Chapter 6

Vowel quality measures

In addition to duration and energy features, it is suggested in Section 3.1 that perceived vowel quality is also correlated with sentential stress. A number of measures of vowel quality are proposed in this Chapter. Vowel quality measures are based on the assumption that the nuclei of prominent syllables are well articulated and are less affected by contextual assimilation, relative to the nuclei of non-prominent syllables. The aim here is to devise a measure of vowel quality based on formant frequency trajectories which is correlated with the prominence of a syllable.

The trajectories of formant frequencies through each vowel of an utterance are estimated using a formant tracking algorithm based on generalised centroids (Crowe & Jack, 1987; Crowe, 1988). The speech waveform, sampled at 20kHz, is Fourier-transformed every 5ms using a 512-point FFT with a 20ms Hann window. This ensures that formant frequency values are synchronised with the energy contour and *FØ* contour calculations. High-frequency pre-emphasis is applied to the resultant spectrum, and a triple centroid is estimated over the frequency range from 234Hz (the 6'th frequency bin) to 2891Hz (the 74'th frequency bin). The formants *F1*, *F2*, and *F3* are captured within this range while *FØ* and higher formants are excluded.

The formant contours generated by the generalised centroids algorithm are post-processed by a five frame median filter and three frame Hann window to remove isolated spurious formant frequency estimates (Section 7.3.1).

6.1 Vowel context and speaker normalisation

The perceptual quality of a vowel is correlated with the trajectories of the formant frequencies *F1*, *F2*, and *F3*. However, these trajectories vary depending on phonetic context and the individual speaker (Peterson & Barney, 1952). Given that there are physical constraints on the articulators, two different vowels pronounced in different phonetic contexts by the same speaker may have similar formant frequency values. Similarly, a vowel perceived as similar when spoken in different contexts by the same speaker may have different formant trajectories (phonetic context dependency). The acoustic realisation of a vowel described by the same phonetic symbol and spoken by two speakers may differ because of anatomical and physiological differences between the speakers. This may result in different vowels pronounced by different speakers having similar formant trajectories, and a vowel perceived as similar when spoken by two different speakers having different formant trajectories (speaker dependency).

A number of normalisation techniques are investigated here which aim to compensate for these context-dependent and speaker-dependent variations in the formant trajectories. In previous investigations of vowel normalisation (Disner, 1980; Syrdal & Gopal, 1986; Hillenbrand & Gayvert, 1993) the focus is placed on minimising the variance within a set of vowels spoken by different speakers which are presumed to correspond to the same vowel target, while maximising the separation of sets of vowels which are presumed to correspond to different vowel targets. Here, the focus of the vowel normalisation is to maximise, for each set of vowels presumed to correspond to the same vowel target, the separation of vowels which form the nuclei of prominent and non-prominent syllables.

The phonetic context dependency of the formant frequency trajectories is compensated for by obtaining one measurement from each trajectory within a vowel. Three different trajectory features $C_{1,2,3}$ are investigated.

- C_1, the peak/valley/mid-point value of a frame-level formant trajectory. The peak/valley/mid-point value of a formant trajectory over a vowel is determined using the method proposed in Section 5.1.
- C_2, the mid-region mean value of a formant trajectory. The mid-region mean value is defined as the average formant frequency value over the central half of a vowel.

Formant frequency values in the initial and final quarters of the vowel are assumed to be highly influenced by segmental context and are disregarded.

- C_3, the *stable*-region mean value of a formant trajectory. The stable-region of a vowel is defined after van Bergem (1988). The *F1*, *F2*, and *F3* contours are initially transformed to a Mel-scale using Equation 6.1.

$$F_{Mel} \; = \; 2595.0 \log\left(1.0 + \frac{F_{Hz}}{700.0}\right) \tag{6.1}$$

A window of quarter of the vowel length is moved through the *F1*, *F2*, and *F3* formant trajectories in search of a region of the vowel for which the pooled variance of these (Mel-scaled) formants is a minimum. The average formant frequency values (in Hertz) are determined within this region. If the duration of the vowel is so short that the window fails to occupy at least two frame-level formant frequency values (and hence one is unable to determine a variance) then a *break* value of infinity is associated with the vowel. Since the frame-level formant frequency values are specified at intervals of 5ms, vowels shorter than 40ms are associated with the break value.

Syrdal & Gopal (1986) propose that the speaker dependency of the formant frequency values can be normalised by use of formant differences on the critical-band-based Bark scale, $F2_{Bark} - F1_{Bark}$, $F3_{Bark} - F2_{Bark}$. The set of formant frequency measurements for each vowel are converted to a Bark-scale using Equation 6.2 (Zwicker & Terhardt, 1980).

$$F_{Bark} \; = \; 13.0 \arctan(0.76 F_{kHz}) + 3.5 \arctan\left(\frac{F_{kHz}}{7.5}\right)^2 \tag{6.2}$$

A problem in using formant differences, however, is that some pairs of phonetically distinct vowels have similar formant differences — for example, the pairs /ɑ, ɔ/ and /u, ʊ/ (Hillenbrand & Gayvert, 1993). Syrdal & Gopal address this problem by including the difference between *F1* and *FØ* ($F1_{Bark} - FØ_{Bark}$) as a parameter in the normalisation.

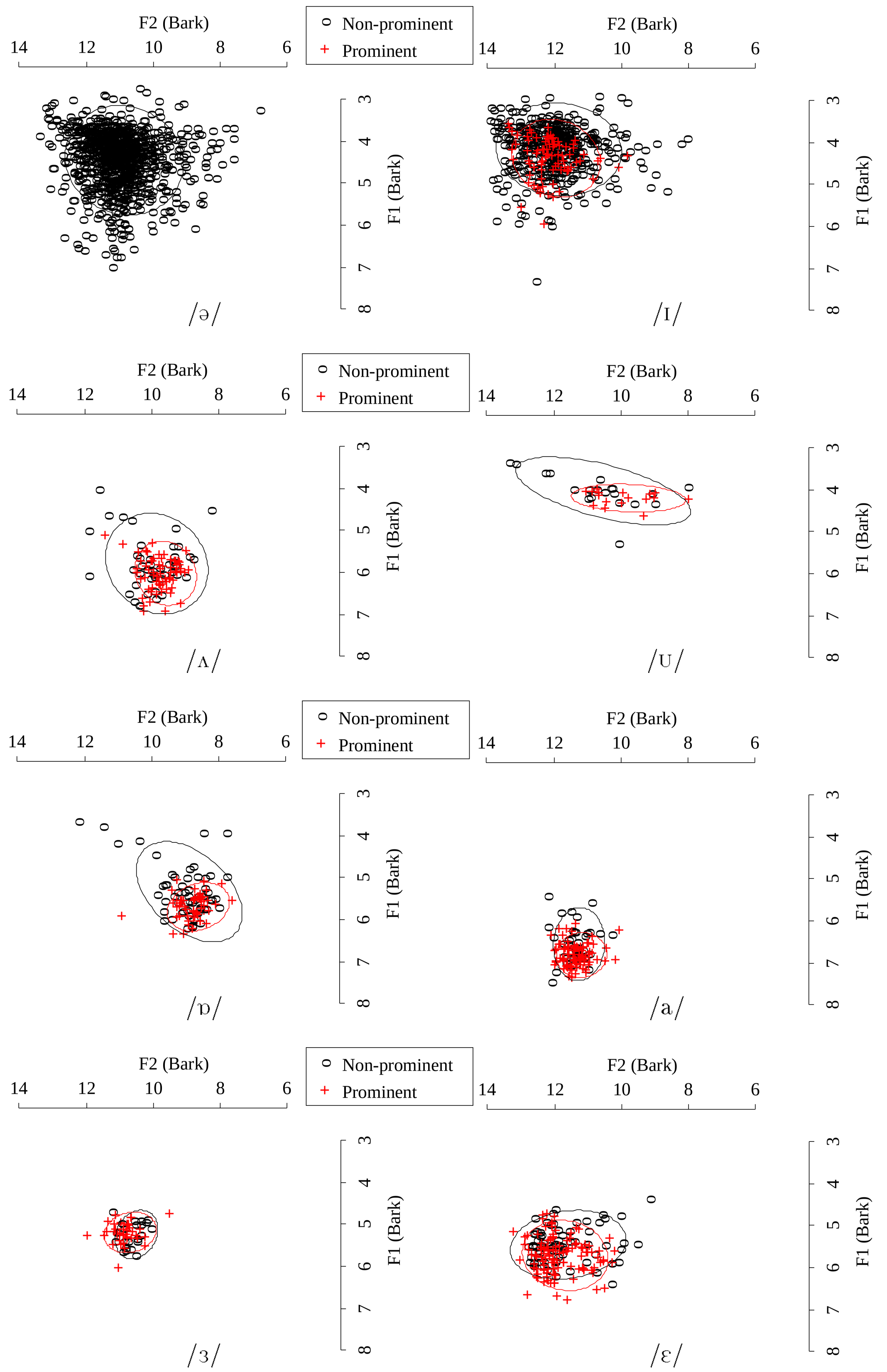
In summary, each vowel in an utterance is characterised by a set of measurements derived from *FØ* and formant frequency trajectories. One value is obtained from each frequency trajectory within a vowel by using one of three methods — $C_{1,2,3}$. For example, four values can be obtained from the *FØ*, *F1*, *F2*, and *F3* trajectories within a vowel by calculating the mid-region mean value C_2 for each trajectory. A vowel can be characterised by these values converted to a Bark-scale or by the difference between them.

6.2 Vowel target model

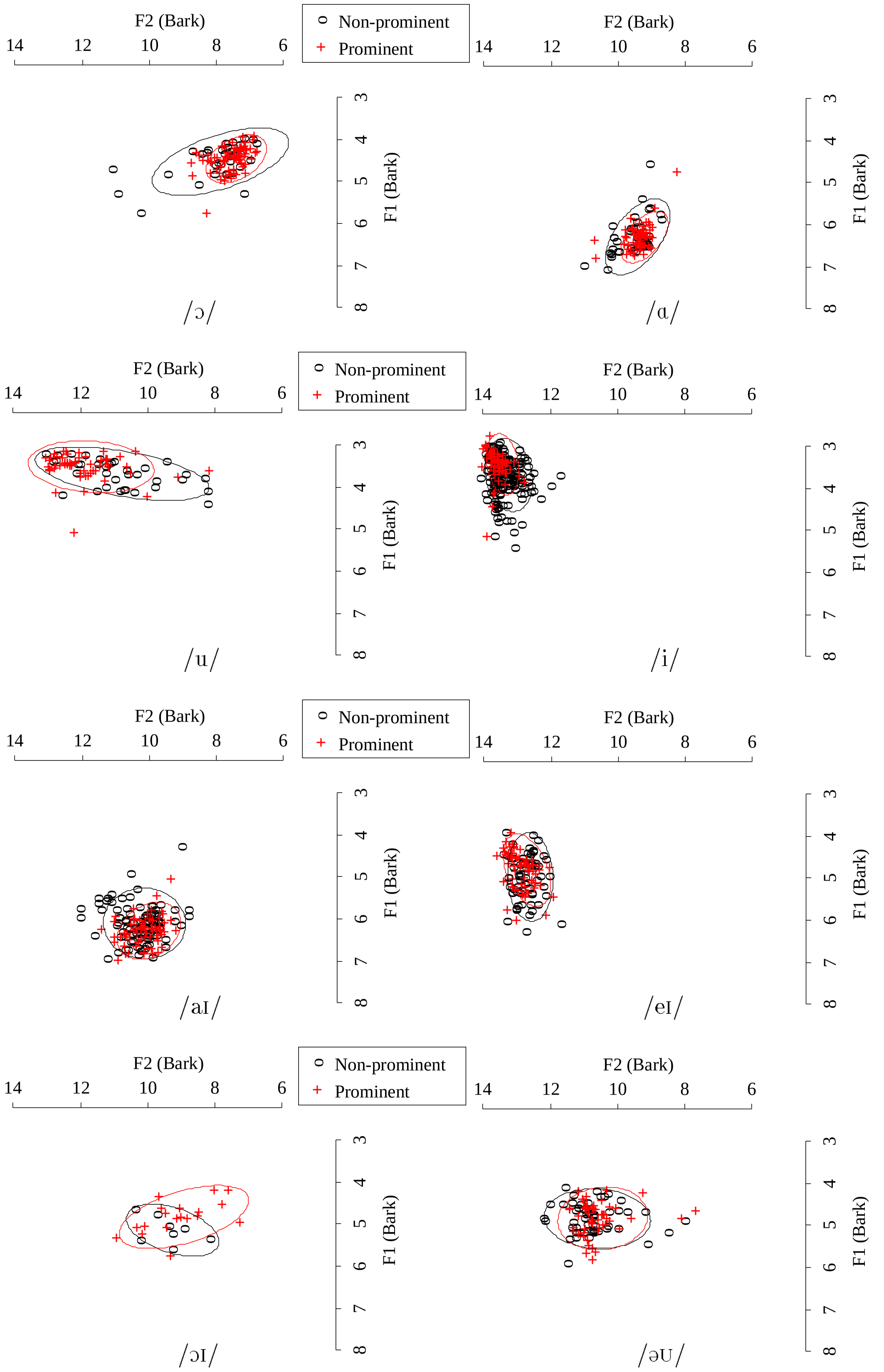
Figure 6.1 shows the clustering of prominent and non-prominent vowels for the twenty vowel-types in the training data of English speech. The ellipses indicate the degree of spread in the clusters of vowels in prominent syllables (shown in red) and in non-prominent syllables (shown in black). A 2-dimensional probability density function (p.d.f.) is associated with the variables plotted along the axes (in this case, the Bark-scaled mid-region mean values C_2 of *F1* and *F2*). The intersection of this p.d.f. (which can be imagined as a 'hill') and a plane at a level of two times the standard deviation (a slice through the hill) forms an ellipse when projected onto the zero-zero plane of the p.d.f. (the floor of the hill). It is these ellipses which are shown in Figure 6.1. There is a tendency for the vowels (at least the monophthongs) in prominent syllables to cluster more tightly than corresponding vowels in non-prominent syllables, and a tendency for the clusters of vowels in prominent syllables to lie within the clusters of the corresponding vowels in non-prominent syllables. This is not necessarily applicable to diphthongs because the inherent glide in their formant trajectories is not modelled by the single measurement associated with each formant trajectory.

The populations of formant frequency measurements approximate a Normal (Gauss-Laplacian) distribution in the Bark domain. The application of the Bayesian classifier to model each type of vowel using Bark-scaled frequency measurements is therefore viable[1]. A Bayesian model is created for each vowel-type. Each vowel is characterised by a number of features. The features are a combination of the absolute or the differences in Bark-scaled formant frequencies and fundamental frequency, where each of the formant frequencies and the *FØ* value associated with the vowel are obtained using any one of

[1]The requirement for the input data to be Normally distributed is generally not a prerequisite of a Bayesian classifier. This constraint is imposed here by the implementation of the classifier as described in Appendix B



continued on next page...



continued on next page...

... continued from previous page

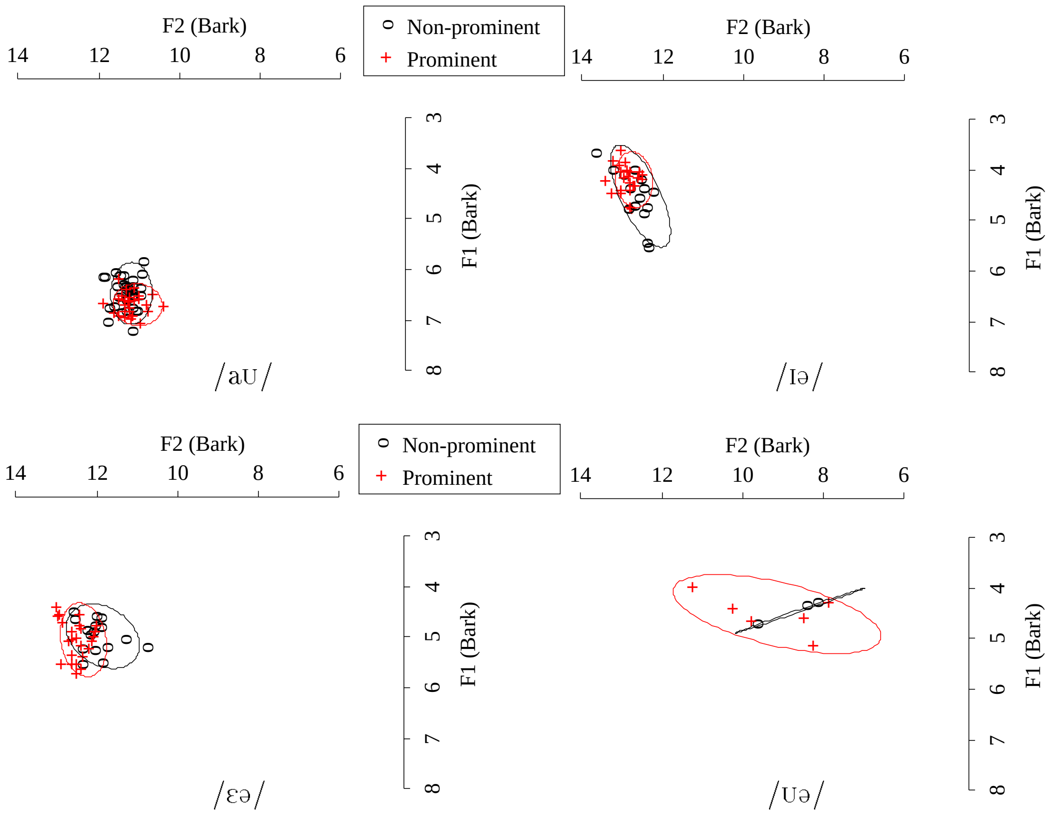


Figure 6.1: Vowel targets in prominent and non-prominent syllables

the trajectory features $C_{1,2,3}$. The vowel target models are calculated from frequency measurements of every vowel token (those corresponding to the nuclei of both prominent and non-prominent syllables) in the training data.

The quadratic discriminant score Q_{vowel} is determined for each vowel token in the training data with respect to its corresponding vowel target model. This score is a weighted variance-normalised distance measure between the vector of features which characterise the vowel and the centroid of the respective vowel target model. Thus, Q_{vowel} is a measure of how close the vowel is to its target.

It is assumed that vowels closer to their respective vowel target model correspond to the nuclei of prominent syllables. Under this assumption, Q_{vowel} for a vowel in a prominent syllable is less than Q_{vowel} for a vowel in a non-prominent syllable. Thus, in a distribution of Q_{vowel} for a given vowel-type containing p percent of vowels corresponding to non-prominent syllables, the lower $(100 - p)$ percent of Q_{vowel} values are assumed to correspond to vowels in prominent syllables. The distance of a vowel from its respective vowel target model is expressed relative to the $(100 - p)$'th-percentile of the quadratic discriminant scores for its vowel-type $\varphi_Q^{(100-p)}(vowel\_type)$.

$$Q' \;=\; Q_{vowel} - \varphi_Q^{(100-p)}(vowel\_type) \tag{6.3}$$

If the assumption is correct then Q' is positive for all vowels in non-prominent syllables and is negative for all others. The population of Q' does not have a Normal distribution. Since the implementation of the Bayesian classifier (described in Appendix B) assumes that the input has a Normal distribution, it is not used to classify the prominence of a syllable on the basis of Q'.

A simple peak-picking algorithm is used to label the contribution made by vowel quality to the perceived prominence of a syllable. A syllable is labelled as "*sq*" to indicated that it is prominent relative to its neighbours on the basis of vowel quality; that is, if Q' for the vocalic nucleus of the syllable is less than zero and if Q' for the nucleus of the syllable is less than Q' for the nuclei of both neighbouring syllables (end-points being inherently higher). Otherwise, a syllable is labelled as "*uq*", indicating that is not prominent relative to its neighbours on the basis of vowel quality. Schwa never forms the nucleus of a prominent syllable. Hence, Q' is positive for all syllables with a schwa nucleus, by definition. They are therefore labelled "*uq*". Q' is set to 'infinity' for any syllables which have a syllabic consonant as the nucleus. They are, therefore, also labelled "*uq*".

Every syllable in the test data is labelled using the above criteria. These labels are compared with prominence levels assigned to syllables by hand and the level of agreement between them is determined. Agreement levels are represented in Figure 6.2 for when the vowel target models are calculated from a number of different combinations of features. The highest level of agreement (71.6%) is obtained when the vowel target models are calculated from Bark-scaled peak/valley/mid-point (C_1) frequency differences $F1_{Bark} - F\O_{Bark}$, $F2_{Bark} - F1_{Bark}$, and $F3_{Bark} - F2_{Bark}$. A confusion matrix for the prominence labels assigned using this method versus prominence labels assigned by hand

		Vowel target label		
		sq	uq	total
Hand Label	P	963 (18.1%)	1135 (21.3%)	2098 (39.4%)
	NP	375 (7.1%)	2846 (53.5%)	3221 (60.6%)
	total	1338 (25.2%)	3981 (74.8%)	5319 (100.0%)

Correct classification rate = 3809/5319 (71.6%)

Table 6.1: Vowel quality: Confusion matrix for peak-picking algorithm (sq — syllable corresponding to a local peak in the vowel quality contour less than 0.0; uq — unstressed syllable on the basis of vowel quality; P — prominent syllable; NP — non-prominent syllable)

is shown in Table 6.1.

In order to place the agreement levels shown in Figure 6.2 into perspective, consider the performance of the peak-picking algorithm given the $z_{percentile}$ measure of a random variable with a Normal distribution. The thresholds for each vowel-type are given by the $(100 - p)$'th-percentile of a Normal distribution, where p represents the percentage of vowels which are non-prominent in the training data. Every syllable in the test data is associated with a random number from a single Normal distribution. A syllable is labelled as prominent if this random number is less than the threshold associated with the vowel-type of the nucleus of the syllable. Approximately $69 \pm 2\%$ of the syllable labels determined by this method are the same as those given by hand.

If the features used to characterise a vowel include an absolute measure of fundamental frequency (cf. $F1_{Bark} - F\O_{Bark}$) or are based on stable-region mean values (C_3) then the correlation of the vowel target distance Q' with sentential stress is worse than the correlation of a $z_{percentile}$-transformed, Normally-distributed random variable with stress.

A measure of the distance from a vowel target can, in some cases, be more effective at distinguishing between prominent and non-prominent syllables than a random variable in the analysis of the speech of a native speaker of English. However, such a measure cannot be used in the analysis of the speech of a non-native speaker because of vowel pronunciation errors.

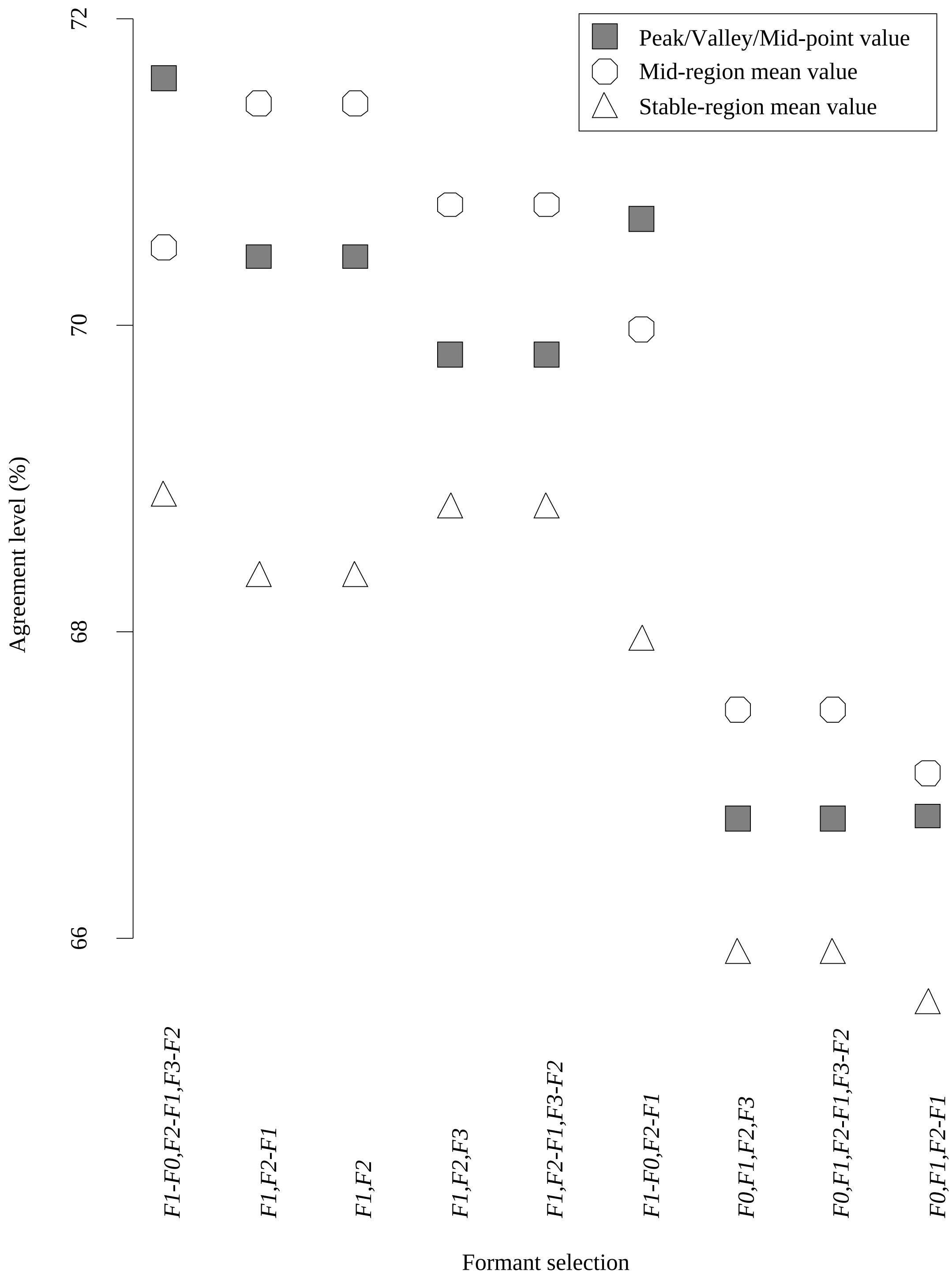


Figure 6.2: Vowel quality feature evaluation

There are two types of vowel pronunciation errors for non-native speakers of a language. Firstly, vowels which are assigned the same phonetic symbol in the target language to be acquired by a learner and in the native language of the learner, can differ in their phonetic realisation (Disner, 1980). Thus, vowel quality in the target language is influenced by the native language of the learner, as well as by context (both segmental context and syllable prominence). Secondly, vowels which do not belong to the vowel system of the learner's native language tend to be substituted by a vowel in the learner's native vowel system which is perceived (in the mind of the learner) to be 'close' to the desired vowel in the target language (di Benedetto *et al.*, 1992).

There is a fundamental problem in using a measure of the quality of a vowel spoken by a foreign language learner as a correlate of the perceived prominence of a syllable. A measure is required to determine if the quality of the vowel spoken by the learner is comparable to that of a native speaker when pronounced in the same context. This may be done by comparing the acoustic features which characterise the learner's vowel with the vowel system of a native speaker. The problem is that a mismatch between the vowel token pronounced by the learner and the intended vowel target for the target language can be due to a number of different reasons, other than the perceived prominence of a syllable. A mismatch can be due to anatomical and physiological difference between the learner and a native speaker, due to differences in the phonetic realisation of the vowel in the source and target languages, or due to a pronunciation error by the learner. This problem can be avoided only by comparing the acoustic features which characterise the learner's vowel with a vowel target modelled in the learner's vowel system (rather than in the vowel system of a native speaker of the target language). Therefore, if vowel quality is to be used in the automatic analysis of stress in utterances spoken by a non-native speaker, the task of learning correct vowel pronunciation and the task of learning the prosodic aspects of the target language must be kept separate.

The results shown in Figure 6.2 indicate that a measure of the distance of a vowel from its respective target is comparable to a $z_{percentile}$-transformed, Normally-distributed random variable when used as a correlate of the perceived prominence of a syllable. In the investigation, a vowel target is modelled from vowel tokens of a native speaker of English. The vowels used in the test data are spoken by the same speaker. In is inferred that even if the tasks of learning correct vowel pronunciation and learning the prosodic aspects of the target language are kept separate and a non-native speaker has successfully learnt to approximate the vowel system of the target language, then it is not possible to use the vowel-to-vowel-target distance as a correlate of stress.

6.3 Vowel stability measure

It is assumed here that a continuum exists in perceptual vowel quality from when a vowel is enunciated in isolation, to when a vowel is pronounced as the nucleus of a prominent syllable, to when a vowel is pronounced as the nucleus of a non-prominent syllable, to a schwa. Van Bergem (1994) proposes that schwa is, "a vowel without articulatory target that is competely assimilated with its [segmental] context" and that reductions in perceptual vowel quality (such as that which can occur in the nuclei of non-prominent syllables) is the partial assimilation of vowels with their segmental context. The formant trajectories of schwa are therefore completely dependent upon the position of the articulators during the production of surrounding phones, and vowels which have a perceptual reduction in quality are subject to some degree of contextual assimilation. The hypothesis that formant trajectories (or the spectra) are more changeable in the vocalic nuclei of non-prominent syllables (which are subject to vowel reduction) than in the vocalic nuclei of prominent syllables, is considered here.

Seven different measures of the degree of change in formant trajectories and one measure of the degree of spectral change through the course of a vowel, are investigated.

- V_1. The peak/valley/mid-point values in the *F1* and *F2* trajectories are determined using the method described in Section 5.1. An accumulative score is associated with each trajectory based on the type of point which is located — peak (+3), valley (+2), and rise or fall (+1). Thus, for example, if the trajectories of both *F1* and *F2* contain local peaks then a score of +6 is assigned to the syllable; and if the *F1* trajectory contains a local peak and the *F2* trajectory contains a fall then a score of +3 is assigned to the syllable.

- V_2. Maximum pooled variance of $F1_{Mel}$, $F2_{Mel}$, and $F3_{Mel}$ is calculated over the duration of the vowel using a window of quarter of its length — an 'instability' measure. This measure approximates a log-Normal distribution.

- V_3. Minimum pooled variance of $F1_{Mel}$, $F2_{Mel}$, and $F3_{Mel}$ is calculated over the duration of the vowel using a window of quarter of its length — a 'stability' measure. This measure approximates a log-Normal distribution.

- $V_{4,5,6,7}$. Maximum V_4, minimum V_5, mean V_6 or standard deviation V_7 (four different measures) is determined for the rate of change of $F1$, $F2$, and $F3$ from one frame to the next, across the entire duration of the vowel.

- V_8. The average Euclidean distance of frame-to-frame log-power spectra is calculated over the central 50%[2] of syllable nuclei — a measure of 'spectral change'. This measure is determined by performing a Fourier transform to frames (20ms in duration) of speech data at regular intervals (every 5ms) in an utterance and calculating a log-power spectrum for each frame. The Euclidean distance (Equation 6.4) is evaluated for adjacent spectra. This distance is averaged over the central 50% of each syllable nucleus in the utterance.

$$Euclidean_{ij} = \sqrt{\sum_{n=1}^{N}(x_{in} - x_{jn})^2} \tag{6.4}$$

where N is the number of frequency bins in a spectrum and x_{in} is the log-power in the n'th frequency bin of the spectrum of the i'th frame ($j = i + 1$).

A syllable is labelled as prominent on the basis of vowel stability if V_1 equals +6; otherwise it is labelled as non-prominent. Each of the measures $V_{2,\ldots,8}$ is used as a unidimensional feature for the Bayesian classification of syllable prominence. None of the measures succeed in classifying syllable prominence better than a $z_{percentile}$-transformed, Normally-distributed random variable. The use of any measure results in more prominent syllables being labelled automatically as non-prominent than being correctly labelled as prominent. The best performance (68.8% of syllables in agreement with the prominent labels transcribed by hand in the test data) is yielded by the measurement V_8.

[2]The central 50% of a syllable refers to the mid-portion of the syllable, thus excluding 25% at the beginning of the syllable and 25% at the end.

6.4 Phonological rules

According to the phonological description of English R.P., none of the lax vowels /ɪ, ɛ, ʌ, ʊ, ɒ, (æ)/ can appear in prominent open syllables (Ladefoged, 1982). It may seem reasonable to use a rule that labels all lax vowels that occur in open syllables as non-prominent. There are, however, two disadvantages to the use of such a rule. Firstly, the rule is dependent upon the definition of a syllable. An open syllable defined phonologically using the principle of maximal onset (Pulgram, 1970) will differ from one defined using the principle of maximal coda or one defined using acoustic-phonetic criteria (Section 8.1.1). Secondly, foreign speakers may stress any of the vowels /ɪ, ɛ, ʌ, ʊ, ɒ, (æ)/ in open syllables when speaking English if such syllable structures are permissible in their mother-tongue. In the training data, spoken by a native speaker of British English, the only vowels which form the nucleus of prominent open syllables in word final position are long, tense vowels and diphthongs /ɑ, i, ɔ, u, aɪ, aɪ, eɪ, ɔɪ, əʊ, ɛə, ɪə, ʊə/. This does not mean that every speaker of English (including learners of English with strong foreign accents) only place stress on word-final open syllables containing this set of phones. Recall that the aim here is to identify the location of prominent syllables so that a foreign speaker can be informed about incorrect stress placement in English, such as placing stress on lax vowels in word-final open syllables. It is therefore inappropriate to assign all word-final open syllables with lax vowel nuclei as non-prominent by default.

6.5 Summary

The vowel quality of a syllable nucleus is assumed to be an acoustic correlate of stress and a number of measures of vowel quality based on formant trajectories are proposed in this Chapter.

A normalised quadratic discriminant score (Equation 6.3) is proposed as a measure of how close a vowel is to its target, where a vowel target is a statistically trained model based on frequency measurements which aim to be normalised for phonetic context and speaker dependencies. This measure is used to label syllables spoken by a native English speaker as either prominent or non-prominent and yields a 71.6% agreement level with prominence labels based on human perception. It is argued that this measure cannot be used as a correlate of stress in utterances spoken by non-native speakers of English because of vowel pronunciation errors.

A number of different measures of the degree of change in formant trajectories (and spectral change) due to contextual assimilation are investigated. The average Euclidean distance of frame-to-frame log-power spectra (Equation 6.4) is used as a measure of the degree of spectral change in a vowel which is due to contextual assimilation. This measure yields a 68.8% level of agreement between prominence labels assigned automatically and as transcribed by hand.

The underlying assumptions made are that, in connected speech, the vocalic nuclei of prominent syllables are closer to their respective vowel targets than the vocalic nuclei of non-prominent syllables, and that formant trajectories are more changeable in reduced vowels than fully articulated vowels because they are subject to greater degrees of contextual assimilation. The proposed measures of vowel quality fail to convincingly support these assumptions since their use to distinguish prominent syllables from non-prominent syllables is comparable with using of a $z_{percentile}$-transformed, Normally-distributed random variable as an input parameter to a peak-picking algorithm. The application of the proposed vowel quality measures as correlates of stress is therefore inappropriate. The use of phonological rules to transcribe syllables as non-prominent under certain conditions, is also dismissed.

The formant trajectories of vowels are known to vary within the vowel system of a speaker. The cause of these variations are not yet fully understood although it has been suggested that segmental context and stress play a role. Duration and articulatory constraints may, however, be the means whereby segmental context and stress influence formant trajectories (Lindblom, 1963). Whether or not stress affects the perception of vowel quality directly, the prominence of a syllable cannot be determined from a measure of vowel quality, in the same way as it is not possible to determine the segmental context of a vowel from its perceived quality, without first understanding the exact nature of the influence of a vowel's context on its formant trajectories in connected speech. It is beyond the scope of this thesis to formulate such an understanding. Research to systematically measure coarticulatory effects on the formant frequency trajectories of the schwa in isolated nonsense words is reported by van Bergem (1994).

To conclude, Chapters 4 & 5 and this Chapter have concentrated on optimising the extraction of acoustic correlates required for the prosodic analysis of English. The underlying principle is to normalise the acoustic parameters for variations due to non-prosodic aspects of speech. The duration, energy and vowel quality features investigated are relevant to spoken English. Thus, the acoustic features presented might not be applicable to other languages. For example, the duration feature $D_{feature}$ (Equation 4.6) might not be related to the syllable lhyme for languages which differ from English in their syllabic structure, rhythm and timing. The methodology outlined in these Chapters can, however, be applied to investigations of the acoustic correlates of prosodic aspects of languages other than English.

Chapter 7

Fundamental frequency extraction

Methods to automatically determine the fundamental frequency of a speech waveform are addressed in this Chapter. 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 occurring during *FØ* extraction from propagating into the prosodic analysis.

A preliminary evaluation of the FDAs reviewed in Section 3.2 provided evidence (which is supported by the evaluation reported in Section 7.2) that the super resolution *FØ* determinator (SRFD) and the feature-based *FØ* tracker (FBFT) perform most reliably and consistently for both male and female speech. A number of modifications to the SRFD algorithm (summarised in Section 3.2.6) are proposed in Section 7.1 in order to reduce the occurrence of errors such that it is optimised for prosodic analysis.

The performances of the *enhanced* SRFD algorithm and the *FØ* determination algorithms summarised in Section 3.2 are evaluated in Section 7.2. The evaluation uses laryngograph data to generate a reference contour and is tailored to identify the most appropriate method of fundamental frequency extraction of adult male and female speech recorded with little ambient interference.

A de-step filter is proposed in Section 7.3.2 to post-process the *FØ* contour generated by an FDA in a way which further reduces the occurrence of errors.

7.1 Enhanced super resolution *FØ* determinator (*e*SRFD)

The speech waveform is initially low-pass filtered to simplify the temporal structure of the sampled waveform. Each frame of filtered sample data is then processed by the 'silence' (low energy) detector described in Section 3.2.6.

The definitions of the segments x_n and y_n (Equation 3.8) are refined in order to improve the synchronisation of the *FØ* contour with the speech data. In the unmodified implementation of SRFD, the r'th frame of an utterance starts at $t_{interval}(r-1)$ milliseconds into the utterance, where $t_{interval}$ is the interval between successive analysis frames ($t_{interval} = 6.4$ms for compatibility with the other FDAs during the evaluation and $t_{interval} = 5.0$ms to synchronise *FØ* values with the energy contour during prosodic analysis). The analysis of each frame yields an estimated value of the fundamental period as n_0. In other words, the fundamental period for the two cycles following the start of the frame is n_0 samples. Thus, the *FØ* value produced for a frame describes the data centred $t_{interval}(r-1) + n_0/F_s$ milliseconds into the utterance (where F_s is the sampling frequency in kHz). This is dependent upon the estimated fundamental period of the analysis frame. An asynchronous *FØ* contour is therefore generated. To rectify this problem, the first sample of the segment y_n is defined as the sample $s(1)$, $t_{interval}(r-1)$ milliseconds into the utterance for the r'th frame. Each frame of speech data contains a set of samples $s_N = \{s(i) \mid i \in -N_{max}, \ldots, N - N_{max}\}$ which is divided into three consecutive segments each containing a variable number of samples, n, as illustrated in Figure 7.1.

$$\begin{aligned} x_n &= \{x(i) = s(i-n) \mid i \in 1, \ldots, n\} \\ y_n &= \{y(i) = s(i) \mid i \in 1, \ldots, n\} \\ z_n &= \{z(i) = s(i+n) \mid i \in 1, \ldots, n\} \end{aligned} \tag{7.1}$$

Frames at the beginning of an utterance, for which x_n is not fully defined, are classified as 'silent'; likewise frames at the end of an utterance, for which y_n and z_n are not fully defined. The use of the third segment of speech, z_n, is described later in this Section.

The analysis of the segments x_n and y_n produces an estimate of the *FØ* value for a frame of speech data centred $t_{interval}(r-1)$ milliseconds into the utterance. This timing

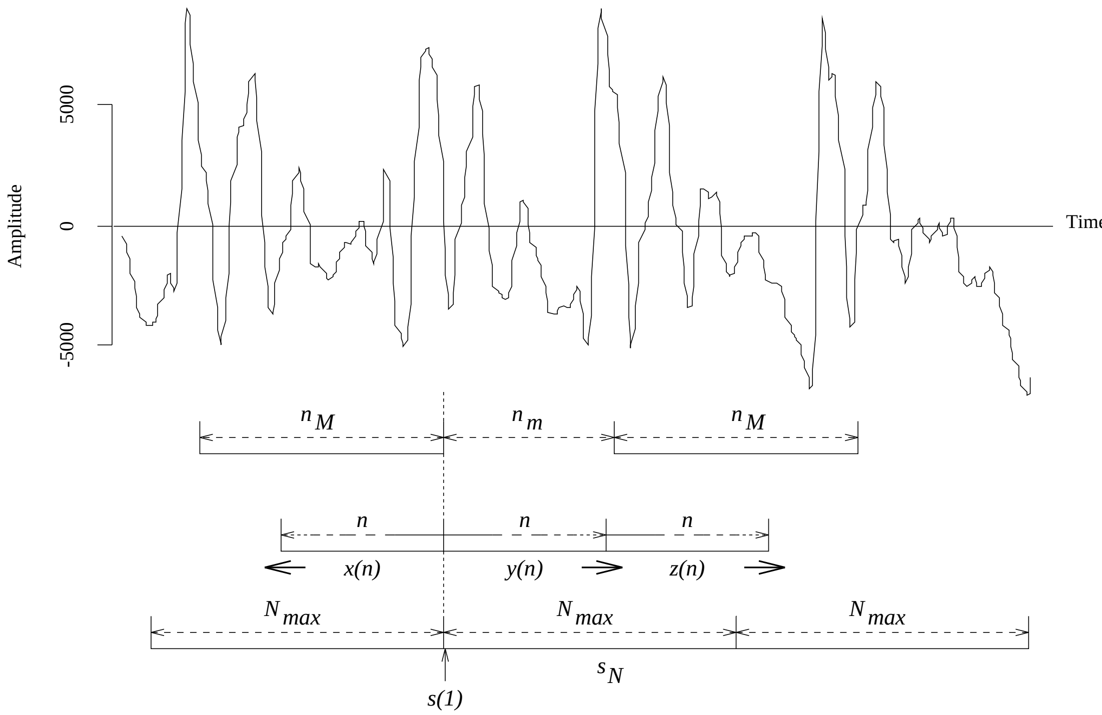


Figure 7.1: Analysis segments for the enhanced super resolution *FØ* determinator (*e*SRFD)

is independent of the estimated fundamental period of the analysis frame.

An *FØ* contour produced by any *FØ* determination algorithm can be expected to contain values which are inaccurate, such as instances of *FØ* doubling and halving errors and instances when voiced sections of speech are classified as unvoiced by an FDA or when unvoiced sections of speech are classified as voiced. *FØ* doubling errors occur when the estimated fundamental frequency is an overtone of the true fundamental frequency. *FØ* halving errors occur when the *FØ* determination algorithm erroneously mistakes the correctly estimated fundamental frequency as an overtone of the true fundamental frequency and then effectively over-compensates for this by dividing by some multiple of two. The remaining modifications made to the SRFD algorithm are designed to reduce the occurrences of *FØ* doubling and halving errors and to improve the voicing classification of frames of speech.

A value of n is sought for each analysis frame such that each segment defined in

Equation 7.1 occupies a single fundamental cycle. Candidate values of n are sought within the range N_{min} to N_{max} by using the normalised crosscorrelation coefficient $p_{x,y}$ defined in Equation 3.10. The locations of local maxima in $p_{x,y}$ with values above the adaptive threshold T_{srfd} (Equation 3.12) form candidate values for the fundamental period.

If no candidates for the fundamental period are found, the frame is classified as 'unvoiced'. Otherwise, the frame is classified as 'voiced' and a second normalised cross-correlation coefficient $p_{y,z}(n)$ is determined for all the fundamental period candidates.

$$p_{y,z}(n) \;=\; \frac{\sum_{j=1}^{\lfloor n/L \rfloor} y(jL).z(jL)}{\sqrt{\sum_{j=1}^{\lfloor n/L \rfloor} y(jL)^2 . \sum_{j=1}^{\lfloor n/L \rfloor} z(jL)^2}} \tag{7.2}$$

$$\{n \mid p_{x,y}(n) > T_{srfd}\}$$

Those candidates for which $p_{y,z}(n)$ also exceeds the threshold T_{srfd} are given a score of 2, while the others are given a score of only 1. Candidates with a higher score are more likely to represent the true fundamental period. If there are one or more candidates with a score of 2, then all those with a score of only 1 are removed from the list of candidates and ignored. Following this, if there is only one candidate (with a score of either 1 or 2) then the candidate is assumed to be the best (and only) estimate of the fundamental period for that frame. Otherwise, an optimal fundamental period is sought from the set of remaining candidates. 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 of n_m. The coefficient $q(n_m)$ (Equation 7.3) is calculated for each candidate. $q(n_m)$ is the normalised crosscorrelation coefficient between two sections of length n_M spaced n_m apart, as illustrated in Figure 7.1.

$$q(n_m) \;=\; \frac{\sum_{j=1}^{n_M} s(j-n_M).s(j+n_m)}{\sqrt{\sum_{j=1}^{n_M} s(j-n_M)^2 . \sum_{j=1}^{n_M} s(j+n_m)^2}} \tag{7.3}$$

As in the unmodified version of SRFD, 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 *a priori* 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.

In the case when there is only one fundamental period candidate with a score of 1 and no candidates with a score of 2, there is only a small probability that the candidate correctly represents the true fundamental period of the frame. If, in such cases, the previous frame was classified as either 'silent' or 'unvoiced', then the *FØ* value describing the current 'voiced' frame is held until the state of the subsequent frame is known. If this next frame is also not classified as 'voiced', then the frame whose *FØ* value is on hold is an isolated frame which is highly unlikely to be voiced. It is therefore re-classified as 'unvoiced'. Otherwise, the held *FØ* value is assumed to be a sufficiently good estimate of the fundamental frequency for that frame.

The above modification dramatically reduces the occurrences of *FØ* doubling and halving errors in the resultant *FØ* contour. However, its implementation also causes a greater percentage of voiced regions of speech to be erroneously classified as unvoiced. In order to counteract this undesirable effect, an additional modification applies biasing to the coefficients $p_{x,y}(n)$ and $p_{y,z}(n)$ for values of n where the fundamental period of a new frame is expected to lie. Biasing is applied if the following conditions are satisfied:

- The two previously analysed frames were classified as 'voiced'.
- The *FØ* value of the previous frame is not being temporarily held.
- The fundamental frequency of the previous frame f_0' is less than $\frac{7}{4}$ times the fundamental frequency of its preceding voiced frame f_0'', and greater than $\frac{5}{8}f_0''$, ie. it is highly probable that the fundamental period estimate of the previous frame is not an *FØ* doubling or halving error.

The fundamental period of the new frame n_0 is expected to lie within the range of n closest to n_0' for which the set of $p_{x,y}(n)$ from the previous frame are greater than zero (Figure 7.2). The normalised crosscorrelation coefficients $p_{x,y}(n)$ and $p_{y,z}(n)$ are doubled for values of n in this range. This effectively applies a bias on the location of a maxima in

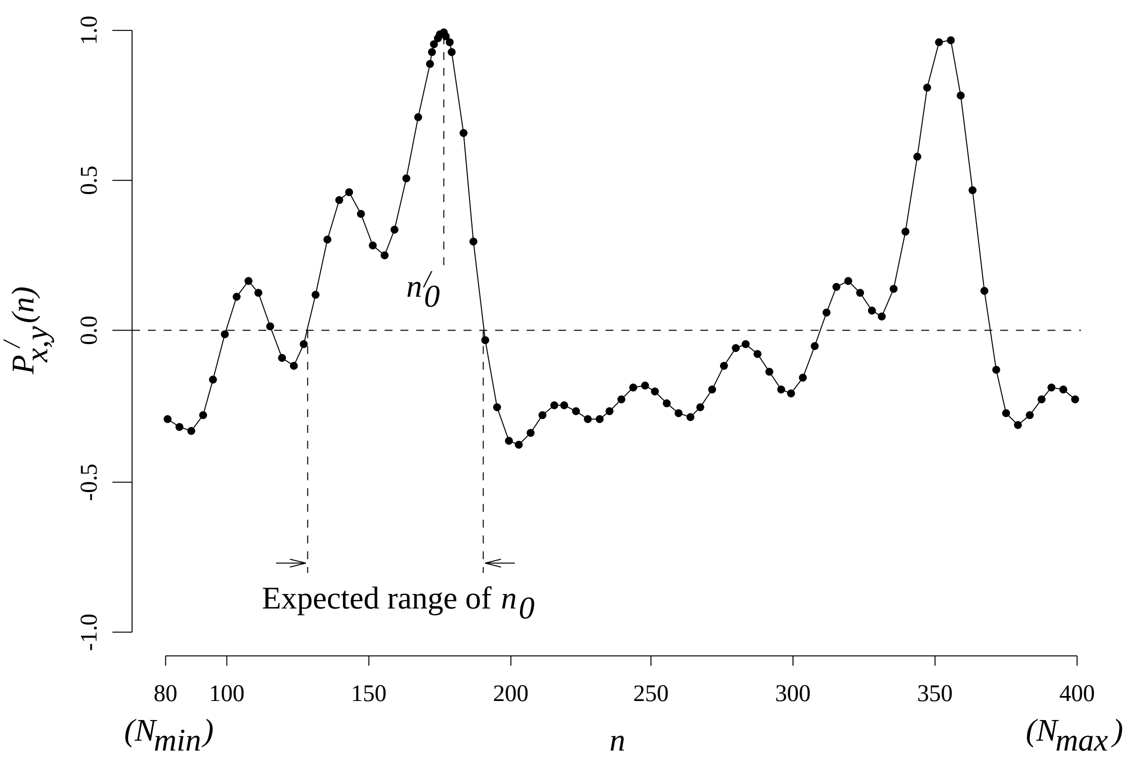


Figure 7.2: Example set of $p'_{x,y}(n)$ from previous frame

the region of the fundamental period for the previous frame to form a candidate for the fundamental period of the current frame. Note, however, that the decision to classify a frame of speech as voiced or unvoiced is based on the presence or absence of local maxima in $p_{x,y}(n)$ which exceed the adaptive threshold T_{srfd}. The biasing therefore tends to increase the percentage of unvoiced regions of speech being incorrectly classified as 'voiced'. In order to minimise this undesirable side effect, if the unbiased coefficient $p_{x,y}(n)$ does not exceed the threshold T_{srfd} for the candidate believed to be the best estimate of the frame fundamental period, then the *FØ* value for that frame is held until the state of the subsequent frame is known. If this next frame is classified as 'silent' or 'unvoiced', the former frame is re-classified as 'unvoiced'.

Finally, as in the unmodified version of SRFD, 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).

7.2 Evaluation of *FØ* determination algorithms

The following evaluation of the algorithms summarised in Section 3.2 and the enhanced SRFD algorithm described in the previous Section, is tailored to identify the most appropriate method of fundamental frequency extraction of adult male and female speech recorded with little ambient interference.

A database containing approximately five minutes of speech is used in the evaluation. The speech is recorded simultaneously with a close-talking microphone and laryngograph in an anechoic studio. The database is formed from fifty sentences each read by one adult male (♂) and one adult female (♀), both with non-pathological voices. The database is biased towards utterances containing voiced fricatives, nasals, liquids and glides, since these phones are aperiodic in comparison to vowels and FDAs generally find them difficult to analyse.

The sentences are recorded with the use of a laryngograph so that a reference laryngeal frequency (Fx) contour can be obtained. The laryngograph measures the impedance between two electrodes placed bilaterally across the larynx. The measured impedance decreases with an increasing degree of vocal-fold contact. Glottal closure is marked in the laryngograph signal (Lx) by a sharp rise to a peak, whereas the opening of the glottis is much slower and less marked by a gradual fall in signal amplitude (Baer *et al.*, 1983). It is not expected that the laryngograph data can be used to give a perfectly 'correct' corresponding Fx contour for the associated speech data. In particular, a contour generated from laryngograph data may be inaccurate due to effects at the end of voiced speech segments for which a small area of vocal-fold contact is insufficient for the glottis activity to be distinguished from noise in the laryngograph signal, but the speech is periodic and low in energy. The extent of such errors will only be over two or three Fx cycles and are thus deemed negligible in this study. The laryngograph data provides a simple and (relative to FDAs or locating glottal closures from the speech waveform by hand) accurate method of producing a contour with which other contours can be compared.

The functionality of all the algorithms in the evaluation is dependent upon certain thresholds and pre-determined parameters, some of which are common across the algorithms. In order to set a degree of similarity between the FDAs, all are required to present a computed *FØ* value at 6.4ms intervals through an utterance. The values are limited to the ranges of 50Hz–250Hz for the male speaker and 120Hz–400Hz for the female speaker. In cases where a fixed-length analysis frame is required by an algorithm, the frame duration has been set to 38.4ms. A frame of this duration enables at least two signal periods to reside within the frame for all fundamental frequencies greater than 52Hz, and allows sufficient data for cepstral and spectral analysis techniques. The speech data applied to the algorithms is sampled at 20kHz using a 16-bit analogue-to-digital converter (ADC). For those algorithms which operate on a low-pass filtered waveform, the 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.

7.2.1 A laryngeal frequency tracker

The reference *Fx* contours are created from the laryngograph data by using a simple pulse location algorithm and deriving the duration between successive pulses. Assume that the laryngograph data contains a positive-going pulse at the instant of each glottal closure and that any acoustic propagation delay between the glottis and the close-talking microphone (approximately 0.5ms) has been compensated.

A 'pulse' is defined as follows (see Figure 7.3): The pulse start time t_{start} is the first sample for which the amplitude is less than zero and less than or equal to the amplitude of following samples. The pulse stop time t_{stop} is the last sample for which the amplitude is less than zero and less than or equal to the amplitude of preceding samples. The pulse width t_{width} is defined as the difference between t_{start} and t_{stop}. The pulse peak amplitude a_{peak} is the maximum amplitude of samples between t_{start} and t_{stop} (always greater than zero). The pulse instant t_{pulse} is defined as the time of the first of these samples with an amplitude a_{peak}.

For laryngograph data sampled at 20kHz by a 16-bit ADC, a pulse at t_{pulse} is classed

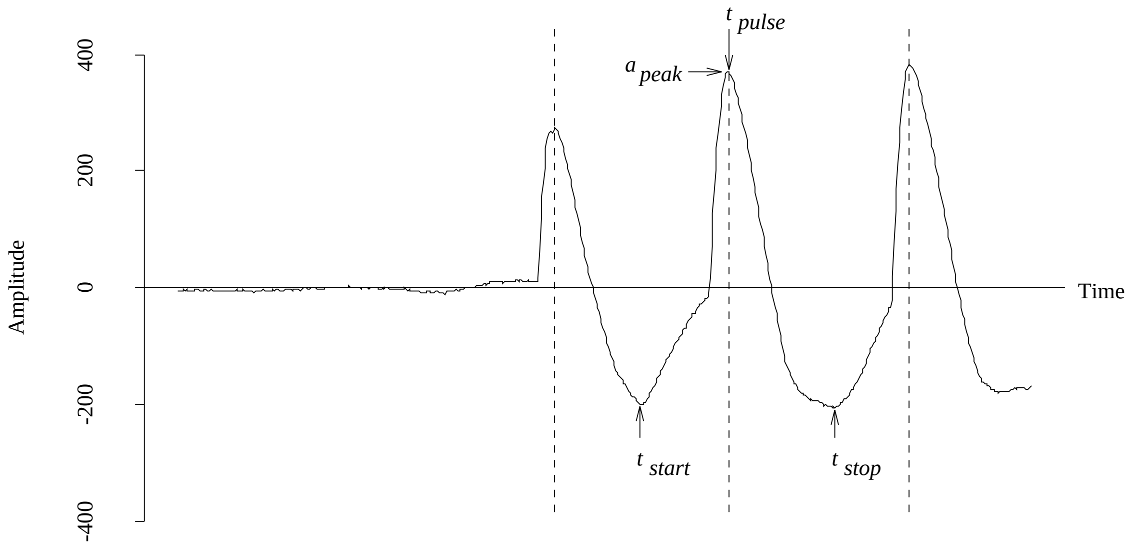


Figure 7.3: Glottal pulses evident in laryngograph data

as a marker of the glottal closure instant if the pulse width t_{width} is greater than four samples (for data from both speakers) and the pulse peak amplitude a_{peak} is greater than a threshold value which is dependent on the recording conditions and which relates to the signal-to-noise level in Lx.

The set of pulse instances generated from the laryngograph data of an utterance are considered in chronological order. The duration between one pulse instant t^n_{pulse} and the next pulse instant t^{n+1}_{pulse} is calculated and converted to Hertz. If the value is greater than some lower limit then it is taken to represent the laryngeal frequency at the time $(t^n_{pulse} + t^{n+1}_{pulse})/2$; otherwise, the duration between the pulses is considered to correspond to an unvoiced region of speech. The Fx values are limited to $\geq$50Hz for the male speaker, and $\geq$120Hz for the female speaker. There must be at least three laryngograph pulses in each voiced section. This final restriction is imposed in order to remove the few errors when a 'pulse' in the laryngograph data is formed by events other than glottal activity.

The Fx determined using this algorithm has a mean of 124.19Hz and a population standard deviation of 25.61Hz for the male speech, and a mean of 256.02Hz and standard

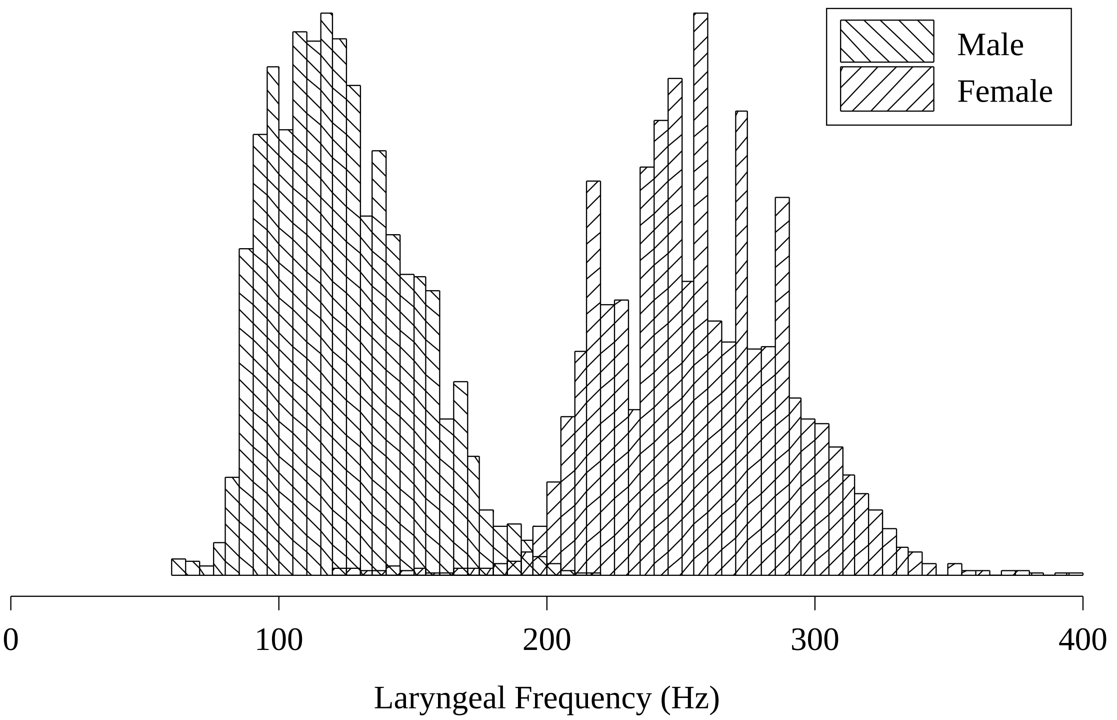


Figure 7.4: Histograms of laryngeal frequency for male and female speech

deviation of 33.39Hz for the female speech. These statistics indicate that the selection of ranges for which *FØ* is to be determined by the FDAs (♂: 50–250Hz; ♀: 120–400Hz) is suitable. The histograms of laryngeal frequency in Figure 7.4 for the male and female speech both show a single prominent lobe of data centred around the mean. There are no smaller lobes to the left or right of the main one, suggesting that Fx doubling and halving is not a problem for the algorithm.

The accuracy with which Fx can be determined by this method is limited by the time quantisation in sampling the laryngograph signal. The pulse instances can only be specified to within a single sample duration. The true laryngeal period T_x estimated by the duration between adjacent pulses d samples apart lies within the range,

$$(d-1).T_s \ < \ T_x \ \leq \ d.T_s \tag{7.4}$$

where T_s is the sampling period. The true laryngeal period can be specified with an error of just T_s seconds — note that the error is independent of d. The true laryngeal frequency $F_x = 1/T_x$ lies within the range,

$$\frac{F_s}{(d-1)} < F_x \leq \frac{F_s}{d} \tag{7.5}$$

where F_s is the sampling frequency. Each value of F_x therefore has an error of,

$$\frac{F_s}{d(d-1)} \text{ Hz} \tag{7.6}$$

This error is dependent upon d and is calculated for the distributions of Fx for the two speakers. The quantisation error in determining Fx by this method has a mean of 0.80Hz and population standard deviation of 0.34Hz for the male speaker, and a mean of 3.33Hz and standard deviation of 0.86Hz for the female speaker. This error cannot be compensated for and affects the evaluation results for the FDAs.

7.2.2 Comparison of asynchronous frequency contours

The reference Fx contour of each sentence, generated from the laryngograph data, is compared with the $F\O$ contours generated by the six algorithms reviewed in Section 3.2 and the enhanced $F\O$ determination algorithm eSRFD described in Section 7.1, all of which operate on the speech data.

Frequency contours are described by a sequence of two variables — a time from the start of the utterance and a frequency value at that time. The frequency value is set to zero to describe regions of silence or unvoiced speech. For the reference Fx contour, the times from the start of an utterance (where an Fx value is stated) are given by the mid-point times between pulses detected in the laryngograph data. The laryngeal frequency between such times must be inferred by interpolation. Similarly, for the $F\O$ contours generated by the FDAs in which an $F\O$ value is only stated at regular intervals; linear interpolation is used to form an approximation of the $F\O$ value between times for which it is stated.

For each time a frequency value is stated in either the reference Fx contour or the $F\O$ contour, the two values for the contours (one of which may be derived by linear interpolation) are compared in the following manner (see Figure 7.5). The frequency value from the reference contour is represented by $Fx_{reference}$, and that from the *FØ* contour is represented by $F0_{FDA}$.

- When $Fx_{reference}$ and $F0_{FDA}$ are zero, both contours describe a silent or unvoiced region of the utterance and no error results.
- If $F0_{FDA}$ is non-zero but $Fx_{reference}$ is zero, then an unvoiced (or silent) region of speech has been incorrectly classified as voiced by the FDA. The duration of the erroneous region is determined by finding the subsequent time at which either the *FØ* contour becomes zero or the reference *Fx* contour becomes non-zero.
- If $Fx_{reference}$ is non-zero but $F0_{FDA}$ is zero, then a voiced region of speech has been incorrectly classified as unvoiced (or silent) by the FDA. The duration of the region in error is determined by finding the subsequent time at which either the reference *Fx* contour becomes zero or the *FØ* contour becomes non-zero.
- When both $F0_{FDA}$ and $Fx_{reference}$ are non-zero, the contours represent the correct classification of voiced speech. In such cases, if the ratio of,

$$\frac{Fx_{reference} - F0_{FDA}}{Fx_{reference}} \tag{7.7}$$

  is greater than 0.2 then the FDA made a gross error in estimating the fundamental frequency of more than 20% of the reference *Fx* and the error is categorised as *FØ* 'halving'. If the ratio in Equation 7.7 is less than -0.2, a gross *FØ* error of more than 20% was made which is categorised as *FØ* 'doubling'[1]. Otherwise, the FDA is assumed to have estimated the fundamental frequency with an acceptable accuracy, and the absolute difference between $F0_{FDA}$ and $Fx_{reference}$ is noted. The $\pm$20% threshold of acceptability is chosen because all FDAs are expected to form an *FØ* value within this range with due consideration of time quantisation errors and the finite frequency resolution of the analysis technique.

The durations of unvoiced or silent regions classified in error and the durations of voiced sections incorrectly classified as unvoiced or silent by the FDA, are accumulated

[1] *FØ* 'halving' and *FØ* 'doubling' terms are used here to distinguish between abnormally low and abnormally high estimates of *FØ* rather than referring to exact octave errors.

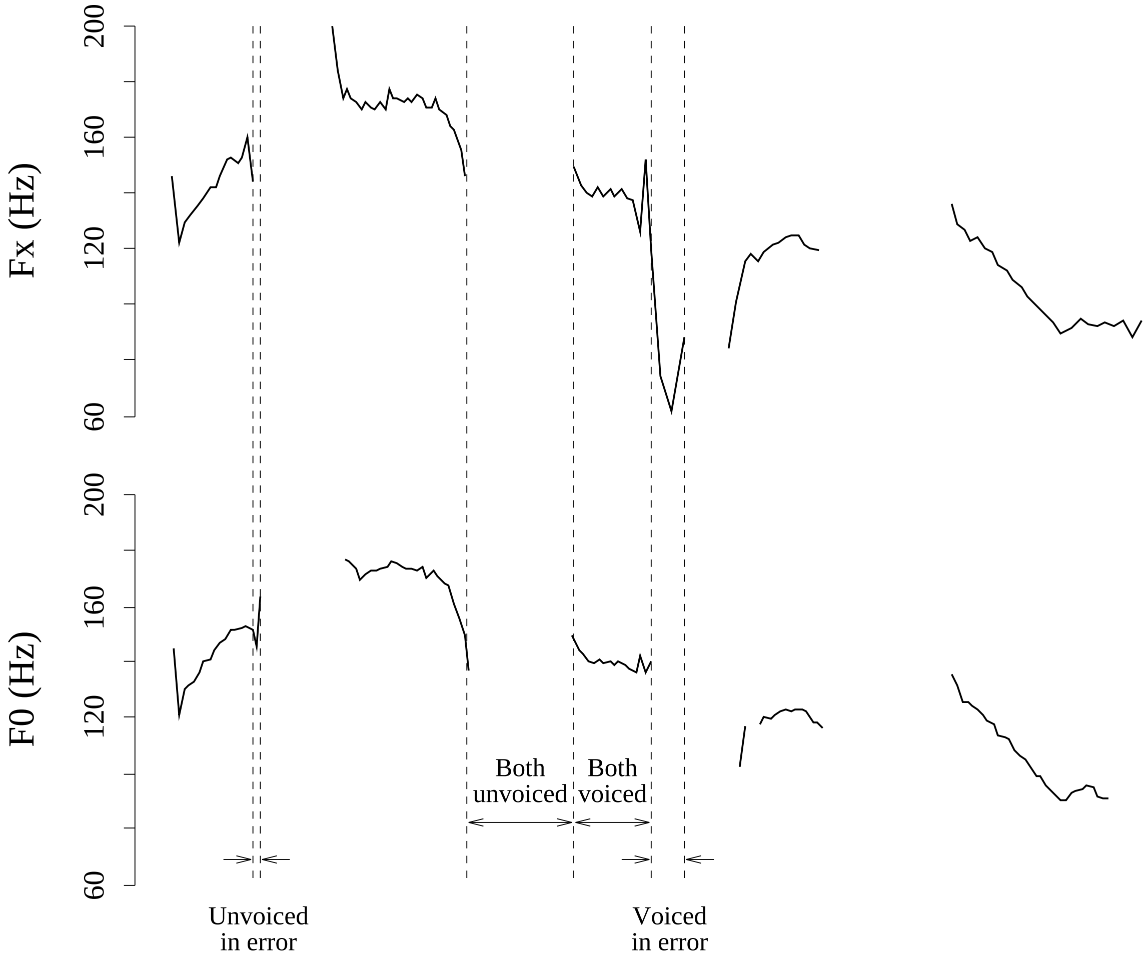


Figure 7.5: Comparison of asynchronous Fx and $F\emptyset$ contours

over all the utterances in the database for each speaker. The sum of the individual durations of such erroneous regions are expressed as a percentage of the total duration of unvoiced (or silent) speech and voiced speech respectively. The total number of comparisons for which a gross *FØ* doubling error occurs and the total number of gross *FØ* halving errors are also determined. These gross *FØ* error rates are expressed as a percentage of the total number of comparisons for which $F0_{FDA}$ and $Fx_{reference}$ are non-zero. The final statistics calculated in the evaluation of an FDA are the population standard deviation (p.s.d.) and the average, absolute deviation of the reference *Fx* and *FØ* contours when both represent voiced speech, and the FDA has not made a gross *FØ* error.

7.2.3 Results and discussion

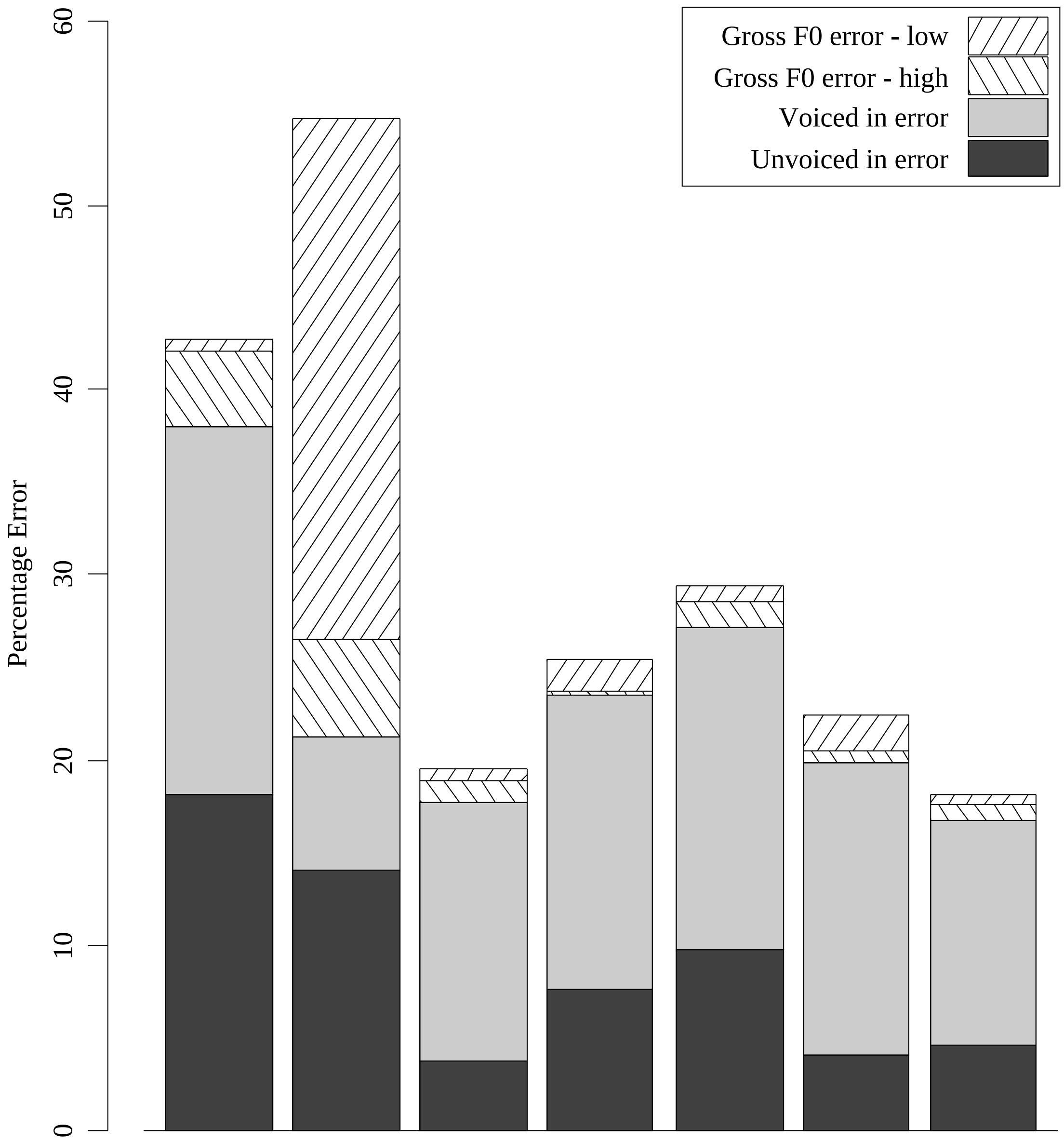
Each of the six FDAs reviewed in Section 3.2 and the enhanced SRFD algorithm described in Section 7.1 is used to extract the fundamental frequency from the speech waveforms of the utterances in the database. The contours generated are then compared with the reference *Fx* contours formed from synchronised laryngograph data, using the method described above. The results of the comparison of each pair of contours are averaged over the entire set of utterances for each speaker so that the evaluation of the FDAs may be assumed to be independent of the segmental and prosodic structures of the read utterances. The resultant statistics are presented in Figures 7.6 & 7.7 for the two speakers.

Voiced/Unvoiced Classification: CFD and HPS (which use a frequency domain representation of speech) perform considerably poorer for the female speech than for the male speech when classifying frames of speech as either voiced or unvoiced. Results reported by Rabiner *et al.* (1976) also find the cepstral-based *FØ* determination to perform much better on lower frequency speech than on higher frequency speech. This is a consequence of the frequency domain algorithms basing the voicing classification on the magnitude of spectral peaks rather than using the temporal structure of the waveform. The maximum cepstral peak magnitude in a given frame is a measure of the degree of voice periodicity; however, voiced speech is only approximately periodic. During sudden changes in articulation, the speech waveform can be aperiodic although voiced, for example in the transition from a vowel to a nasal segment. In such cases, a single feature such as the cepstral peak magnitude is insufficient to distinguish voiced speech from unvoiced speech (Atal & Rabiner, 1976). The time domain approaches, FBFT, PP, IFTA and SRFD, are consistently better, with both FBFT and *e*SRFD having an overall voicing decision error rate of less than 17%.

Gross FØ *Errors:* The outstanding number of *FØ* halving errors produced by HPS for the male speech is unacceptably high. Large sections of aperiodic voiced speech are often analysed by the HPS algorithm to have an erroneously low fundamental frequency. For the male speech, CFD and HPS generate far more *FØ* doubling errors than the time domain based algorithms. The total gross *FØ* error rates for the time domain based algorithms are less than 2.1% for the male speech and less than 4.2% for the female speech, with the exception of the unmodified SRFD which generates 5.6% *FØ* halving errors for the female speech. This error rate is dramatically reduced to 0.2% by the modified algorithm *e*SRFD.

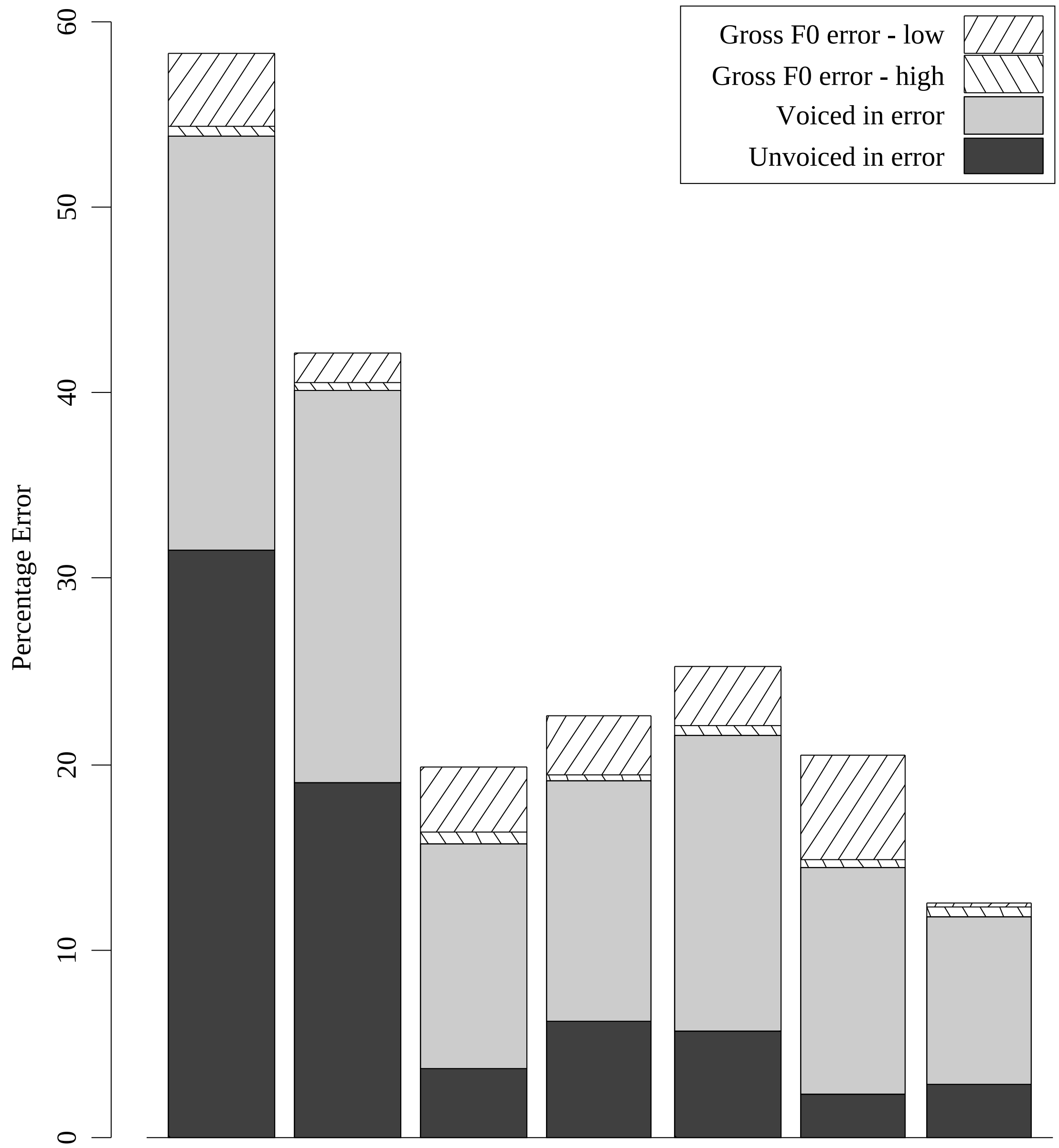
Frequency Accuracy: The accuracy with which the fundamental frequency is determined by the FDAs is not of great importance for the analysis of intonation since only the general trend of *FØ* is required. However, it can be seen that the super resolution *FØ* determinator performs consistently better that the other FDAs in this category, with the accuracy of its contours being comparable with that of the reference *Fx* contours.

Any FDA producing an *FØ* contour suitable for automatic prosodic analysis must perform consistently between male and female speech. The resultant contour must accurately identify voiced sections of speech so that pitch accents are not left undetected, and gross *FØ* errors must be minimal to facilitate correction by *FØ* post-processing techniques. CFD and HPS are therefore unsuitable algorithms for the application of automatic prosodic analysis. Of the remaining FDAs studied here, FBFT and *e*SRFD form the better voiced/unvoiced classifications and generate the least number of gross *FØ* errors. The enhanced super resolution *FØ* determination algorithm (*e*SRFD) is used to generate the fundamental frequency contours for the automatic prosodic analysis described in this thesis.



|Fx-F0|	**CFD**	**HPS**	**FBFT**	**PP**	**IFTA**	**SRFD**	**eSRFD**
mean (Hz)	2.94	3.25	1.86	2.64	2.67	1.78	1.40
p.s.d. (Hz)	3.60	3.21	2.89	3.01	3.37	2.46	1.74

Figure 7.6: FDA evaluation: Male speech



|Fx-F0|	CFD	HPS	FBFT	PP	IFTA	SRFD	eSRFD
mean (Hz)	6.39	4.59	5.40	6.11	4.38	4.14	4.17
p.s.d. (Hz)	7.61	5.31	7.03	6.45	5.35	5.51	5.13

Figure 7.7: FDA evaluation: Female speech

7.3 *FØ* contour post-processing

An attempt is made in the design of the *FØ* extraction algorithm (eSRFD) to minimise the number of *FØ* doubling and halving errors derived from the speech waveform. It is, however, inevitable that some errors of this kind will exist in the *FØ* contour and that they will propagate into subsequent analysis of the contour. The aim of post-processing an *FØ* contour is to eliminate *FØ* doubling and halving errors whilst maintaining any information in the *FØ* contour which is relevant to the prosodic analysis. Whereas the previous attempt (Section 7.1) to reduce the number of errors was based on the characteristics of the speech waveform, the post-processing of the contour aims to reduce the number of errors by taking into account the characteristics of the contour itself.

7.3.1 Non-linear smoother

In an attempt to eliminate *FØ* doubling and halving errors, the contour can be processed by a non-linear filter. Linear smoothing can be applied to remove small perturbations in the contour which are not required in the prosodic analysis. Linear smoothing also reduces the effects of frequency quantisation in the contour. The non-linear filter and linear smoothing proposed by Rabiner *et al.* (1975) is illustrated in Figure 7.8.

A non-linear median filter consists of a shift register of length l_{median}. The *FØ* values (one for each analysis frame) stored in the elements of the register are sorted into ascending order and the middle value is presented at the output of the median filter. The length l_{median} is always an odd number.

A linear Hann window is used as a smoothing filter. The window uses l_h values selected from a shift register of length l_{Hann} and weights the n'th value by a factor $h(n)$.

$$h(n) = \frac{1}{l_h + 1}\left(1 - \cos\frac{2\pi n}{l_h + 1}\right), \qquad n \in 1, \ldots, l_h \tag{7.8}$$

If the shift register contains more than $l_{Hann}/2$ *FØ* values which represent frames of unvoiced speech or silence, then the output of the filter is also set to represent unvoiced speech. Breaks in an *FØ* contour are represented by values equal to zero, which correspond to frames of unvoiced speech and frames of silence. Otherwise, l_h is set to the number of *FØ* values in the shift register which represent frames of voiced speech and

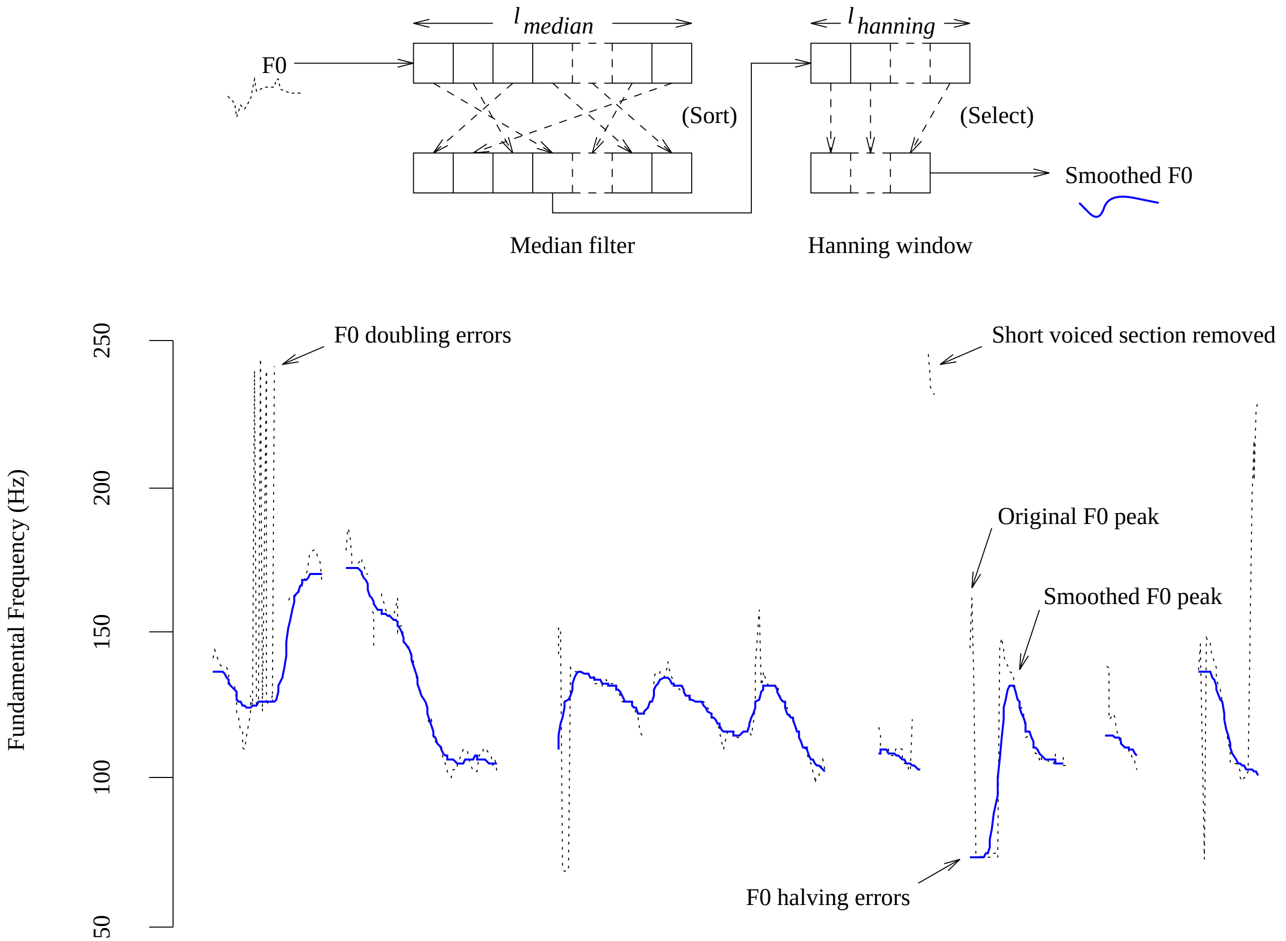


Figure 7.8: Non-linear smoother

the filter output is set to the sum of their weighted values.

The algorithm uses extrapolation for the initial and final values of the contour in order to prevent small time shifts due to the delay imposed by the smoothing process. By using a Hann window for which l_{Hann} is an odd number, the delay $(l_{median} + l_{Hann} - 2)$ imposed by the smoother is always an even number. The extrapolation at the beginning of the track accounts for half of this delay, and extrapolation at the end accounts for the other half.

An example of the effects of this non-linear smoothing process is illustrated in Figure 7.8. In this example, the number of elements in the shift register of the median filter and the maximum number of elements in the Hann window are set to 17 frames. Although some of the short-term irregularities of the *FØ* contour which are not relevant to prosodic analysis are removed by this process, some of the inflections of the *FØ* contour which need to be preserved for prosodic analysis are affected. There are two major disadvantages in applying a non-linear smoother to an *FØ* contour prior to prosodic analysis. Firstly, short isolated regions of the *FØ* contour which represent sections of voiced speech spanning less than $(l_{median} + 1)/2$ frames are removed if they are flanked by sequences of unvoiced speech each spanning $(l_{median} + 1)/2$ frames or more. Secondly, peaks in the *FØ* contour related to pitch accents are time-shifted and altered in magnitude if they occur near *FØ* doubling or halving errors. If the numbers of elements in the shift registers are reduced in order to preserve *FØ* inflections relevant to the prosodic analysis then fewer *FØ* doubling and halving errors are removed. If the numbers of elements in the shift registers are increased in order to remove larger sections of an *FØ* contour which are subject to doubling and halving errors then the magnitudes of *FØ* peaks are reduced. Changing the magnitudes of the *FØ* peaks is undesirable because they are needed for the prosodic analysis. Another method of removing *FØ* doubling and halving errors is required so that the number of elements in the shift registers can be reduced, thus enabling relevant *FØ* inflections to be preserved whilst still being able to smooth *FØ* micro-perturbations which are irrelevant to the prosodic analysis.

7.3.2 De-step *FØ* filter

An algorithm is proposed in this Section which eliminates *FØ* doubling and halving errors in order to overcome the disadvantages inherent to the non-linear smoother. The algorithm is illustrated in Figure 7.9.

It is necessary to distinguish legitimate frame-to-frame variations of the *FØ* contour from changes which are due to errors in estimating the fundamental frequency. Once this distinction is made, the *FØ* changes which correspond to doubling and halving errors can be corrected. It is assumed that *FØ* can legitimately increase or decrease from frame-to-frame by up to 75% during continuously voiced sections of speech, and that *FØ* can change by any amount across unvoiced sections of speech. An *FØ* increase from one frame to the next of more than 75% is defined for this algorithm to be an *FØ doubling* (exact doubling is a 100% increase) superimposed on a legitimate decrease of up to 25%. Similarly, an *FØ* decrease from one frame to the next of more than 75% is defined to be an *FØ halving* superimposed on a legitimate increase of up to 25%.

Consider each section of voiced speech which is uninterrupted by any frames of unvoiced speech. The onset *FØ* value of a section of voiced speech can correspond to an *FØ* doubling or halving error, or can be a legitimate value. A technique is required to determine which of these states the *FØ* value is in. Each *FØ* value in the section of voiced speech is therefore placed into one of a number of pools. The first *FØ* value is placed into pool P_x. If the change in *FØ* from one frame to the next corresponds to an *FØ* doubling (in terms of the definition above) then all *FØ* values up to the next *FØ* doubling or halving are placed into a higher pool (x is increased). Similarly, if the change in *FØ* corresponds to an *FØ* halving then all *FØ* values up to the next *FØ* doubling or halving are placed into a lower pool (x is decreased). The pool containing the greatest number of values is assumed to contain legitimate *FØ* estimates. This assumption is valid only if an FDA makes a majority of the *FØ* estimates without doubling or halving errors in any given section of voiced speech. Let $x = 0$ for the pool containing the greatest number of values. Higher pools ($x > 0$) contain *FØ* values which are doubling errors and lower pools ($x < 0$) contain *FØ* values which are halving errors.

To describe this explicitly in mathematical terms, let f_i represent the *FØ* value of the i'th frame in a section of voiced speech. The first *FØ* value f_1 is placed into pool $P_{x(1)}$.

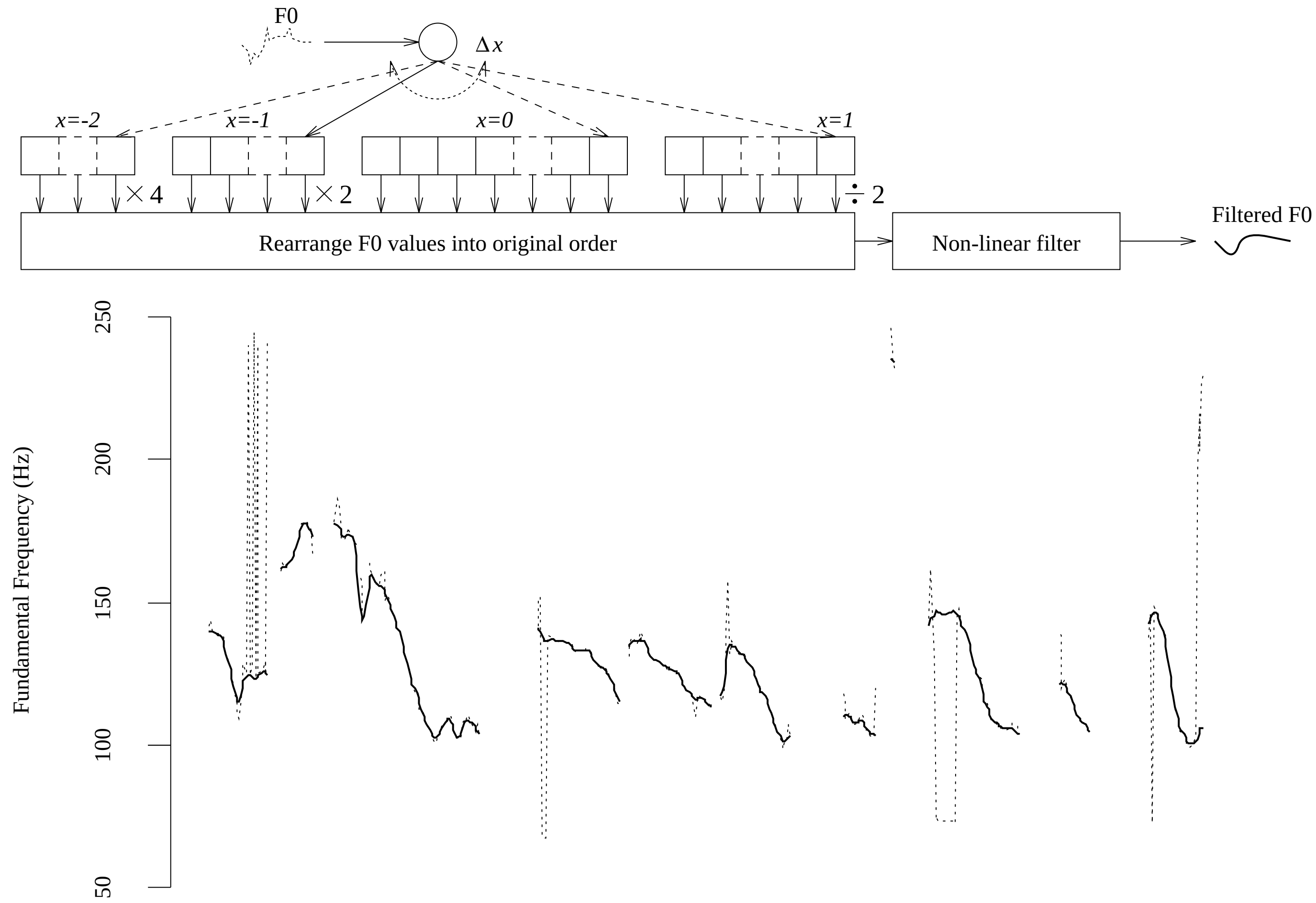


Figure 7.9: A de-step *FØ* filter and non-linear filter

Subsequent values ($i > 1$) are placed into a pool $P_{x(i)}$ where $x(i)$ is given in Equation 7.9. *FØ* values are pooled with the index i in order to keep track of their location in the *FØ* contour.

$$x(i) = \begin{cases} x(i-1) + \left\lfloor \log_2 \left(\frac{4}{7} . \frac{f_i}{f_{i-1}} \right) + 1 \right\rfloor & \text{if } f_i \geq f_{i-1} \\ x(i-1) - \left\lfloor \log_2 \left(\frac{4}{7} . \frac{f_{i-1}}{f_i} \right) + 1 \right\rfloor & \text{if } f_i < f_{i-1} \end{cases} \tag{7.9}$$

Let P_0 represent the pool containing the greatest number of *FØ* values. The *FØ* values in $P_{x(i)}$ are corrected by multiplying each value by $2^{-x(i)}$.

$$f_i' = 2^{-x(i)} f_i \qquad \forall f_i \in P_{x(i)} \tag{7.10}$$

Once the correction factor has been applied, the *FØ* contour for the voiced section is reconstructed from the new *FØ* values f_i' and the stored indices i.

An example of the effects of this de-step *FØ* filter on the performance of the non-linear smoothing process is illustrated in Figure 7.9. In this example, the number of elements in the shift register of the median filter and the maximum number of elements in the Hann window are set to 5 frames. The *FØ* doubling and halving errors are eliminated without affecting inflections of the *FØ* contour relevant to the prosodic analysis. *FØ* micro-perturbations are reduced by applying the non-linear smoother with a reduced number of shift register elements (reduced from spanning 17 frames to 5 frames). Short sections of voiced speech are therefore preserved for further analysis.

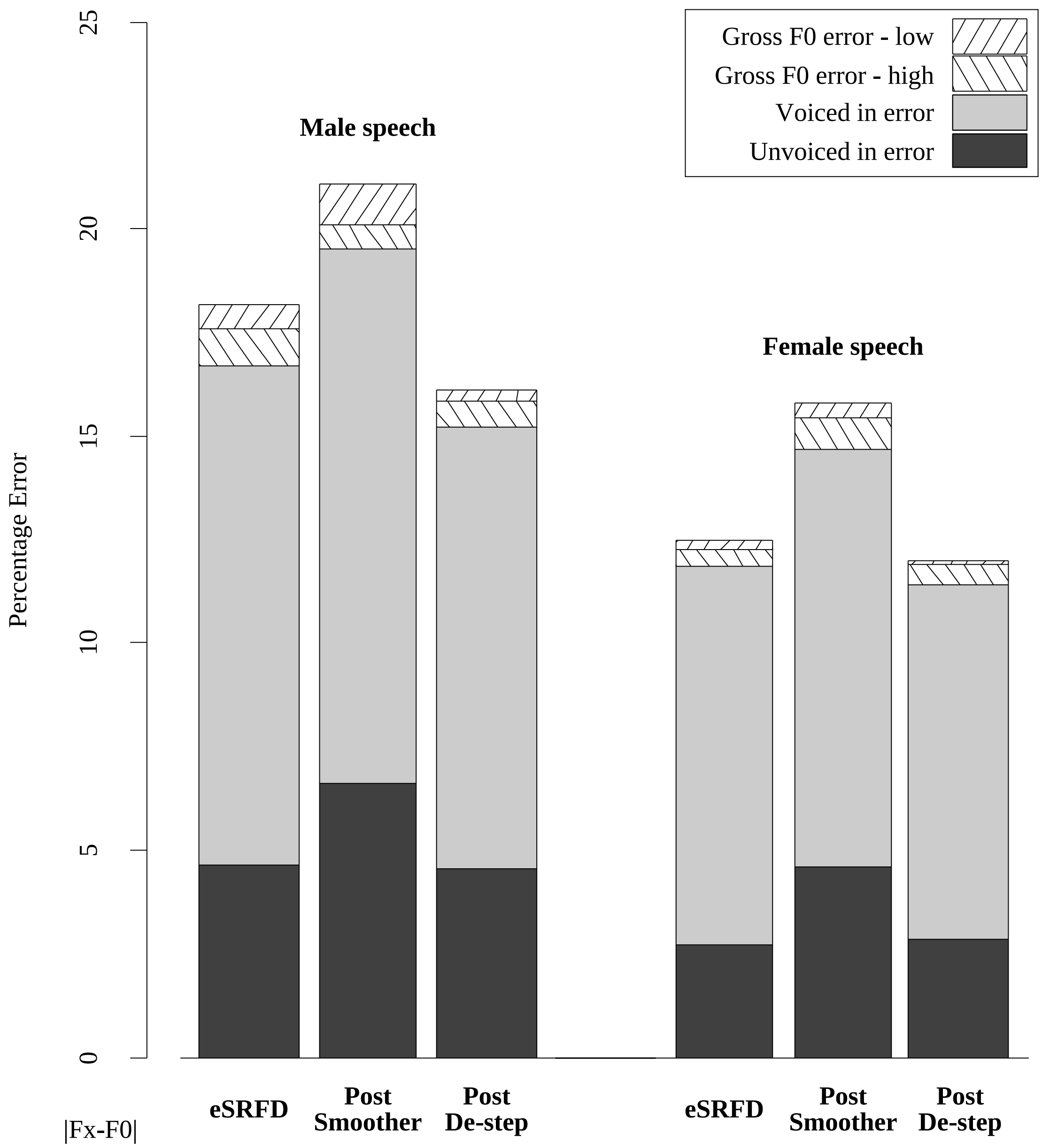
Some errors in the *FØ* contour can remain after this post-processing technique has been applied. If, in the course of a single voiced section of speech, an *FØ* doubling error is detected but a subsequent *FØ* halving error is not detected (or visa versa) then legitimate *FØ* values are altered. This produces doubling and halving errors rather than eliminating them. Legitimate *FØ* values are corrupted in this way only when a doubling error occurs during a legitimate rapid decrease in *FØ* such that the overall frame-to-frame increase is less than 75%, or when a halving error occurs during a legitimate rapid increase in *FØ* such that the overall frame-to-frame decrease is less than 75%. The possibility of this undesirable effect occurring is small and is far outweighed by the benefits of the de-step *FØ* filter.

Three other types of error can remain in the *FØ* contour. Firstly, if an FDA makes either more $F\O$ doubling errors or more $F\O$ halving errors than correct estimates of the fundamental frequency (in any given section of voiced speech) then pool P_0 will not contain legitimate values, as assumed. The $F\O$ values in the voiced section will therefore be erroneously set to an overtone of the their true values. The only way to remove such errors is to improve the $F\O$ determination algorithm. Secondly, the $F\O$ contour may contain rapid frame-to-frame changes in frequency of more than 75% which are accurate estimates of the rate of glottal pulses, but which are detected as doubling or halving errors — such as in creaky voice. Finally, $F\O$ inflections dependent on the segmental content of the utterance remain in the contour (such as $F\O$ peaks at the onset of vowels occurring after unvoiced stop consonants) although they are not required by the prosodic analysis. A further level of post-processing is required, such as $F\O$ contour stylisation (Sections 3.4 & 8.2.1), to eliminate these microprosodic variations in the $F\O$ contour.

An evaluation of the $F\O$ contours produced by the eSRFD algorithm when post-processed by the non-linear smoother and the de-step filter are shown in Figure 7.10. This shows that the non-linear filter only decreases $F\O$ doubling errors in the male speech, while the de-step filter combined with a non-linear filter using fewer frames yields improvements in voicing classification and a reduction in gross $F\O$ errors. These improvements are made at the expense of a slight reduction in the accuracy of $F\O$ estimates.

7.4 Summary

Intonation refers to the manipulation of pitch for linguistic and paralinguistic purposes above the level of phonemic segments. The acoustic correlate of pitch is the fundamental frequency of speech. The extraction of $F\O$ from a speech signal is therefore a prerequisite of prosodic analysis. A number of modifications to the SRFD algorithm (summarised in Section 3.2.6) are proposed in this Chapter, which reduce the occurrence of errors involved in the extraction of $F\O$ such that it is optimised for prosodic analysis, relative to a selection of other algorithms designed to determine $F\O$. A novel de-step filter is proposed to post-process an $F\O$ contour generated by an FDA in a way which further reduces the occurrence of discontinuity errors commonly observed in $F\O$ contours.



|Fx-F0|	eSRFD	Post Smoother	Post De-step	eSRFD	Post Smoother	Post De-step
mean (Hz)	1.40	3.50	1.60	4.17	6.25	4.22
p.s.d. (Hz)	1.74	4.09	2.22	5.13	7.29	5.33

Figure 7.10: Evaluation of post-processed *FØ* contours (for *e*SRFD)