-clipped is highly intelligible, even a few percent of center clipping drastically reduces intelligibility. The explanation is not hard to find. Whereas infinite peak-clipping retains the formants of the speech signal (although it introduces a few secondary "formants"), center clipping destroys formant structure, while retaining the periodicity. It is the removal of formant structure that is so important for a good pitch extractor.

Center clipping, as seen from Fig. 6, is an efficient and exceedingly simple way of formant removal. In conjunction with an autocorrelator and the decision algorithm (to be described in Section VI), it provides a very good pitch extractor. In at least one type of situation, it works more reliably than spectrum flattening, or even the more elaborate cepstrum pitch extractor. This is the case when a voiced segment of speech becomes almost sinusoidal. (This occurs, for example, if the speech signal is the sound /i/ spoken by a female and high-pass filtered with a cutoff at about 200 or 300 Hz. This is not a very unusual situation if the speech has traveled over an ordinary telephone circuit.) Since the success of both the cepstrum and the spectrum

[1] The length of the analysis intervals is not critical. We have successfully tried 40-ms analysis intervals with a 30-ms overlap between consecutive intervals.

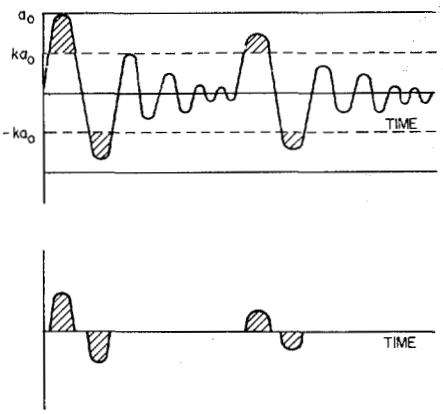

Fig. 5. Illustration of the process of center clipping.

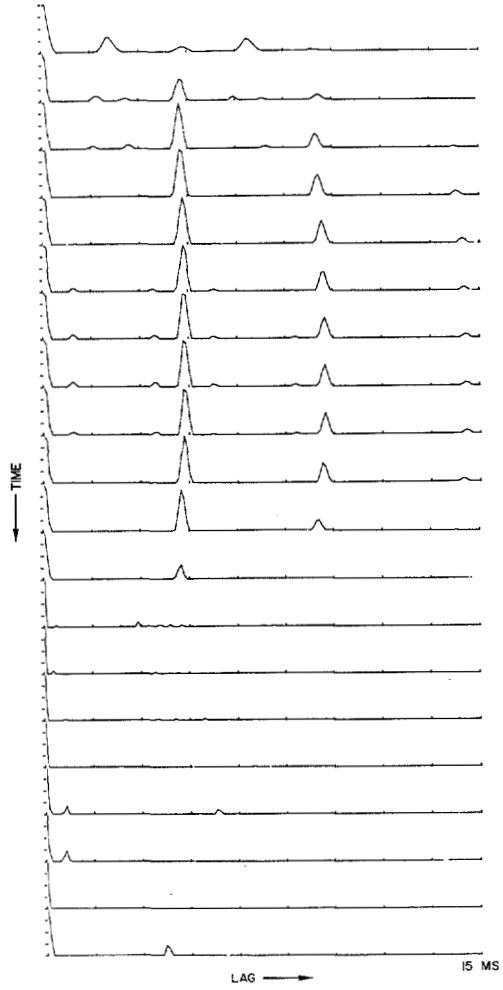

Fig. 6. Sample of correlation functions. Each trace is the correlation function of a 30-ms segment of center clipped speech with a 15-ms overlap with the preceding and succeeding segments.

flattener depends upon the presence of a large number of harmonics, these types of pitch extractors are prone to error in such cases. The absence of a large number of harmonics clearly is not a serious problem for the center-clipping method.

VI. The Decision Algorithm

A relatively simple algorithm suffices to pick the correlation peaks corresponding to the pitch period during voiced intervals, and to make the v/uv decision. The algorithm is adapted from the one used by Noll for his cepstrum pitch extractor.[7] Since it has been described in detail in his paper, only the main points are summarized here.

1) Each correlation trace of the type shown in Figs. 4 and 6 is the basis of a decision which applies to 15 ms of the speech.

2) The correlation functions are weighted to emphasize peaks at large lags (mainly to offset the effect of the Hamming window). The weighting function was determined by trial and error, and is essentially a linear weighting.

3) A threshold is preset, and the location of the first peak in the correlation function exceeding this threshold is accepted as the pitch for the corresponding 15-ms interval. (No acceptable peak indicates an unvoiced interval.)

4) The decision for the current interval (say B) is modified by the decisions for the preceding and succeeding intervals (say A and C, respectively). If A and C are voiced (unvoiced), then B is forced to be declared voiced (unvoiced). If A and C have approximately equal pitch, and the pitch for B differs by more than 60 percent, then the pitch for B is declared to be the average of A and C. If two successive intervals show a large departure from the pitch for preceding intervals, it is accepted as a genuine large change.

5) If a pitch peak is found for one 15-ms interval, the threshold for the succeeding intervals is lowered by a factor of 2 over a ± 1-ms region around the peak. The original threshold is restored if the pitch changes or the voicing interval ends.

These features eliminate the occasional errors of pitch doubling, spurious voicing, and spurious unvoicing that would otherwise arise.

VII. Tests and Conclusions

It is clear that the two preceding pitch extractors can tolerate considerable high-pass filtering of the input signal, as well as addition of broadband noise. In fact, the traces of Fig. 4 were obtained from speech filtered through a band-pass filter from 250 to 3250 Hz, and those of Fig. 6 were obtained from speech similarly filtered with additive white noise at a signal-to-noise ratio of about 18 dB.

As a test of the use of these pitch extractors in vocoder systems, the output of the decision algorithm was converted to two vocoder signals. The first signal had a value 0 during unvoiced intervals, and 1 during voiced intervals. The second signal had a constant value for each 15-ms interval, corresponding to the pitch for that interval. These two signals were low-pass filtered to

about 33 Hz. These low-pass signals (A and B, respectively) were then converted to an excitation signal as follows. Whenever A exceeded $\frac{1}{2}$ (indicating an unvoiced interval), the excitation consisted of white Gaussian noise. In the voiced intervals, pitch pulses were obtained by means of a ramp signal of constant slope. Whenever the ramp reached the level of signal B, a pitch pulse was generated, the ramp was reset to zero, and the sequence was repeated as often as necessary until the end of the voiced interval. Excitation signals generated in this manner were used with a 13-channel vocoder simulated by Golden.[8]

The resulting resynthesized speech was judged excellent by listeners in informal listening tests. None of the usual troubles of pitch doubling and loss of the trailing portions of voiced intervals was noticeable.

In conclusion, we might mention that we have tested the pitch extractors by computer simulation. We have tested them on a number of male and female voices, with speech signals that have been high-pass filtered and have additive white noise, and with signals that have been recorded over a telephone line. Computer simulation is not an entirely adequate test of these systems; however, the success so far indicates them to be worthy of extensive study on hardware models based upon the simulations.

References

[1] M. R. Schroeder, E. E. David, B. F. Logan, and A. J. Prestigiacomo, "Voice-excited vocoders for practical speech-bandwidth reduction," *Proc. Internat'l Symp. on Information Theory* (Belgium, August 1962).

[2] N. P. McKinney, "Laryngeal frequency analysis for linguistic research," Communication Sciences Lab., University of Michigan, Ann Arbor, Rept. 14, September 1965. Chapter 4 gives a comprehensive bibliography.

[3] H. W. Bode, *Network Analysis and Feedback Amplifier Design*. New York: Van Nostrand, 1947, ch. 14.

[4] J. S. Gill, "Automatic extraction of the excitation function of speech," *Proc. 3rd Internat'l Cong. on Acoustics*, vol. 1. Amsterdam: Elsevier, 1961, pp. 217–220.

[5] R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra*. New York: Dover, 1959.

[6] J. C. R. Licklider and I. Pollack, "Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech," *J. Acoust. Soc. Am.*, vol. 20, pp. 42–50, January 1948.

[7] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, vol. 41, pp. 293–309, February 1967.

[8] R. M. Golden, "Digital computer simulation of a sampled-data voice-excited vocoder," *J. Acoust. Soc. Am.*, vol. 35, pp. 1358–1366, September 1963.

Man Mohan Sondhi was born in Ferozepur, India, on December 18, 1933. He received the B.Sc. (Honors) degree from Delhi University, Delhi, India, in 1950, the Diploma of the Indian Institute of Science, Bangalore, India, in 1953, and the M.S. and Ph.D. degrees from the University of Wisconsin, Madison, in 1955 and 1957, respectively.

He has worked at the Bell Telephone Laboratories, Inc., Murray Hill, N. J., since 1962 on problems concerning the processing and transmission of speech signals. He is currently interested in similar problems, as well as in modeling the detection of auditory and visual signals by humans.