CHAPTER 3: ISSUES IN SPEECH FUNDAMENTAL FREQUENCY AND PERIOD ESTIMATION

3.1 INTRODUCTION

This chapter explores some of the issues and problems involved in the estimation of speech fundamental frequency. Firstly there is a discussion of what is meant by the terms fundamental frequency, fundamental period and pitch. Some aspects of human pitch perception and their relationships to the requirements of algorithms that estimate speech fundamental frequency are then discussed. Finally, there is a brief introduction to the basic approaches to speech fundamental frequency estimation by machine.

3.1.1 Fundamental frequency and pitch

Before entering into an in depth discussion of the problems involved in estimating speech fundamental frequency, it is necessary to precisely define the problem. It is also enlightening to investigate the relationship between the parameters fundamental frequency, fundamental period and pitch.

The automatic estimation of fundamental frequency of voiced speech excitation is often misleadingly referred to as pitch analysis. Pitch properly refers to a percept rather than a parameter of speech production (McKinney, 1965), although the term pitch is often used in current technical literature to express both fundamental frequency and fundamental period. Pitch is a subjective phenomenon whereas fundamental frequency is open to physical measurements. There is a relationship between pitch and frequency, but it is rather complex, although pitch is correlated with the physical feature of fundamental frequency. Thus, when one is considering speech at the acoustic level, it is preferable to use the concept of fundamental frequency. It is also useful to distinguish between fundamental period estimation, implying a period-by-period estimation process, and fundamental frequency estimation, which result from short-term analyses.

3.1.2 Approaches to speech analysis

Zwicker, Hess & Terhardt (1967) looked at the problem of speech analysis from three different points of view. All of these are related for stationary periodic signals, but not for real speech.

1] Speech can be considered from the production viewpoint, and it can be analysed using knowledge about the way in which it was generated. The parameters that are estimated using such an approach are related to the control parameters of the speech production process. At the lowest level of speech production, one can define the larynx fundamental periods as the time between successive vocal folds closures. Similarly the larynx fundamental frequency can be defined as rate of the vocal fold vibration.

2] From a perceptual viewpoint, speech may be analysed in a fashion that is similar to the processing that is believed to occur in the human auditory system. In the case of a human listener or by using a model of human pitch perception, the perceived pitch of a speech stimulus can be defined as the frequency of a pure tone that evokes the same perceived pitch.

3] From the signal processing viewpoint, speech analysis does not necessarily take account of speech production or speech perception, but seeks to describe the signal in some mathematically optimal way. If the speech production process is not taken into account, the fundamental frequency and the fundamental period of the speech can be defined in terms of the minimum repetitive period of the signal, or corresponding to the common sub-multiple of a set of harmonics. This task can be carried out using digital signal processing techniques, such as auto-correlation.

3.1.3 Simplified model of speech excitation

A simplified model of the excitation of voiced speech sounds was described by McKinney (McKinney, 1965). In this model, a volume velocity glottal excitation function ug(t) excites a passive linear system. This is illustrated in figure 3.1. The supra-glottal system transfer function represents the characteristics of the vocal tract and radiation at the lips. The glottal wave is often modelled as a pulse train. However, in this model ug(t) will be considered to be due to the sum of a pulse train pg(t) and a slowly varying function vg(t). The latter term is required because the volume velocity at the glottis does not always go to zero during each cycle of vibration. The function pg(t) will be called the excitation pulse function. Each individual excitation pulse has an associated time of occurrence, its excitation pulse time. In order to make this coincide with the principal excitation of the formant resonance in the vocal tract, the excitation time is defined to occur at the time when the excitation pulse function reaches a zero value at the end of each glottal cycle (see figure 3.1). This time is also the instant of glottal closure, and corresponds to the maximum positive gradient in a laryngograph waveform.

3.2 FUNDAMENTAL PERIOD, FUNDAMENTAL FREQUENCY AND PITCH

3.2.1 Definition of fundamental period

Hess (1983) states that there are three possible ways to define T_0, the speech fundamental period.

1] There is a long term definition, whereby T_0 is the period duration of a signal that is strictly periodic.

2] There is a short-term definition, in which case T_0 is due to the average elapsed time between successive excitations, somehow averaged over a specified short-term window.

3] There is a period-by-period definition, where T_0 is the elapsed time between two successive period markers.

Definition 1] cannot be applied to speech, because it is a quasi-periodic signal and this definition only applies for stationary signals. Definition 2] implies a short-term analysis of the speech signal, whereas 3] can be achieved by means of time-domain analysis of the speech signal. In each case, the associated fundamental frequency F_0 to a fundamental period T_0 is defined as

$$F_0 = 1/T_0$$

3.2.2 Period-by-period or average measurements

Hess and Indefry (1987) discuss several basic approaches to estimating the fundamental period and fundamental frequency values of speech. Their analysis is as follows:

Method 1: Ideally an algorithm is required that can locate individual laryngeal cycles as accurately as possible. Such an algorithm will then be able to measure the natural fluctuations in vocal fold vibration. By detecting the "event" points of glottal closure it is possible to generate cycle-by-cycle fundamental period estimates, that are the times between successive points. In this case the period estimates are in correct phase; that is to say, a period is defined with its start located at one excitation point an its end at the next excitation time. Most algorithms operating on the acoustic speech waveform are unable to perform this function. However, laryngograph-based analyses can quite easily follow this definition.

Method 2: The next best approach, in terms of retaining information concerning the excitation, is to use an arbitrary repetitive point in the speech waveform and calculate the successive period spacing between these points. Most time-domain fundamental period estimation algorithms operate in this manner. This again leads to period-by-period measurements. In this case, the repetitive point may not correspond to the point of excitation in the speech waveform, and its location relative to the excitation point may change depending upon the wave-shape. Consequently, the period estimates may not be in-phase with the excitation points, as there were in the previous case.

Method 3: Another method involves determination of the average length of several successive periods. This operation is implicitly carried out by algorithms that use short-term analysis, such as auto-correlation. The inherent smoothing with this approach results in the loss of fine perturbations in the fundamental period values that occur in speech.

Method 4: Finally fundamental frequency can be determined from a short-term frequency representation of the signal. Again, the window of analysis is required to contain at least one period, which gives a minimum window of around 20ms. The detailed nature of the method varies from technique to technique. This approach also results in smoothed frequency estimates.

3.2.3 The perception of spectral and virtual pitch

There is now a brief discussion of the human perception of pitch. This section is included because it is the limitations of the human auditory system and the perception of pitch that provide the ultimate limit on the performance necessary for a speech fundamental period (or frequency) estimation algorithm for general use.

Pitch perception has been investigated by many researchers for a long time. Many of the earlier theories of pitch perception relate to stationary complex sounds. At present, little is known about the perception of non-stationary sounds with changing fundamental frequency of excitation.

In early research, it was believed that the fundamental harmonic played the dominant part in the perception of pitch. However, Schouten (1938) showed that the phenomenon of pitch perception is not only evoked by the fundamental harmonic (at least not over the range of normal speech), an that the pitch of a harmonic complex remains the same when the fundamental harmonic is removed.

In an attempt to explain this phenomenon, de Boer (1956) proposed that this is not due to non-linear reconstruction of the fundamental harmonic within the ear, and that the perception of pitch is due to a pattern matching process. Subsequent work developed this idea further. In these theories, each harmonic evokes a spectral pitch corresponding to its fundamental frequency. All the spectral pitches then contribute to an overall pitch.

This is knows as the residual periodicity (Goldstein, 1973) or the virtual pitch (Terhardt, 1974).

3.2.4 Some important models of pitch perception

Three models that represent different approaches to pitch perception are now described. All models are characterized by a peripheral analysis that is characterized by a frequency analysis and a stage in which low pitch is estimated. However, the final pattern recognition stage is different in each model.

1] Wightman's pattern transformation model (1973).
There are three stages of processing in this model. Stage 1 is a limited frequency resolution power spectrum analyzer which is an approximation to frequency analysis performed by the peripheral auditory system. Stage 2 consists of a Fourier transform, which is assumed to be realised by means of a specially wired network of neural elements. Stage 3 is then a pitch estimator that operates by finding the positions of maximal activity in the output patterns from stage 2.

2] Goldstein's optimal processor model (1973).
In this model the processor is believed to make an optimal estimate of fundamental frequency on the basis of the noisy representations of the harmonics that are resolved. Under the assumption that the input stimulus is periodic and that adjacent harmonics are present, the model calculates the harmonic numbers and makes use of this information to estimate the fundamental frequency.

3] Terhardt's learning matrix model (1974).
This model is centres on a learning matrix that uses spectral-pitch and lowest spectral-pitch cues as its input (the term spectral-pitch refers to an estimate determined from peak in the short-term spectrum of the signal). The model operates in two phases: The first is a learning phase, which is assumed to be part of the childhood learning process in which a subject acquires the ability to recognize speech. In this phase, the correlations between the two input signals make their impression on the learning matrix.

The second is the recognition phase, in which the learned system generates its pitch estimates. During this phase of operation, the previously impressed traces in the learning matrix can be evoked by similar input stimuli to provide a virtual low pitch. Any given stimulus generates an number of such virtual pitch cues and the strongest determined the final pitch estimate.

A single sinusoidal tone evokes a spectral pitch. A signal such as speech is not a single tone, but rather a complex tone. If we assume for the moment that it will have many harmonics, each of which has it associated spectral pitch. The individual spectral pitches due to the harmonics are then centrally combined to give rise to the sensation of virtual pitch. This is the perceptual equivalent of fundamental frequency.

A definition of spectral and virtual pitch based on a quote by Terhardt (1972a) is as follows:
A single sinusoidal tone evokes a sensation known as the spectral pitch, which is related to the greatest place of excitation in the organ of Corti. The spectral pitches due to the partials associated with a complex sound can be individually perceived by a subject, provided that he makes a conscious effort to do so, unless the difference in frequency between the partials does not fall below a certain level. In addition to the spectral pitches, a stimulus generally evokes a dominant global pitch. In the case of harmonic sounds, this corresponds to the fundamental frequency. This due to a completely different phenomenon to that which evokes spectral pitch, and is known as the virtual pitch.

3.2.5 The pitch of speech

For most purposes, it can be assumed that the pitch and fundamental frequency of speech sounds correspond to each other. However, this is only true if fundamental frequency is defined as the reciprocal of fundamental period. This definition of fundamental frequency only corresponds to the largest common divisor of the partials in the case of strictly periodic signals. The definition of pitch by Terhardt (1979b) provides a good way to combine the temporal properties of the stimulus with its perceived pitch. He states that "The extraction of fundamental frequency is in some respect equivalent to extraction of virtual pitch. In a strict sense, however, the frequency which corresponds to virtual pitch, and the fundamental frequency defined as the largest common divisor of the partials) are in general not identical. ...Hence in the analysis of auditory signals such as speech and music actually the extraction of fundamental frequency is not the real aim but rather extraction of the frequency which corresponds to the virtual pitch".

3.2.6 Difference limens for changes in frequency

The smallest detectable change in the frequency of a stimulus is known as the frequency different limen (DL) for frequency change. For synthetic speech stimuli the fundamental frequency DL has a value of about 0.3% to 0.5% of the fundamental frequency for the fundamental frequency range of male voice; that is over about 40Hz-150Hz (Flanagan & Saslow, 1958). This is less than the difference limen for a pure tone within the same frequency range, which correspond to about 3Hz (Zwicker & Feldkeller, 1967).

Even if changes in fundamental frequency are audible, there are not necessarily linguistically significant. The DL for linguistic significance is an order of magnitude larger than the DL for audibility (McKinney, 1965). This is not that surprising if one considers the fact that if the change is important, then it makes sense that is should be easy for the auditory system to detect.

3.2.7 The precision of speech production

Hess (1983) states that unless the output from a speech fundamental frequency estimation algorithm is to be used in synthesis applications (in which case the result is presented to the ear), or for scientific investigations into vocal fold vibration, there is no need to estimate speech fundamental frequency to a higher accuracy than it can be produced by the vocal apparatus. Various researchers have carried out measurements of the cycle-by-cycle changes in location of the glottal pulses. Gill (1962) found that there are more variations in wave-shape than in length of the glottal excitation.

Lieberman (1963) found that for successive periods, there was a relative difference of more than 1% for 30% of all periods and there was a difference of more than 3% for 10% of the periods. Similar results were found by Hollein et al. (1973) and Horii (1979). Horii found that the mean value of the jitter (the absolute difference in time) between two successive glottal pulses had a value of 51 microseconds at 98Hz and 24 microseconds at 298 Hz. In addition, for 10% of the periods in the data used, the jitter exceeded 100 microseconds.

These perturbations in the excitation are large compared to the frequency DLs for steady-state stimuli, and are audible to a listener. They cannot be individually distinguished, but contribute to the sensation of naturalness (Schroeder & David, 1960). Their effect is quite different from that of quantization noise, as has been observed in the context of speech synthesis (Holmes, 1976).

3.3 PROBLEMS IN SPEECH FUNDAMENTAL PERIOD AND FREQUENCY ESTIMATION

3.3.1 Basic difficulties

The determination of speech fundamental frequency is a difficult problem for many reasons. Speech is a non-stationary signal. That is to say, its characteristics change greatly as a function of time. One reason for this is that the shape of the vocal tract can change rapidly even within the space of a single fundamental period. In addition, the vocal tract can give rise to a wide variety of speech sounds, with a multitude of different temporal structures. The glottal excitation of the vocal tract is often only quasi-periodic. This is particularly true in the case of creaky voice. In addition there are acoustic interactions between the excitation from the vocal folds and the vocal tract.

3.3.2 Requirements for fundamental frequency estimation algorithms

There have been many suggestions as to how the ideal fundamental frequency algorithm should perform (Rabiner et al., 1976). It must be free from gross errors, which occur when the frequency or period estimates deviate substantially from their true values. It must be able to retain the irregularity that exists in the vocal fold vibration. The fundamental period or fundamental frequency values should be as accurate as possible. The algorithm must be able to respond rapidly enough to changes in the excitation period. There should be no voicing determination errors. The measurements should be robust over different speakers, noise and environmental conditions. The algorithm should ideally require as little computation as possible, because this makes it easier (and possibly cheaper) to implement in real-time and for non-real time applications it will need less computer time to run (although this is becoming less important as time goes on, because of improvements in computer technology).

The requirements for a fundamental frequency or period estimation algorithm are all dictated by characteristics of the speech production, speech perception, and the particular application for which the algorithm is intended. The human ear is capable of detecting sounds over a wider frequency range than the vocal apparatus can produce, and can detect changes in frequency that are far smaller than the smallest frequency perturbations that a speaker can intentionally generate.

3.3.3 Sources of gross errors in fundamental period and period estimation

There are various reasons why a particular algorithm may generate gross errors. Firstly, when there are adverse signal conditions, which can occur when there is a strong first formant, a rapid change in articulator positions or in the case of band-limited or noisy speech. Secondly, when there is inadequate algorithm performance, perhaps because the analysis window is too small in a short-term algorithm, or because of the absence of some feature used in the estimation process. Thirdly, because the algorithm is unable to deal satisfactorily with creaky voice. In this case, the inherent averaging in some algorithms may cause erroneous output to be generated.

In addition difficulties can arise due to the recording conditions. Quite often the speech signal is degraded by amplitude and phase distortions, and background noise is almost always present to some extent. It is particularly difficult to get algorithms to operate well over telephone lines, because of phase and amplitude distortions, fading, and break-through from other signals.

Strong first formant in vicinity of second harmonic

Gross errors can arise when there is a strong first formant in the vicinity of the second harmonic, which results in its amplitude becoming significant or greater than that of the fundamental harmonic. This can lead to what are known as "doubling" errors, because this leads to a significant second peak in each period, which time-domain algorithms sometimes confused with the main peak. This is illustrated in figure 3.2. For comparison, a temporally simple speech pressure waveform is shown in figure 3.3. Frequency-domain and short-term algorithms face a similar problem with this class of signals, because the second harmonic dominates the short-term spectrum. In this thesis, gross errors that exceed the true values are known as chirp errors.

The complementary type of errors to chirp errors are defined in this thesis as drop errors. In a time-domain algorithm they will occur whenever it misses out a period marker, giving the impression that the period is longer than it truly is. This situation can arise when there are rapid envelope changes in the speech waveform, and it is especially associated with voiced sounds made with articulations that result in obstruction of the vocal tract, such as the sound /r/. It can also occur due to missing secondary excitations in creaky voice quality speech or during diplophonic voicing (which is the tendency to generate pairs of pulses that can occur during even normal voice).

3.3.4 The required operating frequency range

The range of possible fundamental frequencies for human speech is wide. For an arbitrary utterance, the range over a large population of subjects can lie between 33Hz to 3100Hz by Moerner, Fransson & Fant, 1964. However, another investigation due to Catford,1964 (that did not include creaky voice) confined the range to between 70Hz and 1100Hz. For the purposes of singing, a somewhat wider range is required. Hess, 1983 gives the range of 50Hz to 1800Hz to cover a bass to a soprano.

For an individual speaker, the distribution of fundamental frequency depends upon the experimental conditions. It is particularly relevant whether the speech was taken from conversation or from read text. The frequency distributions from read text rarely exceed an octave range. Provided the distribution is plotted on a logarithmic scale, this fundamental frequency distribution comes close to a normal distribution (Risberg, 1961; Schultz-Colson, 1975)

Algorithms that perform speech fundamental frequency estimation usually restrict their operation to a sub-range of the possible fundamental frequency values. A good working range for an algorithm is between 50Hz and 800Hz, because this covers the range of most adult conversational speech (Hess, 1983).

3.3.5 Required measurement resolution and accuracy

The accuracy and resolution requirements for a fundamental frequency algorithm are determined by its intended applications. The human auditory system is more sensitive to changes in absolute frequency at low frequencies, and in general the noticeable difference in frequency is proportional to frequency. The difference limen with respect to the fundamental frequency (DL) for human listeners perhaps represents the ultimate required performance, which is typically 0.3-0.5% resolution of the fundamental frequency for steady state harmonic sounds. Most algorithms do not meet this specification. However, for most applications, less accuracy can be tolerated.

The difference limen for linguistic significance is greater than for that of perception (McKinney, 1965). Thus for prosodic analysis, an accuracy of a few percent may be adequate.

The required frequency (or time) resolution required is dependent upon the required application of the algorithm. For intonation training, a resolution of 3-4% will suffice (for example in a Voicscope, Abberton & Fourcin, 1973). There are also limits on the resolution of fundamental frequency values that can be displayed with such schemes, due to the limited number of pixels available for the graphics display.

Consideration to human frequency difference limens suggest that a frequency resolution of 0.3%-0.4% of the fundamental frequency value would be ideally required by a fundamental frequency or period estimation algorithm.

Requirements for profoundly deaf EPI patients

The required frequency resolution for the profoundly deaf patients for whom high technology signal processing hearing aids are intended is only about 1% of the fundamental frequency values within the male frequency range and poor above about 200Hz, which is several times worse than for normal listeners.

3.3.6 Accuracy limitations due to time quantization of sampled signals

There is an intrinsic accuracy limit in time-domain fundamental frequency estimation algorithms that operate using sampled digital signals which is due to the time quantization of the input signal. This introduces uncertainty into the location of an event in time. For example, at a sampling frequency of 10kHz, it is only possible to locate a time event to 1/10000 = 100 microseconds. For a fundamental frequency of 100Hz, this corresponds to an accuracy of 1%. At higher fundamental frequencies, this percentage error increases still further. Even at 100Hz, this error is greater than the auditory DL for frequency change. The same problem arises for short-term analysis algorithms that operate in the lag domain (for example auto-correlation, cepstral analysis, etc).

There is a similar problem in the case of frequency-domain analyzers. In this case, a sampling rate of 10kHz and an analysis window of 100ms (which is very long for the short term analysis of speech) gives rise to a frequency resolution of 10Hz. Consequently, in this case it is the lower frequencies that give a proportionally larger quantization error. Thus there is a 10% error at 100Hz, and a 2% error at 500Hz. With regard to this accuracy issue, Hess and Indefry point out (1987) that to reduce sampling accuracies to 0.5% up to the fundamental frequency of 500Hz requires a sampling period of 10 microseconds.

Many algorithms use interpolation at their outputs to improve the time or frequency resolution of their estimates. Interpolation can easily be carried out in the case of frequency-domain algorithms and those employing short-term analysis. Interpolation is more difficult to use in time-domain algorithms, although the accuracy of location of peaks and zero-crossings can be increased using interpolation. Another approach to reducing quantization errors is by smoothing the frequency estimates, although this approach is not always guaranteed to improve accuracy.

3.3.7 Required maximum rate of change of speech fundamental period

In regularly excited speech (not creak), the maximum rate of change of period length is typically taken to be a 10% to 15% change between successive periods (Reddy, 1967).

The maximum rate of change of frequency of the normal voice source was found to be about 1% per millisecond by Sundberg (1979). However, in voice qualities such as creaky, as well as in pathological speech, there can be much larger change per period than this figure suggests.

The maximum rate of change on fundamental period usually presents no problems to time-domain analyzers, because they operate on a period-by-period basis. However, they do put an upper time window limit on short-term analysis procedures of around 20ms -30ms.

3.4 CATEGORIZATION OF SPEECH FUNDAMENTAL FREQUENCY ESTIMATION ALGORITHMS

3.4.1 Preliminary classification

McKinney (1965) states that a 'pitch' determination algorithm can be essentially decomposed into three stages. These are the pre-processor, the basic extractor and the post-processor, as illustrated in figure 3.4. The main task of the measurement is performed by the basic extractor stage. The main function of the pre-processor is one of data reduction, and the emphasis of features in the input speech to facilitate the operation of the basic extractor. The post-processor combines many functions, such as error correction and the generation of output in the desired format.

3.4.2 Types of algorithm

The techniques that have been developed to determine speech fundamental frequency are broadly classified into four main groups by Hess, 1983; Those that operate in the time-domain, those that operate over some short-term window of the speech, which he calls short-term analysis, those which are hybrids of the first two, and finally those that operate by direct measurement of vocal fold activity. The is often no clear-cut distinction between the first two types. It is important to understand what is meant by the terms short-term, time-domain and frequency-domain.

Time-domain algorithms employ direct measurements on the speech signal and involve looking for temporal features in the speech pressure waveform (or in the filtered waveform), such as local maxima and minima.

Short-term analysis procedures use some form of transformation of the data within a short (for example, 20ms) time window. The nature of the transformation depends on the particular method used. The estimate obtained with such an approach consists of a sequence of average fundamental period or frequency values obtained over the input interval.

Frequency-domain algorithms make explicit 'frequency' estimates. There may be a frequency-domain interpretation to certain short-term operations which are implicit. For example, the auto-correlation technique can be implemented via a frequency-domain representation.

The time-domain refers to analyses which use the same time base as the input speech signal. A time-domain analyzer gives rise to an output signal that consists of a series of excitation markers that delineate period boundaries. Time-domain operation thus generally presumes the local definition of fundamental period and gives rise to a period-by-period fundamental period estimates.

The next chapter will examine some time-domain, short-term and laryngeal algorithms in more detail.

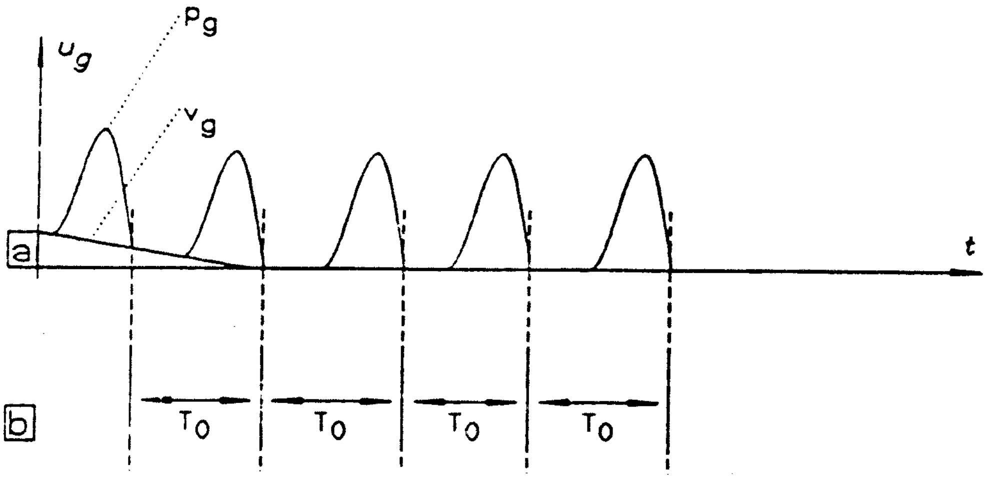

Figure 3.1 Diagram showing voice source parameters.

This illustrates; a) the excitation signal, and b) the corresponding period durations. (After McKinney, 1965).

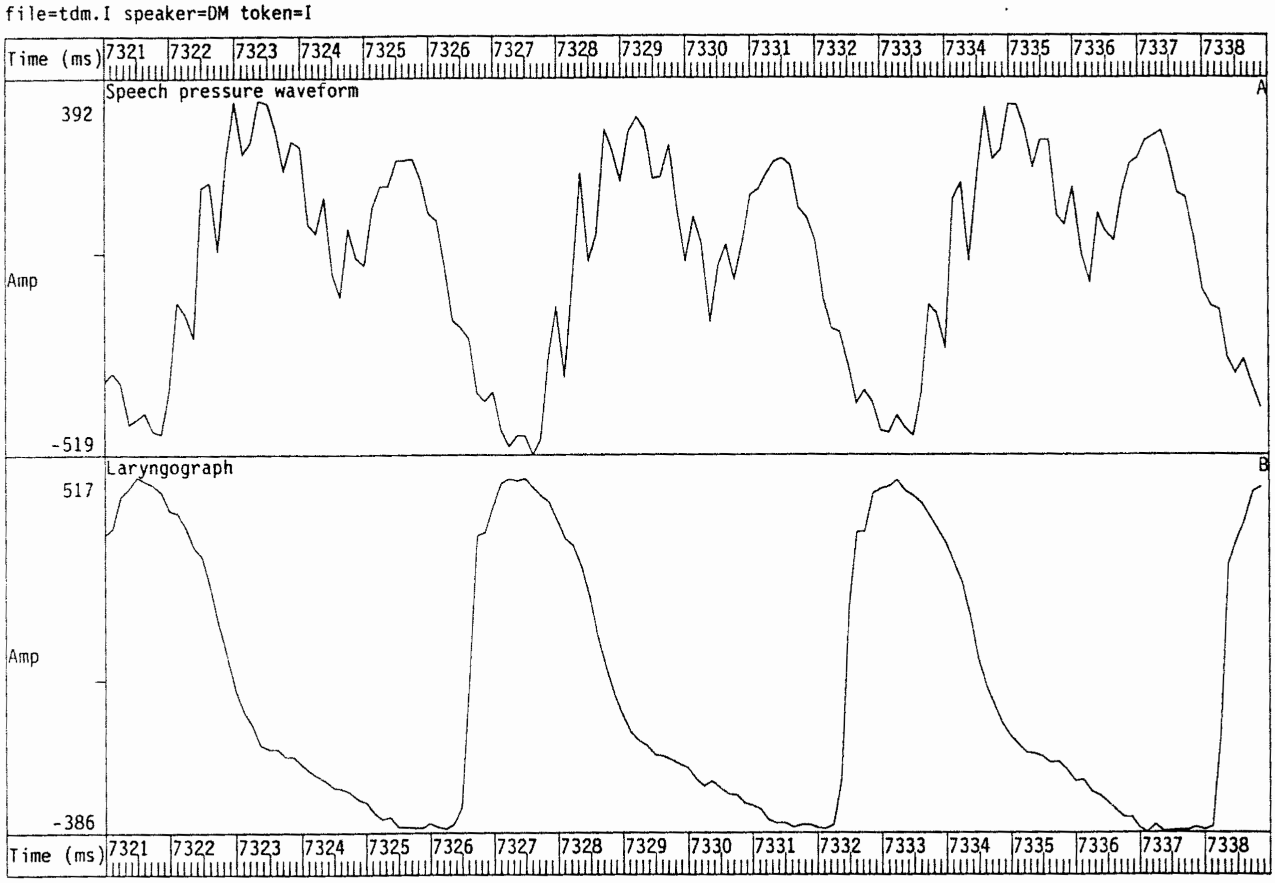

Figure 3.2 Speech pressure waveform exhibiting two peaks per fundamental period. The speech is shown in trace A. The corresponding laryngograph waveform in shown in trace B. This situation arises when the first formant coincides with the second harmonic in the excitation spectrum. This situation can lead to "doubling error" in simple fundamental period estimation algorithms. The speech is the vowel /I/ from a male subject.

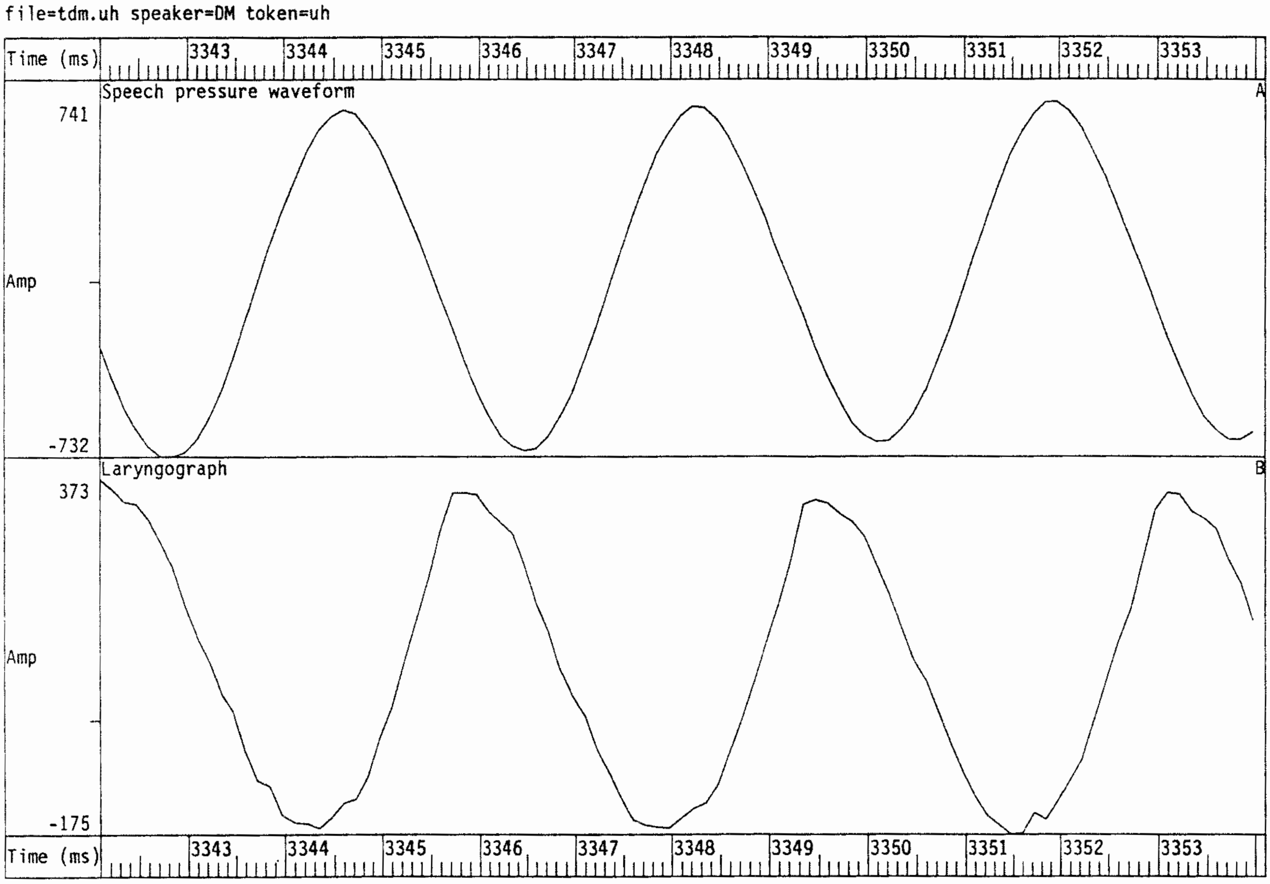

Figure 3.3 Temporally simple speech pressure waveform.

Speech is shown in trace A. The corresponding laryngograph waveform in shown in trace B. It is relatively would be easy to determine the fundamental period of the speech in this case, even with a simple fundamental period estimation algorithm. The speech is the vowel "u", as in the word "but", from a male subject.

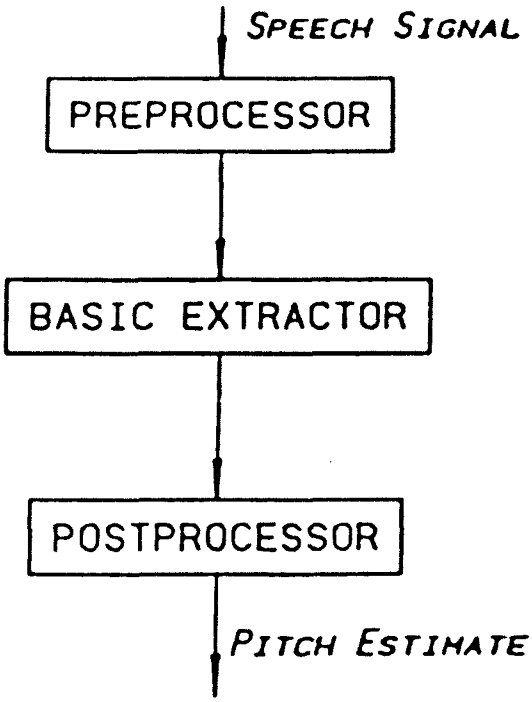

Figure 3.4 Block diagram illustrating the basic stages involved in speech fundamental frequency/period, estimation.

The pre-processing stage is involved with data reduction and extraction of important features of the speech signal. The basic extractor essentially performs the main task estimation of period or frequency. Finally, the post-processing stage converts the output from the basic extractor into a desirable format and may also perform error correction and smoothing of the raw estimates.

(Taken from Hess, 1983; After McKinney, 1965).