Chapter 3

Acoustic-Prosodic Analysis: An overview of related work

The automatic analysis of speech in computer aided pronunciation teaching uses a digitally sampled acoustic waveform as the only input parameter. An acoustic waveform can be easily captured as an electronic signal with a microphone without being cumbersome or obtrusive to a speaker. Other characteristic aspects of speech, such as glottal activity, nasality and air flow, cannot be measured without using cumbersome (sparsely available) transducers such as a laryngograph which ties around the neck, a nasograph which pinches below the bridge of the nose, or an obtrusive face mask.

The acoustic speech signal conveys linguistic and paralinguistic information. The acoustic waveform is perceived as a sequence of phonemic segments which constitute the words of an utterance. It also exhibits suprasegmental phenomena which are related (but not exclusively) to the syntactic and prosodic structures of an utterance. It is emphasised, therefore, that the acoustic waveform is an extremely complex signal which conveys many aspects of speech other than prosody alone. In order to perform automatic prosodic analysis of speech, it is necessary to identify acoustic parameters which are correlated with prosody. Research conducted to identify acoustic correlates of stress is reviewed in Section 3.1.

The fundamental frequency of an acoustic speech signal is correlated with both the perceived prominence of syllables and the characteristic melody of an utterance. Its extraction from the acoustic waveform is therefore a prerequisite for prosodic analysis. The functionalities of a selection of fundamental frequency determination algorithms are reviewed in detail in Section 3.2.

Prosodic aspects of speech are described in a syllabic domain. Thus, the automatic identification of syllables in connected speech forms a part of prosodic analysis. A number of algorithms to partition speech into syllabic units are reviewed in Section 3.3.

It is asserted that the acoustic correlates of stress are influenced by factors other than stress and interact in the manifestation of stress. The fundamental frequency of speech is therefore not a direct representation of intonation and acoustic correlates of stress do not co-exist in a simple relationship to represent stress. The composite structure of fundamental frequency is discussed in Section 3.4. Algorithms proposed to automatically transcribe the prosodic structure of an utterance from acoustic parameters are reviewed in Section 3.5.

3.1 Acoustic parameters for prosodic analysis

Extensive research has been conducted to identify the acoustic correlates of lexical stress in isolated words and of sentential stress in connected speech.

Fry (1955) examines acoustic features correlated with a shift in lexical stress from one syllable to another which is commonly associated with a change in function from noun to verb in disyllabic words in English such as *digest* and *permit*. Fry's results show that the location of the lexical stress is correlated with the ratio of the duration of the vowel in the syllables and the ratio of the peak of energy in the syllables (which occurred in the vocalic part of the syllables). He also claims that the duration feature is a more effective cue to stress location than the energy feature. In a later experiment, Fry (1958) investigates the correlation of fundamental frequency with stress shift in the same type of words. He finds that the syllable with the greatest fundamental frequency is more likely to be perceived as prominent although the extent of the jump in fundamental frequency (*FØ*) between the two syllables is relatively unimportant. Moreover, Fry suggests that the occurrence of an *FØ* glide through a syllable is likely to make it perceptually more salient than the other syllable (in the disyllabic words he examines) and that such glides may "outweigh" the duration cue.

Fry also draws attention to a further correlate of stress, vowel quality.

> "[I]n English ...[t]he substitution of the neutral vowel /ə/ for some other vowel, the reduction of a diphthong to a pure vowel, or the centralisation of

a vowel are all powerful cues in the judgement of stress."

(Fry, 1958, pp.128)

The effect of vowel quality on the perception of lexical stress is also investigated by Fry (1965). The experiment he conducted examines the relation between stress location and changes in the *F1/F2* formant space of monophthongs using synthetic stimuli. Although Fry finds that stress judgement is dependent on vowel formant structure, he suggests that the perception of stress is affected more by duration and energy cues than by vowel quality.

The four acoustic features associated with the perception of stress (identified by Fry as duration, energy, fundamental frequency and vowel quality) have been the subject of investigation for many researches (Lieberman, 1960; Morton & Jassem, 1965; Westin *et al.*, 1966; Nakatani & Aston, 1978; Aull & Zue, 1985).

Lieberman (1960) proposes an algorithm to identify the location of the lexical stress in minimal noun/verb pairs of disyllabic words such as ˈ*compound* and *com*ˈ*pound*, using duration, energy and fundamental frequency (but not vowel quality). Lieberman suggests that the duration of the entire syllable (cf. the duration of the vowel used by Fry) acting as a unidimensional cue for the location of the most prominent syllable is a less effective cue than the peak syllable amplitude. This is the converse to the claim of Fry (1955). Lieberman also suggests that the peak syllable fundamental frequency is a more effective cue to stress location than the peak syllable amplitude. Lieberman recognises the complex interaction that exists between these acoustic features. In some cases, the prominent syllable has a higher peak *FØ*, but the unstressed syllable has a higher peak amplitude; and in other cases, the prominent syllable has a higher peak amplitude, but the unstressed syllable has a higher peak *FØ*. This is analogous to Fry's observation that *FØ* glide through a syllable may "outweigh" the duration cue. In other words, *FØ* may also "outweigh" the energy cue, or visa versa. By combining these acoustic features, Lieberman is able to correctly identify the location of the lexical stress in 99.2% of the minimal noun/verb pairs he studies.

Adams & Munro (1978) examine the acoustic correlates of sentential stress (in connected speech). Their investigation considers fundamental frequency during the course of syllables, the absolute duration of entire syllables and the amplitude envelope in syllables. Adams & Munro do not consider the effects of vowel quality mentioned by Fry. The results of Adams & Munro suggest that the degree of *FØ* change is a more significant correlate of sentential stress than absolute *FØ* height.

> "[T]he initial level and the end level were higher in type I [*rising FØ*] and lower in type II [*falling FØ*] for stressed than for unstressed syllables. Furthermore, the amount of rise in type I and fall in type II was found to be greater in stressed than in unstressed syllables. ... In type V [*level FØ*], the fundamental frequency level was greater in stressed than in unstressed syllables, but *not significantly so*." [my italics]
>
> (Adams & Munro, 1978, pp.140)

The results of Adams & Munro also indicate that the amount of fall from a steady peak in the amplitude envelope is a more significant correlate of sentential stress than the amplitude at the peak of the envelope, and that sentential stress is also signalled by syllable duration.

There is noticeably little reference to vowel quality in these studies on the perception of stress. Most of the work relating stress with vowel quality is found in studies on the perceptual classification of vowels. Such studies take a different view of the relation between vowel quality and stress to the view taken in this thesis, in that they are primarily concerned with understanding how stress, and other factors, affect vowel quality; whereas here an emphasis is placed on understanding how vowel quality, and other factors, affect the perception of stress.

Tiffany (1959) observes a tendency for the vowel diagram (a conventional plot of the first formant *F1* against the second formant *F2*) to grow smaller from vowels enunciated in isolation to prominent vowels to non-prominent vowels. The prominent and non-prominent vowels which Tiffany investigates are in a [h]-vowel-[d] phonetic context embedded in carrier phrases with and without emphatic stress placed on the word containing the relevant vowel. Tiffany suggests that vowels move towards a neutral, or at least a central, point on the vowel diagram as they lose energy. Prominent vowels may not merely be longer in duration, higher in fundamental frequency, and greater in energy than unstressed vowels; they may also be different in terms of vowel resonance patterns as well.

Shearme & Holmes (1962) observe that *F1/F2* measurements for vowels in connected speech and in varying phonetic contexts are displaced from corresponding measurements for vowels in isolated monosyllables towards a central vowel position. They attribute this affect to variability in both stress *and* phonetic context.

It is argued by Lindblom (1963) that as rate of articulation increases, physical limitations impose less time for the articulators to complete their movements within a consonant-vowel-consonant syllable, thus the vowel becomes shorter and hence the vowel becomes more susceptible to contextual assimilation. The acoustic consequence of this is that the placement of the vowel in the $F1/F2$ formant space moves away from a target position, typically towards a more central vowel position. Lindblom argues that vowel reduction is primarily determined by a duration feature and that it is immaterial to discuss whether the duration of a vowel is produced chiefly by the degree of stress or the rate of articulation. If Lindblom's explanation of vowel reduction in terms of a shortening of duration and the physical limitations of the articulators, rather than variations in the degree of stress and the rate of articulation *per se*, is correct, then it is hypothesised that duration acts as a stronger cue to stress than vowel quality and that the correlation between duration and vowel quality is positive in their association with stress. In other words, no more information about the degree of stress of a syllable will be obtained from a measure of vowel quality than can already be obtained from a duration feature.

Moreover, if Lindblom's explanation is correct then it is inferred that variations in the degree of stress and the rate of articulation affect the vowel quality in identical ways. Verbrugge & Shankweiler (1977) report that measurements of vowel formant frequencies reveal relatively little difference between fast and slow speech, but reveal large vowel formant shifts in unstressed syllables relative to stressed syllables. Tuller *et al.* (1982) examine whether stress and speech rate variations involve different transformations of physiological signals. The hypothesis that all changes in vowel duration are the product of the same production rule, as argued by Lindblom, is not supported by their results. Variations in stress and speech rate apparently produced vowel duration changes through different effects on muscle (genioglossus and orbicularis oris) activity. Evidence against Lindblom's hypothesis is also presented by van Summers (1987) who argues that duration lengthening due to final consonant voicing in consonant-vowel-consonant syllables has greater effects on articulatory movements and formant structure in the final portions of vowels than on the initial portions, while duration lengthening due to increased syllable prominence has a more global influence throughout vowel production.

Vowel reduction is associated with the degree of stress and the rate of articulation (Lindblom, 1963; Tuller *et al.*, 1982; van Summers, 1987). Therefore, if vowel quality is to function as a cue to stress location, it is first necessary to understand how vowel quality is affected by factors other than stress. If the reduction in the quality of a vowel can be attributed to variations in the rate of articulation, for example, rather than variations in the degree of stress then there is little possibility of being able to use vowel quality as a cue to stress.

The consensus drawn from these studies is that duration, fundamental frequency, energy and the quality of the syllable nucleus correlate with lexical and sentential stress, although the degree to which each of these parameters correlates with the stress varies considerably from study to study. One of the major problems in the automatic analysis of stress is that there are complex interactions between each of these acoustic parameters in their association with stress, and variations in each of these parameters can be associated with factors other than stress.

For example, the duration of the vocalic portion of a word is influenced by the inherent or intrinsic duration of the intended vowel, phonetic context (Lehiste, 1970), rate of articulation, the sentential stress pattern, the position of the word within the sentence, and other possible factors such as speaker style. Although duration may be ambiguous as an unidimensional cue to stress, syllable final consonant voicing, or any of the other factors affecting duration, it may provide useful information for, say, stress when used in combination with other acoustic cues. If the presence of stress increases vowel duration along with having some specific influence on formant trajectories which final consonant voicing does not, then the stress related vowel lengthening could be disambiguated from voicing related lengthening.

Lehiste & Peterson (1959) investigate intrinsic vowel amplitude and suggest that energy cues to the type of stress investigated by Fry can be more readily identified by first compensating for intrinsic vowel amplitudes. This philosophy can be extended to include any acoustic cue to stress. If acoustic features are normalised for variations which are due to factors other than stress, then the normalised acoustic features can be combined to form a description of the stress pattern of an utterance.

In summary, the phenomenon of stress is realised in the speech waveform by variations in duration, energy, fundamental frequency and vowel quality. The process of locating prominent syllables is complicated both by the interaction of these features in the manifestation of stress and by the fact that all four features are also influenced by factors other than stress. Moreover, there are a number of specific problems in measuring these acoustic features in order to automatically locate prominent syllables. It is not clear how the acoustic features are determined, over what phonetic domains the acoustic features are related with respect to prosody, or how they are best normalised for non-prosodic aspects of speech. A principal aim of this thesis is to address these problems.

Methods to automatically determine the fundamental frequency of a speech waveform are addressed in Section 3.2 and in Chapter 7. The domain of phonetic units whose duration, energy and vowel quality are to be determined as optimal correlates of stress and normalisation techniques applied to them are investigated in Chapters 4, 5 & 6.

3.2 Automatic extraction of fundamental frequency

The *fundamental frequency* (*FØ*) of speech is defined as the rate of glottal pulses generated by the vibration of the vocal folds during the voicing of segments. The *pitch* of speech is the perceptual correlate of *FØ*. The psychoacoustic scales of pitch are linear only at relatively low frequencies. However, it is assumed that there is a linear correlation between pitch and *FØ* at the low ranges of frequency that are relevant to the voicing of male and female speech (approximately 50–250Hz and 120–400Hz respectively).

It can be concluded from the previous Sections that the fundamental frequency of speech plays an important role in the prosodic features of stress and intonation. Its extraction from the acoustic speech waveform is therefore necessary as an initial process for the analysis of these suprasegmental phenomena. However, determining *FØ* is not a simple task, and many approaches (referred to here as Fundamental frequency Determination Algorithms (FDAs)[1]) have been reported (Hess, 1983). The diversity

[1]Methods of extracting fundamental frequency are commonly referred to in the literature as Pitch

and complexity of methods used in an attempt to determine *FØ* stem from the non-stationarity of speech, characterised by non-uniform intensity, by small variations in fundamental frequency between successive periods (such as those particularly noticeable in creaky voice (Laver, 1980)) and by a constantly changing spectrum which is dependent upon articulation.

A selection of FDAs are described below. The choice of FDAs reviewed here is influenced by the availability of existing implementations, by the desire to examine methods of fundamental frequency extraction which use radically different techniques, and by the ease of implementation from the original descriptions of the algorithms. The algorithms investigated are:

- Cepstrum-based *FØ* determinator (CFD) (Noll, 1967).
- Harmonic product spectrum (HPS) (Schroeder, 1968; Noll, 1970).
- Feature-based *FØ* tracker (FBFT) (Phillips, 1985).
- Parallel processing method (PP) (Gold & Rabiner, 1969).
- Integrated *FØ* tracking algorithm (IFTA) (Secrest & Doddington, 1983).
- Super resolution *FØ* determinator (SRFD) (Medan *et al.*, 1991).

The algorithms CFD and HPS make use of frequency domain representations of the speech signal. FBFT and PP produce fundamental frequency estimates by analysing the waveform in the time domain. IFTA and SRFD uses a waveform similarity metric based on a normalised crosscorrelation coefficient. This selection of FDAs is considered to represent a cross-section of the multitude of algorithms developed over recent years. A detailed performance evaluation of these FDAs is described in Section 7.2. The remainder of this Section provides a summary of the operation of each of these algorithms in order to illustrate their diversity and complexity.

Determination Algorithms. However, these algorithms do not determine the linguistic phenomenon of pitch, which is the perceptual correlate of *FØ*.

3.2.1 Cepstrum-based *FØ* determinator (CFD)

Noll (1967) proposes a method to determine the fundamental frequency of a speech waveform based on the use of *cepstra*. A cepstrum is defined as the Fourier transform (FFT) of the logarithm power spectrum of a signal. The method assumes that a speech signal can be represented by the commonly understood source-filter model (Fallside & Woods, 1985) in which the speech signal $f(t)$ is equal to the convolution of the vocal source signal $s(t)$ and the impulse response of the vocal tract $h(t)$.

$$f(t) \;=\; s(t) \bigotimes_{\omega} h(t) \tag{3.1}$$

The objective of the algorithm is to effectively deconvolve the vocal tract response from the source signal, and thus find the fundamental frequency of the speech.

Following from Equation 3.1, the source-filter model can be represented in the frequency domain (by application of the convolution theorem) as,

$$F(\omega) \;=\; S(\omega).H(\omega) \tag{3.2}$$

In order to identify or separate the effects of the vocal tract and the vocal source signal, a Fourier transform is performed on the logarithm of the power spectrum, $|F(\omega)|^2$,

$$\begin{aligned}
\mathcal{FFT}\left\{\log|F(\omega)|^2\right\} &= \mathcal{FFT}\left\{\log(|S(\omega)|^2.|H(\omega)|^2)\right\} \\
&= \mathcal{FFT}\left\{\log|S(\omega)|^2 + \log|H(\omega)|^2\right\} \\
&= \mathcal{FFT}\left\{\log|S(\omega)|^2\right\} + \mathcal{FFT}\left\{\log|H(\omega)|^2\right\} \qquad (3.3)
\end{aligned}$$

Thus, the vocal source and vocal tract effects are now added rather than convolved. The effect of the vocal tract appears in the cepstrum as numerous, closely packed peaks at the lower quefrency[2] end, and the location of a peak occurring at the higher end of the cepstrum marks the fundamental period of the vocal source.

An illustration of the practical application of this method is shown in Figure 3.1. Cepstral analysis is performed on successive frames of data. The duration of the analysis

[2]The quefrency refers to the frequency of the ripples in a spectrum. Since the ripples are in the frequency domain, quefrency is in the time domain.

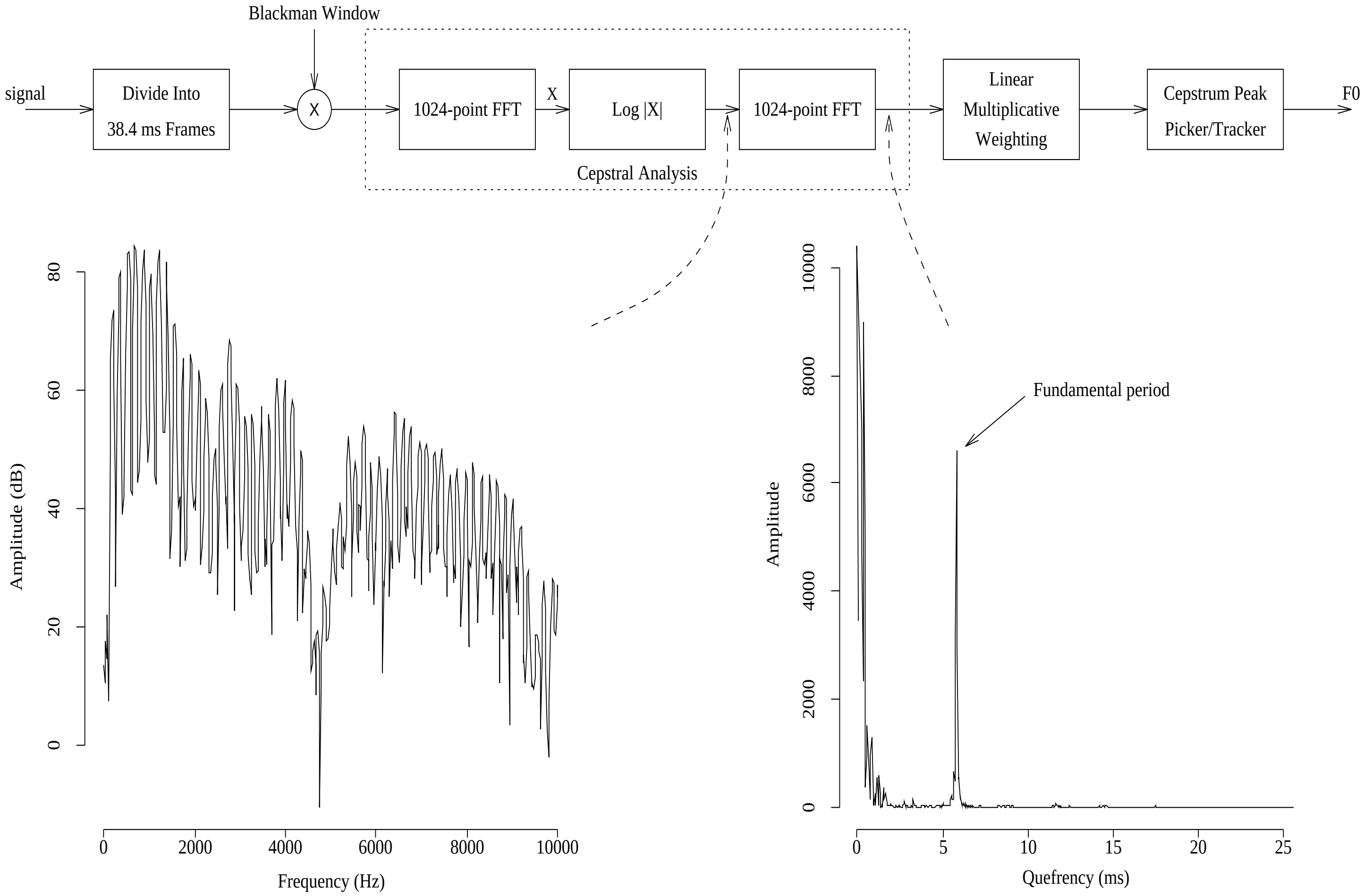

Figure 3.1: Cepstrum-based $F\emptyset$ determinator (CFD)

frame must be chosen with care. If the duration is too long, then the non-stationarity requirement of the analysis frame for the Fourier transform will be violated, particularly during segmental transitions, and spurious artefacts will be produced in the spectra. However, if the duration is too short, then the analysis will be inaccurate. In practice, a duration of 38.4ms is used as a reasonable compromise, as it corresponds to $\frac{3}{4}$ of the data in a 1024-point Fourier transform for data sampled at 20kHz, and corresponds to approximately two fundamental periods at the lower levels of *FØ* for male speech.

Cepstral analyses of sections of a speech waveform are overlapped by spacing the beginnings of successive analysis frames at intervals of 6.4ms ($\frac{1}{8}$ of the FFT data). An estimate of the fundamental frequency at the time which corresponds to the mid-point of the analysis frame is therefore made at regular intervals. The *FØ* estimates are plotted against these times to form an *FØ contour*.

Each analysis frame is applied through a Blackman window (Harris, 1978) and a 1024-point Fourier transform is performed on the data with padded zeros. The use of a 1024-point FFT results in a cepstrum of 512 coefficients (0–25.6ms) with a quefrency resolution of $\frac{1}{20}$ms (for a speech waveform sampled at a frequency of 20kHz). The quantisation error for each estimation of the fundamental period is therefore $\frac{1}{20}$ms. After cepstral analysis of the data, a linear multiplicative weighting is applied over a speaker-dependent quefrency range. For example, a quefrency range of 4.0ms (250Hz) to 20.0ms (50Hz) is selected for the analysis of male speech, and a range of 2.5ms (400Hz) to 8.3ms (120Hz) is selected for female speech. Noll (1967) proposes that the amplitude of the cepstral coefficients is weighted by 1.0 at the lower quefrency limit and on a linear scale up to a weight of 5.0 at the upper quefrency limit. The peak value in the resultant weighted cepstrum is located over the selected quefrency range. If the amplitude of the peak exceeds a speaker-dependent *a priori* threshold, then the peak location corresponds to the fundamental period for that frame; otherwise the frame is assumed to represent unvoiced speech.

3.2.2 Harmonic product spectrum (HPS)

The fundamental frequency of a periodic signal can be determined by measuring the frequencies of its higher harmonic components and computing the greatest common divisor of these harmonic frequencies (Schroeder, 1968). The greatest common divisor can be determined by making an entry to a frequency histogram for each harmonic frequency and at integer divisions of the harmonic frequency. The frequency at the peak of the histogram represents the greatest common divisor of the harmonic frequencies, and hence the fundamental frequency. However, the task of determining the harmonic frequencies is not simple. The problem of determining these frequencies can be avoided by noting that the frequency components of a signal represented in the frequency domain have higher amplitudes at points near to harmonic frequencies. Each entry to the histogram can thus be weighted by a monotonicly increasing function of the component amplitude or of the logarithm of the component amplitude.

The block diagram shown in Figure 3.2 illustrates how the HPS based *FØ* determinator is arranged in practice. The speech signal is divided into 38.4ms frames with an analysis frame shift of 6.4ms, as in the cepstral based approach (CFD). A 4-term Blackman-Harris window (Harris, 1978) is applied to each frame in order to filter out discontinuities (and hence high frequency artefacts) at the limits of the analysis frame. The harmonic histogram is formed from a short term log power spectrum $20\log_{10}|F(nf)|$ and represented as the 'harmonic product spectrum' (Noll, 1970),

$$\log_{10} P(f) = \sum_{n=1}^{N} 20.\log_{10}|F(nf)| \tag{3.4}$$

$$P(f) = \prod_{n=1}^{N} |F(nf)| \quad (+\,10^{20}, \text{ which is ignored}) \tag{3.5}$$

The low-quefrency structure of the log power spectrum of speech only adds confusion when compressed to form the harmonic product spectrum, so such terms are removed by smoothing the spectrum with a window function $W(f)$ (for example, a 4-term Blackman-Harris window) and then subtracting the smoothed spectrum from the original spectrum prior to calculating the harmonic product spectrum. The low-quefrency lifted log spectrum is expressed as,

$$L(f) = 20\log_{10}|F(f)| - W(f).20\log_{10}|F(f)| \tag{3.6}$$

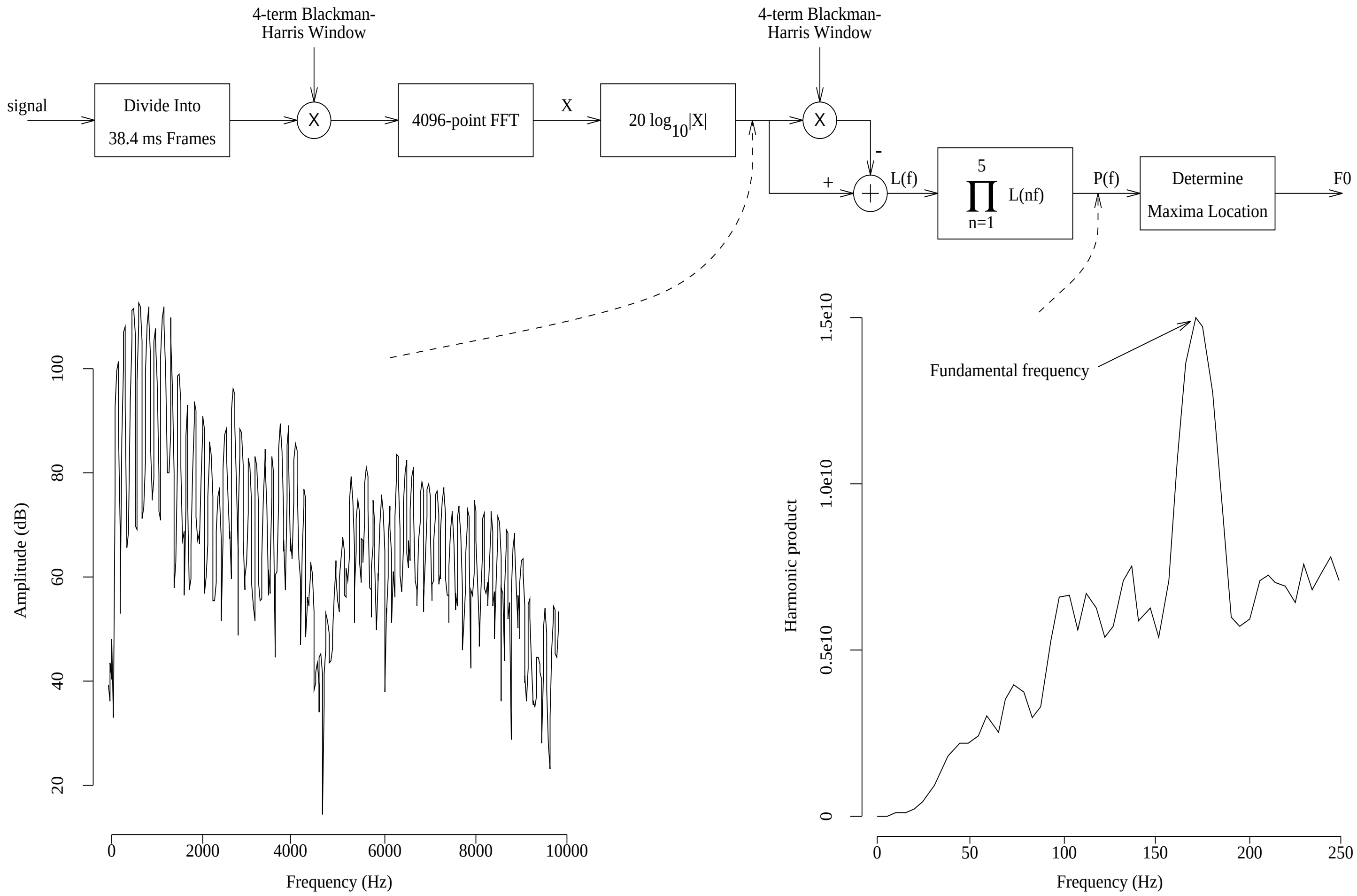

Figure 3.2: Harmonic product spectrum (HPS)

and the harmonic product spectrum, as,

$$P(f) = \prod_{n=1}^{N} L(nf) \tag{3.7}$$

where N is a compression factor (set to 5 in a practical application). If the compression factor is set too low, the estimates of *FØ* will be unreliable, and if the compression factor is set too high, the process of determining *FØ* by this method will be time consuming and computationally expensive.

The fundamental frequency is given by the position of the maxima of $P(f)$, which is searched for in the range of 50Hz to 250Hz for male speech, and between 120Hz and 400Hz for female speech (for compatibility with the other FDAs reviewed in this Section). If this maxima has an amplitude that is greater than some speaker-dependent *a priori* threshold, then the section of speech being analysed is assumed to be voiced. For a speech waveform sampled at 20kHz and with the use of a 4096-point FFT, the fundamental frequency can be estimated by this method with a quantisation error of 4.88Hz (ie. the sampling frequency divided by the number of points in the FFT).

3.2.3 Feature-based *FØ* tracker (FBFT)

An algorithm is proposed by Phillips (1985) which aims to locate glottal pulses in the time domain by using a feature-based statistical approach. The algorithm uses perceptually motivated features that are designed to capture the same information that a person would use to indicate the presence of a glottal pulse from a waveform display.

As is the case with many FDAs which operate on a time domain representation of speech (such as those reviewed in this and the following Sections), the sampled waveform is initially low-pass filtered to simplify its temporal structure. The filtering process reduces the effects of higher formants (*F2*, *F3*, and *F4*) for vowels, and removes the high frequency components of voiced fricatives, thus aiding the classification of voiced speech by the algorithms. In the studies reported here, low-pass filtered speech is produced by a finite impulse response (FIR) filter with a -3dB cut-off at 600Hz and rejection greater than -85dB above 700Hz, though these filter characteristics are not crucial to the performance of the FDAs.

Peaks are located in the speech waveform under the criteria that they correspond to local maxima with signal amplitudes greater than 120 ADC-units[3] (approximately 50dB signal-to-noise level) and that the signal amplitude drops by at least 20% after the local maxima. Peaks satisfying these criteria are represented by the set S_1 and are indicated on the speech waveform in Figure 3.3 by ⇣and ↧.

A member of S_1 is accepted into a subset, S_2, if no other peak occurs within the 3ms region before it, or if its amplitude is greater than 75% of the amplitude of the largest peak within the 3ms region before it. The members of the subset S_2 are indicated in Figure 3.3 by ↧.

Perceptually motivated features are used to identify the first peak of each fundamental period in the speech waveform during voiced portions of speech. The peaks identified in S_2 are considered as candidates for the first peak of a fundamental period. The following nineteen measurements $\{M_i \mid i \in 1, \ldots, 19\}$ are made for pairs of candidates, $P_1, P_2 \in S_2$ (illustrated in Figure 3.3); P_i at time t_i with amplitude v_i:

- $M_1 = v_1$; the waveform amplitude at peak P_1.
- $M_2 = (v_2 - v_1)/v_1$; the relative change in amplitude between the candidate peaks.
- $M_3 = t_2 - t_1$; the duration between the candidate peaks.
- $M_4 = w_1$; the width at the zero crossing point of the peak P_1.
- Locate the largest peak $P_0 \in S_1$ within the region of duration $1.5(t_2 - t_1)$ before P_1. $M_5 = v_0/\min(v_1, v_2)$; the amplitude of peak P_0 relative to the minimum of the amplitudes of the candidate peaks.
- Locate the largest peak $P_3 \in S_1$ within the region of duration $1.5(t_2 - t_1)$ after P_2. $M_6 = v_3/\min(v_1, v_2)$.
- $M_7 = (v_3 - v_2)/v_2$.
- $M_8 = (t_3 - t_2)/(t_2 - t_1)$.
- Locate the largest peak $P_a \in S_1$ within the 3ms region before P_1. $M_9 = (v_1 - v_a)/v_1$.

[3] Speech signals are quantised by a 16-bit analogue-to-digital converter (ADC).

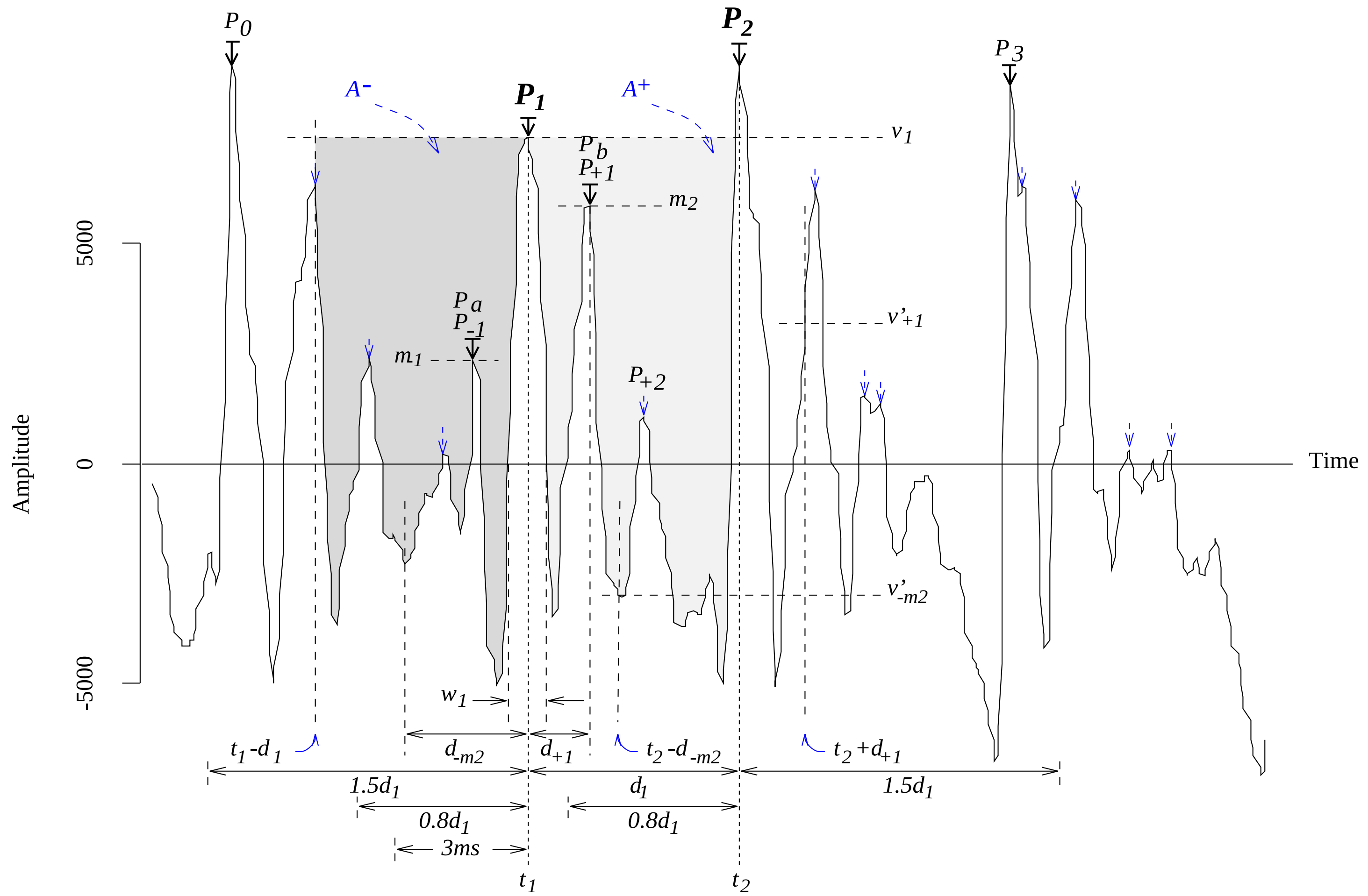

Figure 3.3: Feature-based $F\emptyset$ tracker (FBFT)

- Locate the largest peak $P_b \in S_1$ that lies between the two candidate peaks. $M_{10} = v_b / \min(v_1, v_2)$.
- Locate the first peak $P_{+1} \in S_1$ after P_1 and determine the waveform amplitude v'_{+1} at the same time after P_2. $M_{11} = (v'_{+1} - v_{+1})/v_1$.
- Locate the second peak $P_{+2} \in S_1$ after P_1 and determine the waveform amplitude v'_{+2} at the same time after P_2. $M_{12} = (v'_{+2} - v_{+2})/v_1$.
- Locate the first peak $P_{-1} \in S_1$ before P_1 and determine the waveform amplitude v'_{-1} at the same time before P_2. $M_{13} = (v'_{-1} - v_{-1})/v_1$.
- Locate the first local minimum after P_1 and determine the waveform amplitude v'_{+m1} at the same time after P_2. v_{+m1} is the amplitude of the first local minimum after P_1. $M_{14} = (v'_{+m1} - v_{+m1})/v_1$.
- Locate the first local minimum before P_1 and determine the waveform amplitude v'_{-m1} at the same time before P_2. v_{-m1} is the amplitude of the first local minimum before P_1. $M_{15} = (v'_{-m1} - v_{-m1})/v_1$.
- Locate the second local minimum before P_1 and determine the waveform amplitude v'_{-m2} at the same time before P_2. v_{-m2} is the amplitude of the second local minimum before P_1. $M_{16} = (v'_{-m2} - v_{-m2})/v_1$.
- Determine the maximum waveform amplitude, m_{-1} over the region t_1 to $t_1 - 0.8(t_2 - t_1)$. $M_{17} = (v_1 - m_{-1})/v_1$.
- Determine the maximum waveform amplitude, m_{-2} over the region t_2 to $t_2 - 0.8(t_2 - t_1)$. $M_{18} = (v_2 - m_{-2})/v_2$.
- The area A^+ is defined as the sum of differences in waveform amplitude and the amplitude of the peak P_1 over the region t_1 to $t_1 + (t_2 - t_1)$. The area A^- is defined as the sum of differences in waveform amplitude and the amplitude of P_1 over the region t_1 to $t_1 - (t_2 - t_1)$. $M_{19} = A^- / A^+$.

These measurements are combined in a statistical classifier that derives the location of glottal pulses in the waveform. The set of glottal pulse times generated by this method are considered in chronological order. The duration between adjacent glottal pulses is calculated and converted to Hertz. If the value lies within a limited range, it is taken to represent the fundamental frequency at the time located between the two glottal pulses; otherwise, the duration between the marks is considered to correspond to an unvoiced region of speech. The *FØ* values are limited to lie between 50Hz and 250Hz for male speech, and between 120Hz and 400Hz for female speech.

3.2.4 Parallel processing method (PP)

The parallel processing approach to *FØ* extraction proposed by Gold & Rabiner (1969) uses multiple peak picking in the time domain. An outline of this method is shown in Figure 3.4. The speech signal is initially low-pass filtered to simplify the temporal structure of the sampled waveform, as in the feature-based algorithm (FBFT). In order to make an unbiased comparison of the algorithms being reviewed here, it is necessary to set parameters which are common across algorithms to the same values. Thus, the signal is divided into a series of 38.4ms duration analysis frames with successive frames being separated by a 6.4ms frame shift, as with the other FDAs. Each analysis frame is processed by a 'silence' (low energy) detector. If two or more samples in a frame have magnitudes that exceed 120 ADC-units (approximately 50dB signal-to-noise level) then the frame is assumed not to be silence, and the frame is processed through six simple peak/valley detectors, each examining a different aspect of the waveform. The six detectors each generate a pulse train in which the magnitude of each pulse is governed by the particular aspect of the waveform being measured. The six measurements made by the detectors for the local maximum (peak) and the local minimum (valley) in the filtered waveform are:

- M_1, magnitude of each local maximum (peak).
- M_2, magnitude of each local minimum (valley).
- M_3, absolute difference in the amplitude of each peak and the amplitude of the previous valley.
- M_4, absolute difference in the amplitude of each valley and the amplitude of the previous peak.

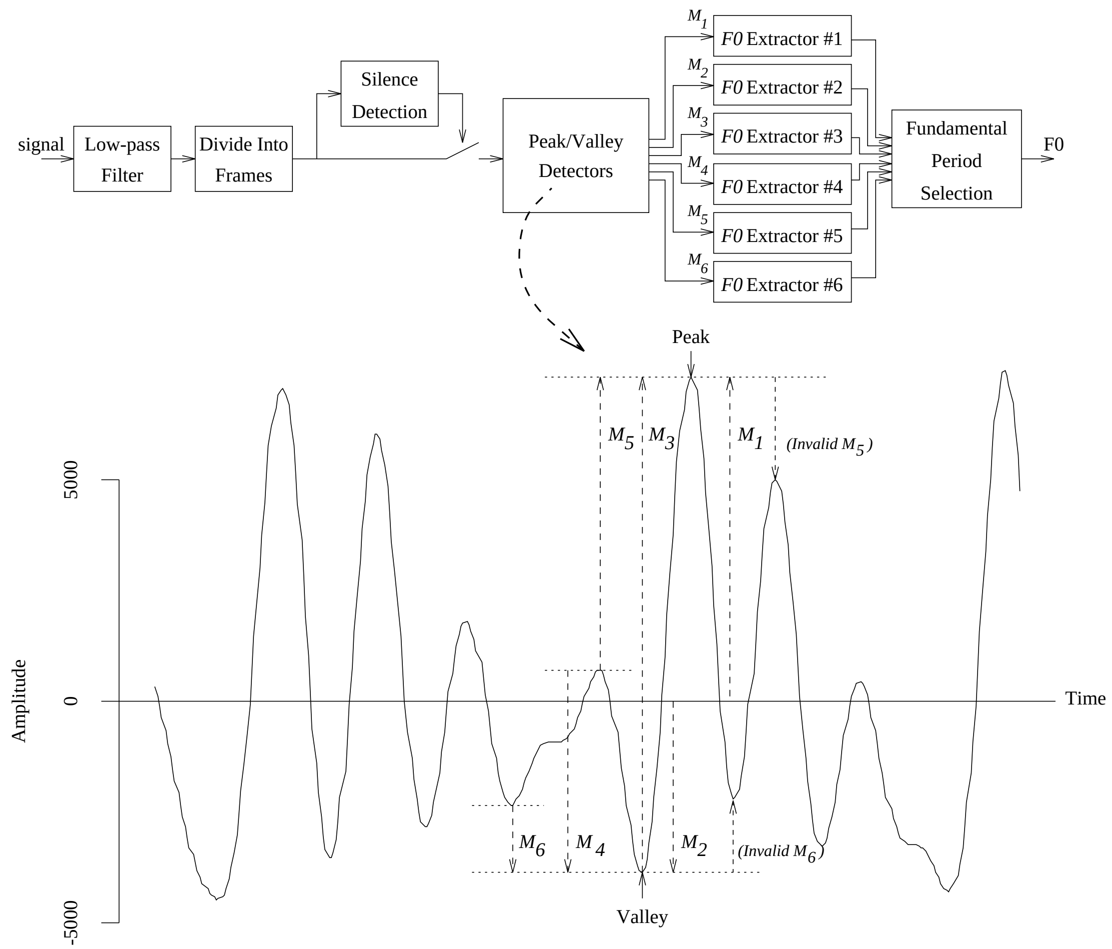

Figure 3.4: Parallel processing method (PP)

- M_5, difference in the amplitude of each peak and the amplitude of the previous peak (never permitted to be negative — ie. this measure is invalid when the peak amplitude is less than the amplitude of the previous peak).
- M_6, difference in the amplitude of each valley and the amplitude of the previous valley (never permitted to be negative – ie. this measure is invalid when the valley amplitude is greater than the amplitude of the previous valley).

The peak/valley detectors are followed by six individual *FØ* extractors operating in parallel. The *FØ* extractors work in the following way. After each detection of a pulse from the preceding peak/valley detector i, a threshold is set to the magnitude of the measurement M_i. The threshold remains constant for a blanking interval τ (during which time no subsequent pulse detection is allowed) followed by a single exponential decay (with time constant β). Whenever, the magnitude of an incoming pulse exceeds the level of the decay threshold value, it is detected and the process is repeated. The estimate of the fundamental period is the duration between each consecutive pair of detected pulses. The two parameters, τ and β, are made dependent upon a running average of the fundamental period. Finally, an overall estimate of fundamental period is computed by using a complex majority-wins type algorithm on the three most recent estimates of fundamental period from the six extractors (and estimates derived from these). A voiced/unvoiced decision is made on the basis of the degree of agreement between the six *FØ* extractors. If four or more agree, the frame being analysed is classified as voiced.

This algorithm is somewhat restricted by the fixed analysis frame length of 38.4ms in that at least two fundamental periods must reside within this duration for the *FØ* extractors to function adequately. For the frame length specified, the lowest fundamental frequency which the algorithm can be expected to extract reliably is approximately 52Hz.

3.2.5 Integrated *FØ* tracking algorithm (IFTA)

Secrest & Doddington (1983) propose that the linear prediction coding (LPC) (Markel & Gray, 1976) residual error signal mainly contains excitation information and therefore should theoretically provide the best signal from which to extract estimates of fundamental frequency. Inadequacies of the LPC model cause high frequency noise to be introduced into the residual signal, thus complicating the temporal structure of the signal and hindering *FØ* estimation. If the residual signal is low-pass filtered in an attempt to overcome this problem, then high frequency energy present in unvoiced regions of speech are also removed, rendering the signal suboptimal for making decisions as to whether or not the speech is voiced or unvoiced. Secrest & Doddington note that the first reflection coefficient of the LPC, K_1, is correlated with the high/low frequency energy in the signal and propose a single pole de-emphasis filter, $F(z) = 1/(1 - K_1 z^{-1})$, to low-pass filter the residual signal in voiced (low frequency energy) regions of speech and high-pass filter it in unvoiced (high frequency energy) regions.

Estimates of *FØ* are made from the de-emphasised residual error signal by quantifying the degree of similarity between two adjacent, non-overlapping sections. The signal is analysed frame-by-frame at intervals of 6.4ms (for compatibility with the other FDAs reviewed here). Each frame contains a set of samples, $s_N = \{s(i) \mid i \in 1, \ldots, N\}$, which is divided into two consecutive sections each containing a variable number of samples, n.

$$\begin{aligned} x_n &= \{x(i) = s(i) \mid i \in 1, \ldots, n\} \\ y_n &= \{y(i) = s(i+n) \mid i \in 1, \ldots, n\} \end{aligned} \tag{3.8}$$

Candidate values of the fundamental period are obtained by locating peaks in the normalised crosscorrelation coefficient given by,

$$r_{x,y}(n) = \frac{\sum_{j=1}^{n} x(j).y(j)}{\sqrt{\sum_{j=1}^{n} x(j)^2 . \sum_{j=1}^{n} y(j)^2}}, \qquad n \in N_{min}, \ldots, N_{max} \tag{3.9}$$

where N_{min} and N_{max} are the minimum and maximum expected fundamental period values (in number of samples) for a given speaker.

The amplitude of $r_{x,y}(n)$ quantifies the degree of similarity between x_n and y_n independently of variations in the signal energy. If x_n and y_n are in phase and have similar temporal structures, then $r_{x,y}(n)$ tends towards $+1$; and if x_n and y_n are out of phase or have different temporal structures, then $r_{x,y}(n)$ tends towards -1.

Secrest & Doddington propose an algorithm to determine the optimal fundamental period from the set of candidate values. The approach requires several frames of data to be analysed before a decision is made upon the optimal period and voicing classification of a frame. Every fundamental period candidate of a frame is compared with each candidate obtained for the previous frame. A penalty score is made for each comparison. The score penalises large deviations in fundamental period from one frame to the next, low values of $r_{x,y}(n)$ during voiced regions of speech and high values of $r_{x,y}(n)$ during unvoiced regions of speech, and uses a measure of the difference between the spectra of the two consecutive frames to penalise changes in the voicing classification when the spectrum is relatively unaltered from one frame to the next. For each candidate in a frame, there is a candidate in the previous frame for which this penalty score is minimal. A trajectory is formed by back-tracking from each candidate in a frame to the optimal candidate of the previous frame over a series of frames. The penalty score is cumulated at each step in the trajectory. The candidate selected as the optimal fundamental period for a frame is the one associated with the trajectory which has the minimum cumulative penalty score.

3.2.6 Super resolution *FØ* determinator (SRFD)

Medan *et al.* (1991) propose an algorithm which also uses a normalised crosscorrelation coefficient to quantify the degree of similarity between two adjacent, non-overlapping sections, as in the integrated *FØ* tracking algorithm (IFTA). The speech signal is initially low-pass filtered to simplify the temporal structure of the sampled waveform, as in the feature-based algorithm (FBFT) and the parallel processing method (PP).

Each frame of speech data is initially processed by a 'silence' (low energy) detector. The silence detector used in this algorithm differs somewhat from the silence detector used in the parallel processing method, which is not robust to d.c. offset. The minimum and maximum values of the sample sets $x_{N_{min}}$ and $y_{N_{min}}$ (see Equation 3.8) are determined. If the sum of their absolute values is less than some preset threshold (say, 240 ADC-units to be consistent with the 50dB signal-to-noise level used in PP) for either set of samples, then the current frame under analysis is classified as silence and no further analysis is done on the frame.

If the frame of data is not classified as silence, then candidate values for the fundamental period (in number of samples) are sought from values of n within the range N_{min} to N_{max} by using the normalised crosscorrelation coefficient $p_{x,y}(n)$ defined in Equation 3.10. This coefficient is determined for values of n in steps of a decimation factor L (a positive integer excluding zero).

$$p_{x,y}(n) = \frac{\sum_{j=1}^{\lfloor n/L \rfloor} x(jL).y(jL)}{\sqrt{\sum_{j=1}^{\lfloor n/L \rfloor} x(jL)^2 . \sum_{j=1}^{\lfloor n/L \rfloor} y(jL)^2}} \tag{3.10}$$

$$= r_{x,y}(n) \qquad \text{if } L = 1 \tag{3.11}$$

$$\{n = N_{min} + i.L \mid i \in 0, 1, \ldots; N_{min} \leq n \leq N_{max}\}$$

The decimation factor L is used to reduce the computational load of the algorithm. If L is set too low, then calculation of the normalised crosscorrelation coefficient will be computationally expensive and time consuming. If L is set too high, then candidate values for the fundamental period of the frame will be determined with less accuracy. As a compromise, L is set to four for this investigation (equivalent to down-sampling the speech waveform to 5kHz).

Values of $p_{x,y}(n)$ are only valid if there are four or more zero-crossings in the sample set $s_{2n} = \{s(i) \mid i \in 1, \ldots, 2n\}$. This ensures that at least two oscillations occur within the section of data s_{2n}. If there are less than two oscillations in the data, such as for low frequency nasal segments, then the value of $p_{x,y}(n)$ will be high for low values of n. This would erroneously promote fundamental period candidates which correspond to high frequencies.

Candidate values of the fundamental period are obtained by locating peaks in the normalised crosscorrelation coefficient for which the value of $p_{x,y}(n)$ exceeds a specified threshold, T_{srfd}. The threshold is adaptive and is dependent upon the voicing classification of the previous frame and three empirically determined values. If the previously analysed frame is classified as 'unvoiced' or as 'silence' (which is the initial state) then the threshold is set to 0.88. Otherwise, the previous frame must have been classified as 'voiced' and the threshold is set equal to 0.85 times the value of $p'_{x,y}(n'_0)$; where n'_0 is the fundamental period estimate (in number of samples) of the previous frame, and $p'_{x,y}(n)$ is the set of normalised crosscorrelation coefficients calculated for the previous frame. This product is not permitted to drop below 0.75.

$$T_{srfd} = \begin{cases} 0.88 & \text{if previous frame 'unvoiced' or 'silent'} \\ \max(0.75, 0.85 p'_{x,y}(n'_0)) & \text{if previous frame 'voiced'} \end{cases} \tag{3.12}$$

If no candidates for the fundamental period are found, the frame is classified as 'unvoiced'. Otherwise, the frame is assumed to contain 'voiced' speech. In order to find the optimal fundamental period from the set of candidate values, the candidates are listed in order of increasing fundamental period. The candidate at the end of this list represents a fundamental period of n_M, and the m'th candidate represents a period n_m. A second coefficient, $q(n_m)$, is calculated for each candidate. $q(n_m)$ is the normalised crosscorrelation coefficient between two sections of length n_M spaced n_m apart.

$$q(n_m) = \frac{\sum_{j=1}^{n_M} s(j).s(j+n_M+n_m)}{\sqrt{\sum_{j=1}^{n_M} s(j)^2 . \sum_{j=1}^{n_M} s(j+n_M+n_m)^2}} \tag{3.13}$$

The first coefficient $q(n_1)$ is then assumed to be the optimal value. If a subsequent $q(n_m)$ exceeds this optimal value when multiplied by 0.77 (an empirically determined value) then it is in turn assumed to be the optimal value. The candidate for which $q(n_m)$ is believed to be the optimal value forms the estimate for the fundamental period, n_0, of the frame being analysed.

Finally the algorithm obtains an estimate of the fundamental period with a fine resolution. A more accurate fundamental period estimate for the frame is determined by calculating $r_{x,y}(n)$ (Equation 3.9) for n in the region $n_0 - L$ to $n_0 + L$. The location of the maximum within this range corresponds to a more accurate value of the fundamental period. This final estimate is then refined to eliminate the effect of time quantisation errors, by using an interpolation method which is described in Medan *et al.* (1991).

3.2.7 FDA summary

Methods of determining the fundamental frequency of a speech waveform operate either on a frequency domain representation, a time domain representation or on a function of correlation. Time domain and correlation based algorithms make use of filtering techniques in an attempt to enhance performance, whereas the frequency based algorithms use the raw speech waveform. The cepstrum-based (CFD) and harmonic product spectrum (HPS) algorithms operate on a frequency domain representation and differ considerably in the way they estimate the fundamental frequency of a signal from its spectral characteristics. The feature-based (FBFT) and parallel processing (PP) methods operate on a time domain representation. These two methods take a number of measurements directly from the temporal structure of the sampled speech waveform. This has the advantage of being less computationally expensive than the frequency domain based algorithms. The integrated *FØ* tracking algorithm (IFTA) and the super resolution *FØ* determinator (SRFD) use a normalised crosscorrelation function. Although the underlying function is similar for these two algorithms, the way in which they determine fundamental period estimates from the correlation function differs considerably.

The most reliable and accurate method of determining the fundamental frequency of a speech waveform is sought in order to minimise the number of errors, which occur during *FØ* extraction, propagating into the prosodic analysis. A detailed performance evaluation of *FØ* determination algorithms is described in Section 7.2.

3.3 Syllabification

The concept of a syllable revolves round the perceptual prominence of speech sounds in sequence (ie. prominence at a segmental level). The number of syllables in a sequence of sounds is judged by a listener on the basis of the number of peaks of prominence perceived. In general, a peak of prominence is perceived where a speech sound carries greater sonority relative to its nearest neighbours. The production of vowels usually involves less constriction of the vocal tract than in the production of consonants. Vowels, therefore, carry greater sonority and constitute the nucleus of the syllables in which they occur. However, relatively prominent consonants can also take on the nucleus of a syllable containing no vowels. The behaviour of the speech sounds between syllable nuclei is of no consequence to the number of perceived prominences. The boundaries between syllables are therefore ill-defined.

Each syllable in an utterance holds the potential of being prominent at a suprasegmental level. Hence, the automatic identification of syllables in connected speech forms a part of prosodic analysis. Several schemes have been devised to partition the speech signal into syllable-sized units.

Mermelstein (1975) uses a *loudness function* as a measure of sonority across phones. The loudness function is a time-smoothed and frequency-weighted (-12dB/octave outside the range 500Hz–4kHz) summation of a speech signal's short-time power spectrum. A "convex-hull" algorithm (*ibid.*) is employed to locate dips in the loudness function, which form potential syllable unit boundaries. The algorithm is dependent upon a *significance* parameter which eliminates small dips (below 2dB) in the loudness function as potential boundaries, and a minimum syllable duration of 80ms is imposed. A 90.5% syllable detection rate is reported (6.9% syllables missed and 2.6% extra syllables).

Lea (1980) locates syllable nuclei by seeking dips (4 or 5dB) in a low-band energy contour (summation of energy in the frequency band 60Hz–3kHz). The energy dips are assumed to be associated with pre-vocalic and post-vocalic consonants. The high energy region between dips is considered to be a syllable nucleus if it consists of at least 30ms of voiced speech. The algorithm is reported to identify 90% of syllable nuclei (with 1% extra and 9% missing).

The automatic syllabification of isolated words is conducted by Aull & Zue (1985). Speech is initially segmented into broad phonetic classes — sonorants (vowels, nasals, liquids and glides), unvoiced obstruents, voiced obstruents, voice bars and silence. Segments classified as sonorant are further processed to locate intervocalic nasals, /l/ and /r/, and vowel-vowel transitions. The division of sonorant segments uses spectral weighting functions across selected frequencies which are aimed at capturing formant movements. All undivided sonorant segments and each vocalic region of the divided sonorant segments are taken as syllable nuclei. A performance evaluation of this syllabification scheme is not reported.

Waibel (1988) defines the syllable boundary, for his purposes, as the onset of the vocalic nucleus and proposes two algorithms to determine such boundaries. He reports a rule-based algorithm which yields a 90–96% syllable detection and a "Zapdash" based algorithm which achieves 90–93% syllable detection. Waibel states, however, that these are results of an informal experiment and gives no indication of the criteria used to syllabify the test data by hand.

The syllabification algorithms cited above are used in speech recognition applications to aid lexical access. Reliable information about the segmental content of an utterance is not available to the syllabification algorithms. In fact, additional information such as syllable structure is needed in speech recognition applications to aid lexical access because reliable information about the segmental content of an utterance is not available. In the application of prosodic analysis for computer aided pronunciation teaching, however, the orthographic transcript of an utterance is known because a foreign language learner is asked to read a given sentence in the course material. Reliable segmental information is therefore available. A syllabification algorithm is described in Section 8.1.1 which uses phone boundary and label information in conjunction with a low-band energy contour.

3.4 The composite structure of *FØ* contours

The raw fundamental frequency contour of a speech waveform does not form an acoustic-phonetic representation of the utterance intonation. It is assumed that a raw *FØ* contour constitutes four superimposed components.

- The intonation pattern of a sentence uttered by different speakers produces *FØ* contours with different fundamental frequency ranges despite the fact that the underlying intonation pattern is identical. Thus, an *FØ* trajectory in different fundamental frequency ranges from different speakers can correspond to the same phonological pitch accent. This speaker-dependent component of an *FØ* contour is a consequence of the anatomical and physiological differences of talkers.

- Macroprosodic variations in an *FØ* contour reflect a speaker's choice of intonation pattern for an utterance. The automatic prosodic analysis of speech aims to isolate this component in order to locate and identify pitch accents.

- Microprosodic variations are imposed on an $F\O$ contour by the segmental content of an utterance (Silverman, 1987). Silverman argues that microprosodic variations of fundamental frequency[4] occur during and in the immediate vicinity of phonetic segments, and that these variations may be greater than macroprosodic $F\O$ variations.

- An $F\O$ contour (produced by any fundamental frequency determination algorithm) can be expected to contain values which are inaccurate. An $F\O$ contour also consists of cycle-to-cycle jitter ($F\O$ perturbations) generated at the vocal folds (Hiller, 1985). Methods to reduce errors resulting from the automatic determination of fundamental frequency and the smoothing of $F\O$ perturbations are investigated in Section 7.3.

Two schemes to compensate for between-speaker differences in fundamental frequency range are reported as relatively successful in a review of normalisation methodologies by Rose (1987).

The first scheme, fraction of range normalisation, involves expressing an observed $F\O$ value as a fraction of the difference between range-defining $F\O$ values,

$$F\O_{norm} = \frac{F\O_{input} - F\O_{min}}{F\O_{max} - F\O_{min}} \tag{3.14}$$

This scheme has the disadvantage that calculating satisfactory range-defining parameters (that are assumed to be equivalent between speakers) is difficult without first completing the normalisation.

In the second scheme, z-transform normalisation, an observed $F\O$ value is expressed as a multiple of a measure of dispersion relative to the mean fundamental frequency. The normalised $F\O$ value is given by,

$$F\O_{norm} = \frac{F\O_{input} - \overline{F\O}}{\sigma_{F\O}} \tag{3.15}$$

[4] Silverman (1987) refers to microprosody as *segmental perturbations* in order to emphasis that this component of an $F\O$ contour can influence the perceived naturalness of synthesised speech. He argues that the term *microprosody* misleadingly implies that this component is smaller than suprasegmental (macroprosodic) variations. Whilst this thesis acknowledges the importance of microprosody in the perceived naturalness of synthesised speech, the term segmental perturbations is not used in order to avoid confusion with $F\O$ perturbations related to cycle-to-cycle jitter.

where $\overline{F\O}$ is the long term mean fundamental frequency for a given speaker, and $\sigma_{F\O}$ is the long term population standard deviation. Thus the normalised $F\O$ contour is centred around the zero with each unit representing a frequency which is one population standard deviation from the mean. This scheme has the advantage of reflecting a relatively stable statistical distribution from a large number of fundamental frequency values, rather than just the two range-defining parameters as is the case with fraction of range normalisation.

The *stylisation* of an $F\O$ contour aims to isolate the macroprosodic component by removing the microprosodic component for a given speaker, under the assumption that no $F\O$ errors exist because of malfunctions in an FDA. A requirement of the $F\O$ stylisation is to ensure that any microprosodic $F\O$ variations which are larger than macroprosodic $F\O$ variations are prevented from being confused as pitch accents, and that any macroprosodic $F\O$ variations which are smaller than microprosodic $F\O$ variations are not removed.

A number of algorithms have been proposed to automatically stylise an $F\O$ contour as a sequence of piece-wise straight lines (Scheffers, 1988), quadratic spline curves (Hirst & Espesser, 1993) or rise-fall-connection elements (Taylor & Isard, 1992; Taylor, 1993) where the rise elements and fall elements are modelled by a polynomial function, and the connection element is a straight line. $F\O$ contours which are stylised as a sequence of straight lines by hand are used in perceptual studies of the intonation of several European languages, including Dutch (Cohen & 't Hart, 1967; 't Hart & Cohen, 1973), British English (de Pijper, 1983) and French (Beaugendre *et al.*, 1992). It is argued in these studies that a sequence of straight lines produces a "close-copy" stylisation such that speech resynthesised using the stylised $F\O$ contours cannot be distinguished perceptually from speech resynthesised using the original $F\O$ contours. A comparative investigation into the use of straight line and parabola stylisation is reported by 't Hart (1991).

In the generation of an acoustic-phonetic representation of prosodic aspects of speech for computer aided pronunciation teaching, the stylisation of an $F\O$ contour aims to remove the microprosodic component of the contour. It is therefore immaterial as to whether or not the stylisation is based on straight lines or on polynominal functions. Linear piece-wise stylisation is employed in Section 8.2.1.

3.5 Automatic transcription of prosodic structure

Algorithms to extract prosodic features from the speech waveform have previously been developed for speech recognition applications. The motivation behind the development of such algorithms has been the notion of islands of reliability (Lea, 1980; Aull & Zue, 1985) where phonetic information is assumed to be more robust. The locations of such islands of reliability are provided by prominent syllables. Stress information is also required to aid lexical access. A review of the use of prosodic structure in automatic speech recognition systems is presented by Vaissière (1988).

The need to automatically extract prosodic features from a speech waveform is motivated by requirements other than speech recognition and computer aided pronunciation teaching applications. It is also potentially valuable for linguistic studies of corpora which are too large to analyse consistently by hand. The automatic transcription of prosodic aspects of speech, if designed on a cogent linguistic basis, is a useful tool for developing a better understanding of the relation between prosodic and syntactic structures and their phonetic realisation.

Lea (1974; 1980) describes an algorithm to annotate syllables as either prominent or non-prominent. The algorithm partitions an utterance into *syntactic constituents* on the basis of fluctuations in its *FØ* contour. Constituent boundaries are placed both at the minimum point in a valley of the *FØ* contour for which there is at least a 7% change in *FØ* at both sides, and at periods of unvoiced speech greater than 350ms. The *FØ* contour within each constituent is assumed to initially rise to some maximal point, then gradually fall. A straight *threshold* line (on a logarithmic scale) is taken from the initial peak in *FØ* to the end of the final fall in the constituent. Automatically located syllable nuclei (see Section 3.3) are annotated as prominent if they are associated with the initial *FØ* rise of a constituent, or if the local *FØ* rises above the threshold line. This algorithm is dependent on the performance of the *FØ* determination algorithm and is prone to error because of the unreliable automatically identified location of constituent boundaries. Lea's algorithm is reported to annotate 85% of syllables as prominent or non-prominent in agreement with human perception.

The algorithm proposed by Lea does not use duration, energy or vowel quality to locate prominent syllables. It terms of the definition of prominence outlined in Section 2.2.4, this algorithm locates pitch accents which are associated with prominent syllables, but it does not locate prominent syllables *per se*.

Aull & Zue (1985) propose an algorithm to determine lexical stress from the acoustic waveform for multi-syllabic isolated words. Syllables are annotated as prominent, non-prominent or reduced with a reported 87% accuracy. A vector of five normalised features is calculated for each syllable. The duration of the syllable nucleus (adjusted for pre-pausal lengthening according to phonetic context); the logarithmic average syllable energy in the frequency bands 400Hz–5.0kHz and 1.2–3.3kHz; the peak fundamental frequency; and a measure of spectral stability are used as features. Syllables bearing stress are believed to have spectrally stable nuclei (except diphthongs). A reference vector is formed from the maximum value of each vector across the syllables of a word. The syllable with the minimum Euclidean distance from its feature vector to the reference vector is designated as prominent. The distinction between non-prominent and reduced syllables is made using the duration and energy features alone. The Euclidean distance provides a measure of the *degree of prominence* of a syllable relative to other syllables in a word. This can be used to provide second candidates for the prominent syllables in multi-syllabic words (secondary stress).

The use of a Bayesian classifier (Appendix B) to assign a *probability of stressedness* to syllables in connected speech, rather than a binary decision (prominent or non-prominent) is described by Waibel (1988). In evaluating the classifier, however, any syllable with a higher probability of being prominent than non-prominent is assigned as prominent, otherwise it is assigned as non-prominent. A vector of acoustic features is used as the input to the classifier. There are a number of possible acoustic features which can be selected to form the input vector for each syllable.

- Integral of the peak-to-peak amplitude over the vocalic portion of the syllable (*ptpint*).
- Duration of the vocalic portion of the syllable (*sondur*).
- Duration of the entire syllable from the onset of the initial vowel to the beginning of the next syllable (*syldur1*).
- Duration from the end of the initial vowel of the previous syllable to the end of

the initial vowel in the syllable concerned (*syldur2*).

- Maximum *FØ* during the syllable (*F0max*).
- Average *FØ* during the syllable (*F0ave*).
- Offset in *FØ* from the syllable concerned to the following syllable (*F0offs*).
- Average spectral change (Aull & Zue, 1985) over the syllable nucleus (*spchave*).

Waibel uses a training database of connected speech consisting of syllables assigned as either prominent (primary and secondary stress) or non-prominent from their lexical assignment to investigate the correlation of the acoustic features with lexical stress. This approach differs from the need for prosodic analysis in computer aided pronunciation teaching proposed in Chapter 2 in that not all lexical stress placements appear as sentential stress in phrases of connected speech. Waibel clearly stipulates that the stress detection algorithm he proposes, "ignores sentential stress, emphasis, phrase level phenomena, and rhythmic/syntactic/semantic phenomena" (*ibid*, pp.108).

Bayesian classification is initially performed with an input vector containing just one of the acoustic features. Waibel's experimental results show that for this unidimensional classifier, *sondur* and *F0max* generate the least number of classification errors of all the duration and fundamental frequency features, respectively. However, in Waibel's experiments using a Bayesian classifier with a multi-feature vector, he does not report the combination of *ptpint*, *sondur*, *F0max* and *spchave* as a feature vector. A recalculation of Waibel's performance evaluation was conducted, as part of this thesis, to determine the performance of the Bayesian classifier under an open test (by excluding the database used in training the Bayesian classifier). The recalculation shows the Bayesian classifier, at best, to achieve 86.6% correct classification by using the features, *ptpint*, *syldur1*, *F0max* and *spchave*. The use of any additional features reduces the percentage of correct classification. The next best performance of 86.4% is achieved by using the features *ptpint*, *syldur1*, and *spchave*. This is contrary to Waibel's conclusion which is based on an evaluation which includes the training database (closed test). There is no evidence to support that the difference in performance of just five syllables (out of 2368) is statistically significant. Therefore, the introduction of the *FØ* feature has little effect on the classifier's performance. This is surprising given that so much literature (Section 3.1) has indicated that fundamental frequency and the placement of pitch accents are correlated with stress. It is possible that the features *F0max*, *F0ave* and *F0offs* do not capture the particular aspects of an *FØ* contour which are associated with pitch accents and subsequently with the location of prominent syllables.

The features that are selected to give the above 86.6% correct classification rate describe only the acoustic parameters for the syllable in question. The variation of these parameters relative to neighbouring syllables is not included as an input parameter to the classifier. This exclusion cannot be beneficial to the classifier, given that a syllable's prominence is relative to its neighbours by definition. (The prominence of syllable in isolation has no meaning.) In a system designed to extract prosodic information in French, Vaissière (1989) uses acoustic parameters which also describe parameter variations from one syllable to the next.

A move towards linguistic knowledge-based rules to locate prominent syllables is made by Hieronymus (1989; 1991). The acoustic parameters, duration, energy, and fundamental frequency are processed in parallel to determine their unidimensional contribution to the perceived stress of each vowel in an utterance. A measure of vowel quality or spectral stability is not used and a syllabification stage is not included. It is assumed that every vowel in an utterance forms the nucleus of a syllable and that no syllabic consonants form the nucleus of a syllable. The analysis system proposed by Hieronymus manipulates the duration, energy and *FØ* acoustic parameters in the following way.

Duration: The duration of the final vowel preceding either a section of speech labelled as a pause or the end of an utterance, is reduced by a fixed factor (0.6) to compensate for pre-pausal lengthening. Each vowel is classed on a broad phonetic basis as either a short vowel, a long vowel or a diphthong, and its duration is regarded as contributing to the prominence of vowel if it is greater than an *a priori* threshold duration respective of vowel-type (short, long or diphthong). A limit is imposed on the percentage of the vowels in the utterance which are marked as potentially prominent on account of duration.

Energy: The maximum low-band energy in each vowel is used as a measure of its intensity. The average of the two highest vowel intensities in the utterance is used as a reference. A vowel with an intensity within 7dB of the reference is regarded as having sufficient energy to contribute to its perceived prominence. Vowels with an intensity more than 20dB from the reference are taken to be definitely non-prominent, regardless of the other acoustic parameters.

Fundamental frequency: The *FØ* contour is divided into sections, one section for each vowel. The *FØ* section associated with a vowel V_0 is a region of continuously voiced speech which runs from the end of the first unvoiced consonant preceding the vowel to the beginning of the first unvoiced consonant succeeding the vowel. If another vowel V_{-1} is encounted between the preceding unvoiced consonant and vowel V_0, then the *FØ* section starts from the mid-point between the end of vowel V_{-1} and the beginning of vowel V_0. Similarly, if another vowel V_{+1} is encounted between the succeeding unvoiced consonant and vowel V_0, then the *FØ* section ends at the mid-point between the end of vowel V_0 and the beginning of vowel V_{+1}.

The peak *FØ* value and the *FØ* values at the beginning and at the end of each section are determined. The initial and final two *FØ* estimates are ignored during this process because they can be highly influenced by the potentially unstable behaviour of the vocal folds at the onset and offset of voicing. The overall slope is calculated from the locations and values of the *FØ* readings at the beginning and the end of each section. Each *FØ* section is categorised as a 'fall', 'rise', or 'level' on the basis of the gradient of this slope.

$$\text{Slope-type} = \begin{cases} \text{`fall'} & \text{if } gradient < -100\text{Hz/second} \\ \text{`rise'} & \text{if } gradient > 100\text{Hz/second} \\ \text{`level'} & \text{otherwise} \end{cases} \tag{3.16}$$

Each vowel is classified as either 'pitch accented' ("PA") or 'unaccented' ("UA") using the decision filter illustrated in Figure 3.5. The decision filter examines three consecutive *FØ* sections for each vowel — the *FØ* sections associated with the vowel and its left and right contexts. The first and last *FØ* section of an utterance do not have a left and a right context respectively. These end points are represented by 'NA' (not available). Figure 3.5 shows all three-way combinations of fall, rise, level and 'NA' which can arise and the corresponding output of the decision filter.

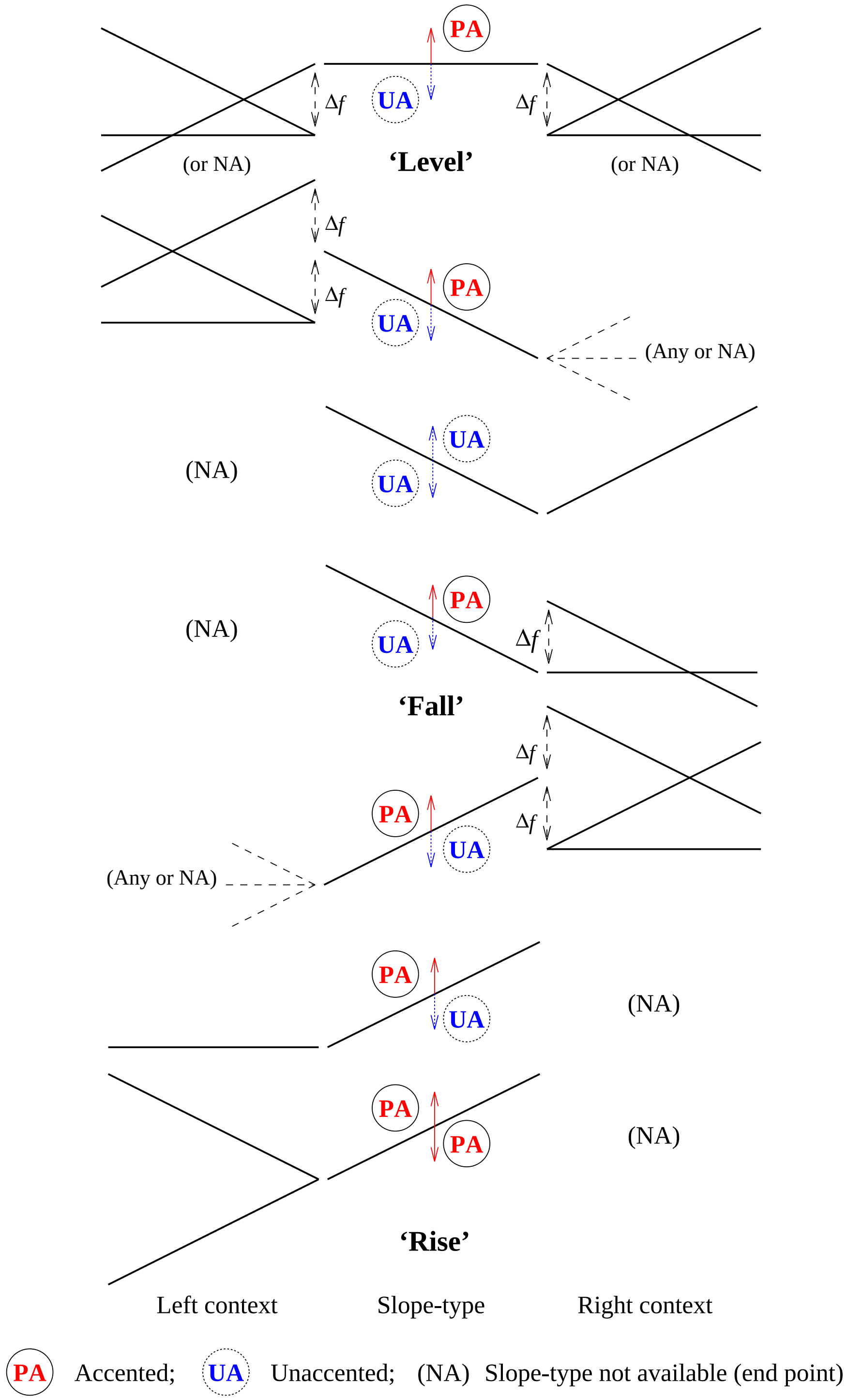

Figure 3.5: Pitch accent decision filter

A vowel associated with a level *FØ* section is classified as pitch accented if, for example, its left context is either falling or level and there is a step up of at least Δf (Hz)[5] from the end of the left context to the beginning of the level *FØ* section, and if the right context is a fall and the end of the level *FØ* section is not less than the beginning of the right context. If, however, the step up from the end of the falling or level left context to the beginning of the level *FØ* section is less than Δf (Hz) or if there is a step up from the end of the *FØ* section to the beginning of the right context, then the vowel is classified as unaccented.

Similarly, a vowel associated with a falling *FØ* section is classified as pitch accented if, for example, its left context is a rise and the beginning of the falling *FØ* section is no more than Δf (Hz) less than the end of the rising left context (the beginning of the falling *FØ* section can be any amount higher than the end of the rising left context). The right context is ignored in cases of a falling *FØ* section. If, however, the step down from the end of the rising left context to the beginning of the falling *FØ* section is more than Δf (Hz) then the vowel is classified as unaccented.

If a vowel is associated with a falling *FØ* section and the left context is not available (because the first vowel of an utterance is being classified) then the vowel is always classified as unaccented when it is followed by a rise.

The classification of a vowel as either pitch accented or unaccented can be determined from Figure 3.5 for any combination of fall, rise, level and 'NA' in the *FØ* sections associated with a vowel and its left and right contexts.

The pitch accent decision filter developed by Hieronymus has a number of advantages over the *FØ* features used in the Bayesian classifier technique. Firstly, judgements of vowel accentuation are based on trajectories of *FØ* in and around the vowel, rather than being based on static, isolated *FØ* measurements. Secondly, listeners classify syllables as accented if a pitch discontinuity is perceived. The pitch accent decision filter is designed to capture such discontinuities. This technique of locating pitch accents is therefore used in the system described in Section 8.2.2.

Prominence judgements for vowels are based on combining the unidimensional contribution to the perceived stress of each vowel obtained from the analyses of duration,

[5] In the implementation of the pitch accent decision filter, Δf is set to 9Hz.

energy and fundamental frequency outlined above. A vowel in an utterance is categorised as prominent if at least two-out-of-three of the acoustic parameters are indicative of stress in an equal-weight voting procedure. Only the energy parameter is given the power to exercise veto against a combined vote for prominence by the duration parameter and an *FØ* 'pitch accented' label. Hieronymus (1991) reports that this procedure agrees with 77.2% of human judgements of vowel prominence in connected speech.

Wightman & Ostendorf (1991; 1992) focus on the automatic recognition of prosodic boundaries and intonational features as well as determining the relative prominence of syllables. Linguistically motivated features are used to automatically determine the break index between adjacent words and to label each syllable as one of four categories — 'default', pitch accented syllable, phrase-final syllable with boundary tone, and boundary tone also perceived as pitch accented. The features used are:

Word dependent information: The lexical stress of a syllable; and a flag indicating the location of word-final syllables. The phonetic transcription of an utterance is determined by an automatic segmentation algorithm given an orthographic transcription of the sentence. This provides word boundary information which is combined with a lexicon to determine the lexical stress of syllables. This algorithm therefore uses the information which Waibel's algorithm tries to determine, as one of its input features.

Duration: The duration of a pause (if one exists) at the end of a word; the average z_{mean} duration measure of phones in the rhyme of the final syllable in a word; and the difference in the average z_{mean} duration measure for the syllable rhyme and the syllable onset (see Chapter 4 for details of syllable structure and z-score duration measures).

Energy: The mean energy in a syllable. The energy at the end of an utterance is often observed to reduce simply because the speaker is 'running out of breath'. It therefore has the potential to being used as a cue for the end of a phrase.

Fundamental frequency: The initial, final, maximum, minimum and average *FØ* values are determined for each syllable i. The following relative values are calculated:

$$F\O_{max}(i)/F\O_{average}(i+1)$$

$$F\O_{max}(i)/F\O_{max}(i-1)$$

$$F\O_{max}(i)/F\O_{average}(i)$$

$$F\O_{min}(i)/F\O_{average}(i)$$

$$F\O_{final}(i)/F\O_{average}(sentence)$$

The shape of the $F\O$ contour during the course of a syllable is described as either a rise, a fall, a rise-fall or a fall-rise by comparing the initial, average (mid) and final $F\O$ values.

These features form the input to a seven-state (one state for each level of break index) continuous Hidden Markov Model (HMM) and to a four state (one state for each syllable category) discrete HMM. The HMMs are fully connected in the absence of a theory either on the interaction between break indices or on the sequences of syllable categories. In evaluating this technique, Wightman & Ostendorf report that 77% of boundary tone are correctly detected with a 3% false detection rate, and that 86% of pitch accents are correctly detected with a 14% false detection rate.

3.6 Summary

The acoustic parameters which are identified by researchers as correlates of prosodic phenomena are duration, energy, fundamental frequency, and vowel quality. There is relatively little consideration of the effects of vowel quality on the perception of (lexical or sentential) stress compared with the extensive research relating duration, energy and $F\O$ with stress.

Studies of the acoustic correlates of stress are not consistent in the features they measure. 'Duration' is used by different researchers to refer to the duration of the vocalic portion of a syllable (Fry, 1955; Fry, 1958; Waibel, 1988) and to the duration of an entire syllable (Lieberman, 1960; Adams & Munro, 1978) with little consideration to the definition of a syllable (neither a phonetic definition nor a phonological definition). 'Energy' is used to refer to the peak amplitude in a syllable (Fry, 1955; Lieberman, 1960), to the amount of fall from a peak in the amplitude envelope of a syllable (Adams & Munro, 1978), to the integral of the peak-to-peak amplitude over the vocalic portion of a syllable (Waibel, 1988), to the average low-band syllable energy (Aull & Zue, 1985), and to the maximum low-band vowel energy (Hieronymus, 1989; Hieronymus & Williams, 1991). 'Fundamental frequency' is used as an acoustic correlate of stress. However, fundamental frequency is principally related to the characteristic melody of an utterance. Measures of *FØ* in a syllable are used as acoustic correlates of stress with little or no consideration of the intonational role of fundamental frequency. This approach is only viable in identifying the prominent syllables of isolated words. Fundamental frequency can be used as a secondary cue to the location of prominent syllables in connected speech because pitch accents are observed to fall on prominent syllables.

The fundamental frequency of a speech waveform must be determined as an initial process for prosodic analysis. The functionalities of a selection of *FØ* determination algorithms are described in detail. The performance of FDAs must be considered with respect to their application in systems of prosodic analysis and errors arising from the malfunctions of FDAs must be prevented from propagating into the subsequent prosodic analysis of speech.

Many of the duration and energy features are related to the definition of a syllable. Furthermore, prosodic aspects of speech are described in a syllabic domain. The automatic syllabification of speech is therefore an important part of prosodic analysis. Algorithms devised to partition a speech signal into syllable-sized units use measures of sonority based on low-band energy and use spectral characteristics to locate boundaries between adjacent sonorants. The reviewed algorithms are applied to speech recognition systems. Reliable information about the segmental content of an utterance is therefore not available to them. In the application of prosodic analysis for computer aided pronunciation teaching, however, reliable segmental information is available and may be used to enhance the syllabification process.

An emphasis is placed on the fact that the acoustic parameters which are associated with prosodic aspects of speech are also influenced by non-prosodic aspects of speech and that they interact in the manifestation of stress. It is highlighted that the *FØ* contour of an utterance is affected by the talker's anatomy and physiology (speaker-dependent

FØ range), the segmental content of an utterance (microprosodic variations), and cycle-to-cycle jitter (*FØ* perturbation) together with errors involved in its determination from the speech waveform.

The reviewed algorithms for prosodic analysis have not addressed the problems of determining the fundamental frequency of speech and have not comprehensively addressed the need to normalise for non-prosodic variations in the acoustic parameters which are used in the prosodic analysis. The system of prosodic analysis described in this thesis addresses the problems related to the determination of the fundamental frequency of speech and focuses on techniques of normalising for variations in acoustic parameters which are due to non-prosodic aspects of speech.