Chapter 8

Automatic Prosodic Analysis

The features of duration and energy, and the fundamental frequency extracted from the speech waveform need to be abstracted to form an acoustic-phonetic representation of sentential stress patterns and intonation in a syllabic domain. This is the aim of the automatic prosodic analysis system proposed here.

The syllabification of speech forms a central part of automatic prosodic analysis, taking on three roles. Firstly, the prosodic aspects of speech are described in a syllabic domain. Stress refers to the relative perceptual prominence of syllables and the description of intonation involves the association of pitch accents with prominent syllables. Secondly, the extraction of acoustic parameters is dependent upon the definition and identification of syllables. The duration feature found to optimally correlate with stress in Chapter 4 is dependent upon identifying the syllable lhymes in an utterance. The optimal energy feature determined in Chapter 5 is dependent upon identifying the syllable nuclei in an utterance. The third role of the syllabic domain is related to the integration of the acoustic parameters. The syllabic domain inherently encompasses energy and segmental information which can be passed on to the analyses of energy, duration and *FØ* measures. For example, the location of syllable nuclei can serve as possible islands of reliability in the extraction of fundamental frequency. Section 8.1 addresses the automatic syllabification of an utterance from its phonetic realisation, and the syllabification of an utterance on the basis of a set of abstract phonological rules.

Microprosodic variations in the fundamental frequency of a speech waveform cause inflections in an *FØ* contour which are independent of the underlying intonational component. An *FØ* contour also includes speaker-dependent effects, fluctuations related to cycle-to-cycle jitter (*FØ* perturbations) generated at the vocal folds, and erroneous *FØ* estimates generated by malfunctions in an FDA. It is proposed that an acoustic-phonetic representation of intonation can be derived from a raw *FØ* contour by removing components of the contour which are due to microprosody, cycle-to-cycle jitter, speaker dependent *FØ* range and fluctuations induced by erroneous *FØ* extraction from the speech signal. The task of reducing the occurrences of *FØ* perturbations and extraction errors is largely accomplished by the post-processing techniques described in Section 7.3.

The stylisation of an *FØ* contour aims to prevent microprosodic variations from being confused as pitch accents, and hence effectively isolate the microprosodic and intonational components. The process of piece-wise linear stylisation of an *FØ* contour is presented in Section 8.2.1. *FØ* estimates made within syllable nuclei are generally reliable because the speech signal within syllable nuclei is quasi-periodic and has a relatively high signal-to-noise ratio. These conditions are generally well suited to the assumptions made by *FØ* determination algorithms. Furthermore, microprosodic variations tend to dominate an *FØ* contour within the vicinity of short syllable nuclei in the context of unvoiced consonants. Syllable information can therefore be used to aid the stylisation of an *FØ* contour.

A stylised contour may contain some microprosodic variations given that the stylisation process will not be faultless. A stylised contour also contains speaker-dependent and intonational components. The stylised contour must be processed with respect to the syllables of an utterance in order to eliminate any remaining microprosodic variations, to compensate for speaker-dependent effects and to form an acoustic-phonetic representation of intonation. A schematisation process aims to manipulate a stylised contour with these goals in mind. A schematic representation of an *FØ* contour is independent of a speaker's fundamental frequency range, and aims to exhibit only the intonational component of the contour. An algorithm to generate *FØ* schemata is presented in Section 8.2.2. The *FØ* trajectories of a schema can be associated with syllables to locate the syllables which are pitch accented.

The processed acoustic parameters need to be combined to form a description of the intonation and the sentential stress patterns of speech, in a syllabic domain. The modular analyses proposed in this thesis are integrated to form a prosodic analysis system which generates such a description. An overview of the system is described in Section 8.4 and its performance is evaluated in Section 8.5 relative to two algorithms formerly proposed in the literature.

8.1 Syllabification

The 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. An algorithm is proposed in Section 8.1.1 for the syllabification of speech from acoustic-phonetic parameters. The algorithm groups phones according to the phonetic realisation of an utterance rather than on the basis of a set of abstract phonological rules. A procedural definition of a syllable based on phonological rules is presented in Section 8.1.2. The groups of phones produced by the acoustic-phonetic syllabification are shown in Section 8.1.3 to correlate with syllables defined on the basis of phonological rules.

8.1.1 Automatic syllabification from acoustic-phonetic parameters

The following algorithm is used to group phones into syllable-sized units using a (frame-level) utterance-normalised low-band energy contour which is generated using the method described in Section 5.1, and the phone boundary and label information of an utterance provided either by hand or by an automatic segmentation system.

In the application of prosodic analysis for computer aided pronunciation teaching, the orthographic transcript of an utterance is known because a foreign language learner is asked to read a given sentence in the course material. An automatic segmentation algorithm can therefore use a network of variant pronunciations (which may also include optional interword silences and segmental assimilation) for the known utterance (McInnes *et al.*, 1992). The constraints of such a network enable automatic segmentation to be performed more reliably than in speech recognition systems where segmentation is either constrained only by vocabulary and syntax, or even worse, no vocabulary and syntax constraints are imposed (as in phoneme lattice generation).

The (auto-)segmentation data gives the location of phones and classifies phone-types. The low-band energy contour of an utterance is used to determine the degree of syllabic cohesion between two adjacent phones.

All local minima (valleys) in the low-band energy contour are located. The locations of the minima form candidates for syllable boundaries. The areas of silence identified by the (auto-)segmentation are respected and the energy minima within such areas are believed to be due to variations in background noise. Each segmentation boundary at the junction of a silence and a phone label is regarded as a syllable boundary (at either the beginning or at the end of a syllable). The nearest candidate (energy minimum) to such a segmentation boundary is therefore moved to align with it and all the candidates residing within the area of silence are disregarded.

If word boundary information is known, then syllable boundaries can be forced to occur at word boundaries. The nearest syllable boundary candidate to a word boundary can be aligned with the word boundary. Word boundary information is not used in this study and so phones which are grouped using this algorithm may span across word boundaries. None of the analysis techniques described in Chapters 4, 5, 6 & 7 make use of word boundary information.

The regions between all the remaining energy minima (after those within silences have been disregarded) are taken to be potential syllables, with a start time given by the location of the nearest minimum on the left-hand side, and the location of the nearest minimum on the right-hand side giving the stop time. It is then determined whether or not the location of each of these potential syllables overlaps more than 50% of any vowel segment. If a potential syllable overlaps more than one vowel segment in this way, then the vowel segment with the maximum low-band energy is taken to be the nucleus of the syllable. If there is more than one vowel in a group of phones then any reduced vowels (/ə/) take a lower precedence in forming the syllable nucleus than non-reduced vowels, even if the reduced vowel has a greater low-band energy than the other vowel or vowels in the group of phones. (It is unlikely that a schwa will really have a higher low-band energy than a neighbouring full vowel, but this situation may arise because of errors in the automatic segmentation of an utterance.)

If no such overlap occurs, then it is determined whether or not the location of the potential syllable overlaps more than 50% of one of the possible syllabic consonant segments /l, m, n, (r)[1]/. Again, if the potential syllable overlaps more than one of these, the one with the maximum low-band energy is selected as the syllabic nucleus.

If there is insufficient overlap, the region between the minima does not correspond to a syllable unit (because no syllable nucleus can be found between the low-band energy minima) and either the left-hand side or right-hand side minimum is disregarded as a syllable boundary candidate — whichever has the highest energy and does not correspond to a phone/silence boundary or (if known) a word boundary. The newly formed region is then taken to be a potential syllable and the process is repeated.

The resultant syllabification has boundaries located in the utterance at positions of local minima in the low-band energy contour. These boundaries are aligned with the (auto-)segmentation by moving each of the boundaries to the nearest phone boundary.

8.1.2 Syllabification based on phonological rules

The following procedural definition is used to group phones into syllables on a phonological basis rather than on an acoustic-phonetic basis (Bagshaw & Williams, 1992).

i) Consonantal phones (such as /m, n, l, r, s/) which may result in schwa deletion (Gimson, 1970; Dalby, 1986) and take on the syllabic nucleus, are syllabified as if the underlying schwa is present. Hence, in rapid speech, *shortest* may be syllabified as [ˈʃ ɔ – t s̩ t] and *additional* may be syllabified as [ə – ˈd ɪ – ʃ n̩ – l̩]. A glottal stop /ʔ/ that may occur before or instead of a word-final stop consonant is treated as an instance of the underlying stop consonant, and the glottalised onset to a vowel is considered to be part of the vowel.

ii) Syllable boundaries are formed from the boundaries of words considered in isolation. Although in connected speech consonants at the end of one word can syllabify with the initial vowel of the following word (Maddieson, 1985), such resyllabification is not necessary in forming a domain in which to describe prosodic events. Thus, for example, the syllabification of *at all* differs to that of *a tall* even if [t] is aspirated in both cases and they are phonetically identical. Similarly, resyllabification is unnecessary across words that appear to blend together due to vowel deletion, as may be the case in *under a*, which is syllabified as [ˈʌ n – d r̩ – ə] rather than [ˈʌ n – d r ə]. The syllabification of

[1]/r/ is included for American English and for rhotic dialects of British English.

under a and *tundra* /ˈt ʌ n – d r ə/ therefore differ even though they may be phonetically similar. This approach is adopted because the exact boundaries between syllable nuclei are not of critical importance, although identifying the syllable nuclei is important.

iii) The boundaries between syllables are also determined by the presence of morphological boundaries. The boundary between a free morpheme and an inflectional suffix (except *-s* — eg. *-tion*, *-tician*) or a class-II derivational affix (eg. *un-*, *non-*, *dis-*) is taken to be a syllable boundary (see Fudge (1984) for details of morpheme prefixes and suffixes in English). Thus, *hopeless* is syllabified as /ˈh əʊ p – l ə s/ rather than /ˈh əʊ – p l ə s/; and *uninteresting* is syllabified as /ʌ n – ˈɪ n – t ə – r ɛ s t – ɪ ŋ/ rather than /ʌ – ˈn ɪ n – t ə – r ɛ – s t ɪ ŋ/.

iv) On the basis of English phonotactics, any cluster of phones forming the onset or the coda of a syllable must also be a permissible word-initial or word-final cluster. According to this rule, *extra* may be syllabified as /ˈɛ k – s t r ə/, /ˈɛ k s – t r ə/, or /ˈɛ k s t – r ə/.

v) The 'maximal onset (and minimal coda) principle' (Pulgram, 1970; Couper-Kuhlen, 1986) arbitrates between competing syllabifications. According to the principle, as many consonantal phones as possible form a syllable onset. Using this principle, *extra* would be syllabified as /ˈɛ k – s t r ə/. However, in cases when alternative boundaries are possible, stressed syllables tend to attract consonants more than unstressed ones, particularly in the case of ambisyllabic consonants such as /s, f/ (Fudge, 1984). When this final criterion is applied, the syllabification adopted for *extra* becomes /ˈɛ k s – t r ə/.

8.1.3 Evaluation procedure

An example of the syllabifications produced by the above methods is shown in Figure 8.1 for the word *international*, in the phrase "...at the Kyoto International Conference Centre." The upper part of the Figure shows the speech waveform and its corresponding segmental transcription (obtained by auto-segmentation). Phone boundaries are shown by dotted lines and the continuous lines show the syllable boundaries derived on a phonological basis. The lower part gives the utterance-normalised low-band energy contour and transcription-aligned syllable boundaries derived using the acoustic-phonetic syllabification algorithm described in Section 8.1.1. The phonologically based syllabification gives five syllables, /ɪ n – t ɚ – n æ – ʃ ə n – ə l/, but the acoustic-phonetic syllabification gives only four syllables, [ɪ n – t ɚ – n æ ʃ – ə n ə l]. The final syllable [– ə n ə l] may initially appear to illustrate an error in the acoustic-phonetic syllabification algorithm (that of a *missing* syllable boundary). However, the low-band energy contour shows only one peak of intensity in this section of speech, and it is transcribed phonetically by a phonetician as [– n ə l] (with schwa deletion). There is an error in the automatic segmentation of the utterance, but not in its subsequent syllabification.

A database of 591 utterances from the English language ATR conference-registration dialogues is syllabified automatically using the above acoustic-phonetic algorithm and by hand using the syllabification based on phonological rules. The data consists of a series of telephone dialogues spoken by a female bilingual speaker of American English and Japanese in a variety of styles and registers, ranging from read speech to spontaneous speech (see Campbell (1992) for a brief description of this material).

In order to compare the human and machine syllabifications, it is necessary to define when they are regarded as *sufficiently similar* and when they are not. At the most stringent level of comparison, a syllable located automatically is regarded as correct only if its boundaries (both at the beginning and at the end of the syllable) match exactly with those defined on a phonological basis. This method of comparison is unrepresentative of the algorithm's performance as just one miss-matched boundary would correspond to two miss-matched syllables. Furthermore, the boundaries between syllables vary even between phonological definitions — only, the syllable nuclei are well defined. Therefore, the following procedure is used to evaluate the automatic syllabification algorithm.

Initially assume that all of the syllables located automatically are *extra* syllables — ie. that each auto-syllable is not the single match of a phonological syllable. Also assume that all of the phonological syllables are *missing* in the automatic syllabification — ie. that each hand-syllable is not matched by any syllables located automatically. The phonetic transcription of every syllable is known. By definition, there is at least one vowel and/or syllabic consonant in each syllable (defined either automatically or on a phonological basis). However, the phonetic units which form each syllable nucleus are not stipulated (as they may differ from one syllabification scheme to another). The comparison proceeds by considering each hand-syllable in turn. Locate the first vowel

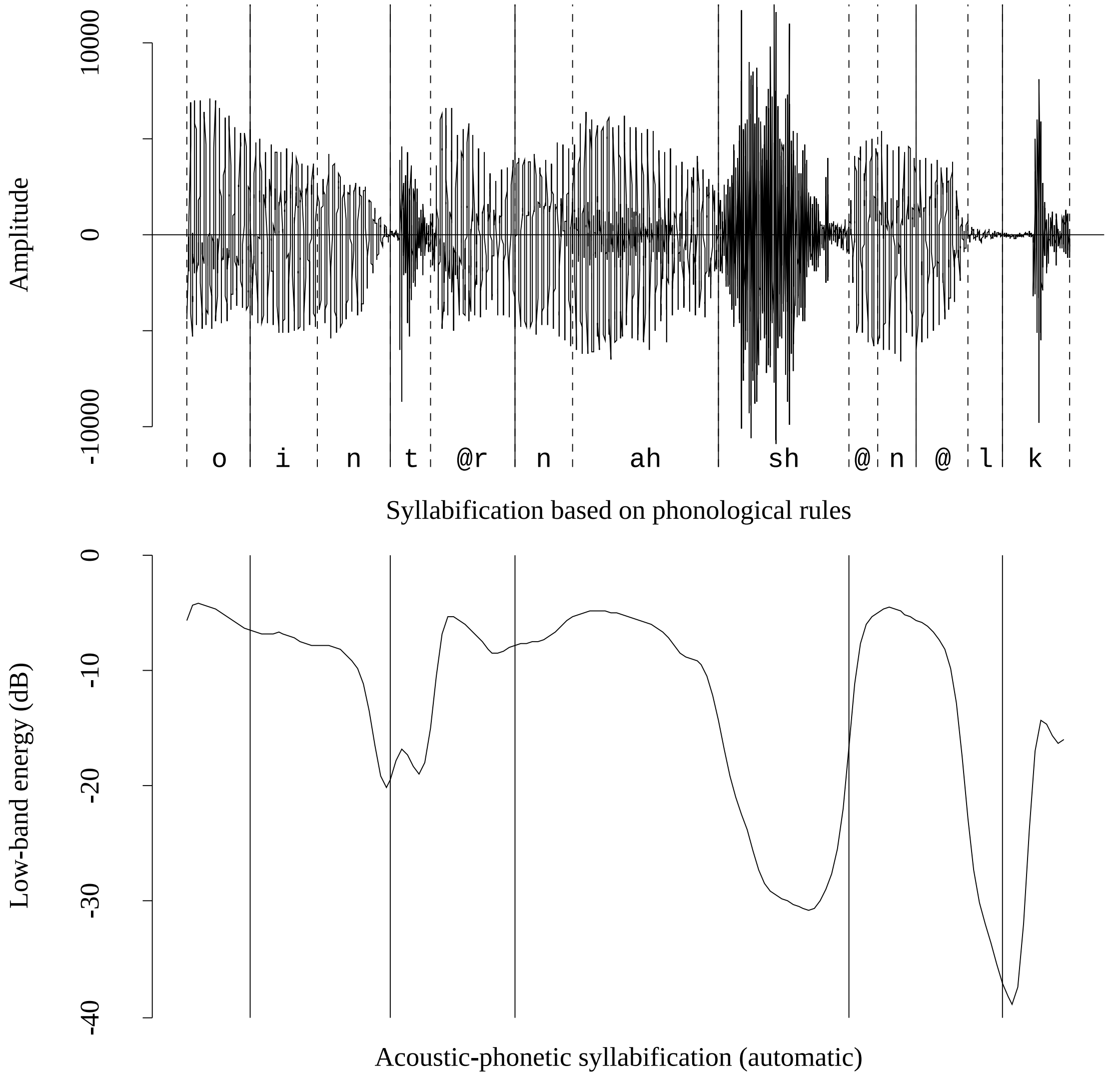

Figure 8.1: Syllabification of the word *international**

or syllabic consonant (a potential nucleus) in the hand-syllable. Then consider each auto-syllable from the beginning of the utterance to determine whether or not it overlaps with the potential nucleus. If the auto-syllable does overlap and if it has previously been assumed to be an *extra* syllable, then mark the auto-syllable as not being an *extra* syllable, mark the hand-syllable as not *missing*, and consider the next hand-syllable in the utterance. Otherwise, repeat the search for a matching auto-syllable using another

Total number of syllables				
from a phonological basis (by hand)	from an acoustic-phonetic basis (automatic)	Match	Missing	Extra
9322 (100.0%)	8910 (95.6%)	8875 (95.2%)	-447 (4.8%)	+35 (0.4%)

Table 8.1: Comparison of syllabifications based on phonological rules (by hand) and on acoustic-phonetic parameters (automatic)

potential nucleus in the hand-syllable until all possible nuclei have been considered. If none of the possible nuclei in a hand-syllable are overlapped by an auto-syllable previously assumed to be an *extra* syllable, then the hand-syllable is genuinely *missing* in the automatic syllabification. Once every hand-syllable in the utterance has been considered in turn, any remaining auto-syllables which have not been found to correspond with a hand-syllable are genuinely *extra* syllables.

There is a large level of agreement between the syllabic domains generated automatically and by the procedural definition based on phonological rules (see Table 8.1). The *missing* syllable boundaries are due to the occurrences of vowel/vowel boundaries between which there is no valley in the low-band energy. When this case arises, often one of the vowels is a schwa; for example, the phonological syllabification of *my address* as /m aɪ – ə – d r ɛ s/ can be grouped on an acoustic-phonetic basis as [m aɪ ə – d r ɛ s]. Conversely, *extra* syllable boundaries occur when the low-band energy dips within the phonologically based syllable at a vowel/vowel boundary or vowel/syllabic consonant boundary; for example, the phones in *tour* /t ʊ ɚ/ can be grouped as [t ʊ – ɚ] on an acoustic-phonetic basis, and the phones in the word *forms* /f ɔ r m s/ are grouped as [f ɔ – r m̩ s] when its phonetic realisation tends towards the pronunciation of *forums* with schwa deletion, but with a fall in low-band energy remaining between /ɔ/ and /m/.

The algorithm described in Section 8.1.1 forms groups of phones with a vowel or syllabic consonant as the nucleus of each group. There are no vowel/vowel pairs or vowel/syllabic consonant pairs with dips in low-band energy between them, in any of the groups. The boundaries between groups are positioned at the point of minimum low-band energy between nuclei (aligned to the nearest phone boundary). It is assumed that such groups of phones can only ever be perceived as single prominent units in connected speech. These units are shown to correlated closely with syllables defined using phonologically based rules. They are, therefore, referred to as *syllables*.

8.2 Processing of *FØ* for an acoustic-phonetic description of intonation

The fundamental frequency contour of a speech waveform does not form an acoustic-phonetic representation of the utterance intonation because microprosodic variations are also present and the contour is dependent upon a speaker's fundamental frequency range. The process of piece-wise linear stylisation of a contour presented in Section 8.2.1 aims to prevent microprosodic variations from being confused as pitch accents. A stylised contour is processed with respect to the syllables of an utterance in Section 8.2.2 to form a schematic acoustic-phonetic representation of intonation. The schematic representation is independent of a speaker's fundamental frequency range.

8.2.1 Linear piece-wise stylisation of an *FØ* contour

The algorithm used to perform the stylisation of an *FØ* contour is based on the technique proposed by Scheffers (1988). As with other *FØ* contour stylisation techniques (Hirst & Espesser, 1993; Taylor, 1993), the *FØ* contour is the only input parameter to Scheffers' algorithm. Information about the segmental content of an utterance is not included. The algorithm described here to stylise an *FØ* contour also uses syllable nucleus boundaries (which are automatically located using the acoustic-phonetic syllabification algorithm described in Section 8.1.1).

Scheffers partitions an *FØ* contour into linear piece-wise sections using the least (mean of) squares method of regression analysis. The least (mean of) squares method of regression analysis produces its best results when the underlying error distribution is Gaussian. This method of regression analysis becomes unreliable if the data contains spurious *outliers* (samples which deviate from the local trend by a large amount). Such data almost invariably exists amongst *FØ* contours, where gross measurement errors and microprosodic variations can, and often do, occur. A more robust method of regression analysis is therefore required. The least median of squares (LMedS) method of regression analysis (Appendix A) is more robust to the presence of outliers (Rousseeuw & Leroy, 1987). The algorithm described here to stylise an *FØ* contour incorporates the robust least median of squared residuals regression analysis.

An account is given below which describes how *significant* turning-points are located in sections of an *FØ* contour (partitioned using syllable nucleus boundaries) then modified to prevent contour discontinuities other than at the boundaries between unvoiced and voiced speech, and how a new, stylised contour is generated by interpolating between the turning-points.

The *FØ* values describing a contour (excluding values which equal zero to represent unvoiced speech) are initially converted to the semitone scale using Equation 8.1.

$$F\O_{semitone} \;=\; 12.0 \log_2 \left(\frac{F\O_{Hz}}{55.0} \right) \tag{8.1}$$

An *FØ* contour is partitioned into sections which represent portions of continuously voiced speech and which overlap any part of at least one syllable nucleus. Partitioning an *FØ* contour in this way eliminates isolated voiced sections of the contour which do not overlap a syllable nucleus from the subsequent processing (which locates significant turning-points in the contour). This is required because isolated voiced sections in the *FØ* contour which do not overlap a syllable nucleus are likely to be unreliable since they may correspond to unvoiced sections of speech erroneously classified as voiced by an FDA.

Short sections (less than 40ms in length) of an *FØ* contour partitioned as above correspond to short syllable nuclei with unvoiced left and right segmental contexts. Thus, microprosodic variations at the onset and offset of voicing dominate the *FØ* trajectory in such short sections. Significant turning-points are therefore not located in these short sections of a contour. They are stylised by setting the *FØ* values to the mean value of the *FØ* trajectory within each short section.

The following process (adapted from Scheffers, 1988) is used to identify the turning-points in each section of an *FØ* contour which is longer than 40ms. Starting with the first voiced frame, LMedS regression analysis is applied to a window of w frames corresponding to voiced speech, where w is initially set to 5. The final frame in this window is taken to be a turning-point candidate. The *FØ* value of the subsequent frame is predicted using the coefficients of the LMedS regression analysis. If the absolute difference between the actual and predicted *FØ* values is less than or equal to some level of permitted variation in *FØ* (1 semitone), then the candidate is not a turning-point, the window length w is incremented to include the next voiced frame, and the above process is repeated. The repetition of this process terminates when the turning-point candidate is the final voiced frame in the section of the *FØ* contour under analysis. Otherwise, when the absolute difference is greater than the permitted *FØ* variation, either this subsequent *FØ* value constitutes some type of irregularity in the *FØ* contour or the candidate could be a genuine turning-point. In order to determine which is the case, the *FØ* value of the next voiced frame is also predicted. If the absolute difference between the predicted value and the actual value is once again greater than the permitted *FØ* variation, and if this situation arises for all following frames up to either the final voiced frame in the section of the *FØ* contour under analysis or such that the duration of this discontinuity is greater than some minimum permitted level (100ms), which ever occurs first, then the candidate is said to be a genuine turning-point. Otherwise, the length of the window w is increased to include the first frame for which the absolute difference in the actual and the predicted *FØ* values was less than or equal to the permitted variation, but not to include those for which it was greater than the permitted variation, and the LMedS regression analysis process is repeated. If the candidate was found to be a turning-point and if it corresponds to a voiced frame immediately preceding a frame of unvoiced speech, then the first frame of the next voiced region is also designated as a turning point. This entire process is then repeated with the length of the window w reset to 5 and with the first frame of the window set to the frame of the most recent turning-point found. The first and final voiced frames of the non-stylised contour are also assigned as turning-points.

In order to ensure that discontinuities in the stylised *FØ* contour only occur at unvoiced sections of speech, the fundamental frequency at each turning-point of the new contour is determined in a way which depends upon the voicing state of the frames adjacent to it. However, a discontinuity in the stylised *FØ* contour is allowed within a voiced section of speech if the turning-point is an *outlier* for either of the piece-wise sections it joins. (The detection of outliers is a part of the LMedS regression analysis.) In such situations, the turning-point is treated as being adjacent to an unvoiced frame of speech. For any given turning-point (tp) at frame f_{tp} with original fundamental frequency $F\O_{tp}$, the LMedS coefficients s_{tp} (slope) and i_{tp} (intercept) of selected points preceding the turning-point are known. The modified fundamental frequency $F\O'_{tp}$ is calculated from Equation 8.2.

$$F\O'_{tp} = \begin{cases} 0.5(s_{tp}.f_{tp} + i_{tp} + s_{tp+1}.f_{tp} + i_{tp+1}) & \text{if frames } f_{tp}-1 \text{ \& } f_{tp}+1 \text{ are voiced} \\ s_{tp+1}.f_{tp} + i_{tp+1} & \text{if frame } f_{tp}-1 \text{ is unvoiced \& frame } f_{tp}+1 \text{ is voiced} \\ s_{tp}.f_{tp} + i_{tp} & \text{if frame } f_{tp}-1 \text{ is voiced \& frame } f_{tp}+1 \text{ is unvoiced} \\ F\O_{tp} & \text{if frames } f_{tp}-1 \text{ \& } f_{tp}+1 \text{ are unvoiced} \end{cases} \tag{8.2}$$

The new stylised contour is then created by linear interpolation of $F\O$ between each turning-point (f_{tp}, $F\O'_{tp}$) and by resetting each frame that is unvoiced in the non-stylised contour to an unvoiced state in the new contour. The resultant data is then coverted back to a Hertz scale.

An example of a linear piece-wise stylised contour produced using the above algorithm is illustrated in Figure 8.2. The stylised $F\O$ contour is aligned with the speech waveform and phonetic transcription. The dotted lines across the speech waveform represent phone boundaries and the solid lines represent syllable boundaries, which are used by the stylisation algorithm. Many of the microprosodic $F\O$ variations and $F\O$ perturbations are removed by the stylisation process, however, some remain in sections of the contour associated with voiced consonants in a vocalic context. The overall melody of the utterance is retained.

8.2.2 Prosodic schematisation of a stylised *FØ* contour

Each piece-wise section in the stylised $F\O$ contour forms a possible pitch accent or a possible part of an accent. Some piece-wise sections do not correspond to part of any pitch accent, such as those which are a direct consequence of erroneous $F\O$ extraction or stylisation. Thus, a piece-wise section is treated as being part of a possible pitch accent if either at least 50% of the piece-wise section overlaps a syllable nucleus (where $F\O$ estimation is expected to be reliable) or at least 50% of a syllable nucleus is over-

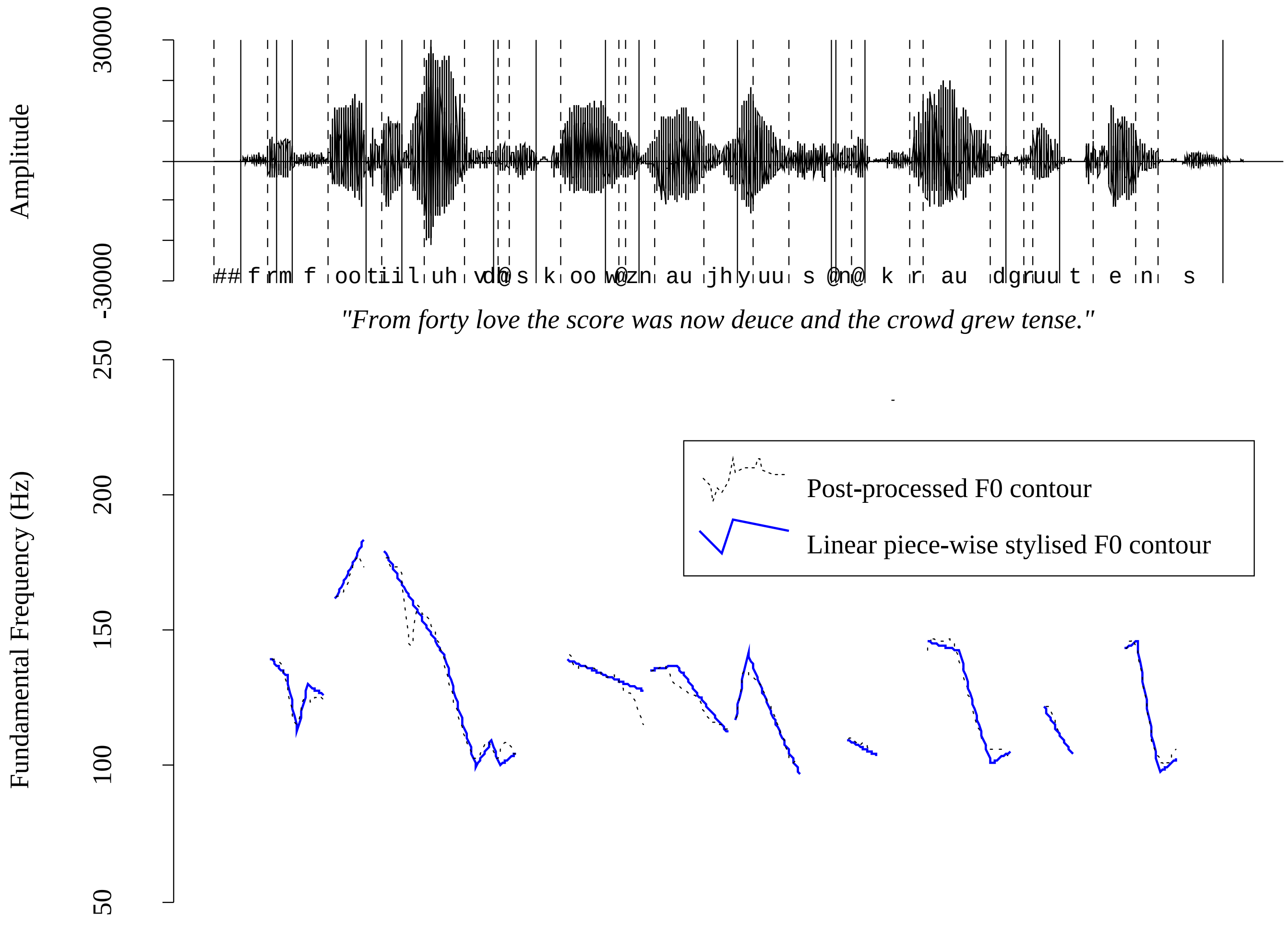

Figure 8.2: Linear piece-wise stylisation of an $F\emptyset$ contour*

lapped by the piece-wise section. The piece-wise sections may therefore extend beyond a syllable nucleus but only those crossing a syllable nucleus by a substantial amount are selected. This approach compromises between using information about the trajectory of *FØ* through vowels alone (which may be limiting for short nuclei) and using the *FØ* contour of an entire syllable (where *FØ* discontinuity errors may occur).

The absolute *FØ* range in an utterance will vary from speaker to speaker and from utterance to utterance (Section 3.4). *FØ* piece-wise sections are, therefore, normalised for each utterance to give relative *FØ* heights. The relative height of each piece-wise section is calculated by first locating a regression line which best fits the contour turning-points using LMedS analysis. A by-product of the LMedS is the standard deviation σ_{LMedS} of the points from the resultant linear model. The absolute *FØ* at each turning-point is then converted by subtracting its modelled value and dividing by the standard deviation, σ_{LMedS}. This effectively compensates for any long term declinative tendency that may be exhibited in the fundamental frequency contour and expresses the *FØ* values relative to an utterance dependent datum.

Once the relative height of each piece-wise section has been established, the piece-wise sections are combined to form *FØ* trajectory descriptors. The *FØ* trajectory descriptors used are level, fall, rise, fall-rise and rise-fall "$-$, $\backslash$, $/$, $\vee$, $\wedge$". Each piece-wise section which crosses a substantial part of a syllable nucleus (as described above) is classified as either level, a fall, or a rise. If there is no piece-wise section which crosses a particular syllable nucleus by a substantial amount, then the syllable is classified as having an "unknown" *FØ* trajectory. Let $F\O_{start}$ represent the relative *FØ* height at the start of the piece-wise section and that at the end of the section be represented by $F\O_{end}$. A piece-wise section is classified on the basis of Equation 8.3.

$$F\O \text{ trajectory} = \begin{cases} \backslash & \text{if } F\O_{start} - F\O_{end} > 0.75\sigma_{LMedS} \\ / & \text{if } F\O_{start} - F\O_{end} < -0.75\sigma_{LMedS} \\ - & \text{otherwise} \end{cases} \tag{8.3}$$

When two or more piece-wise sections cross any particular nucleus, they are combined by initially taking all adjacent sections with the same *FØ* trajectory descriptor and joining them into one. A join is made by setting $F\O_{start}$ to that of the first section, $F\O_{end}$ to that of the second section, and reclassifying using Equation 8.3. In the combined training data and test data of 660 utterances (described in Section 4.1) consisting of 8546 syllables, there are only six syllables for which more than two piece-wise sections remain after this process. One of these six syllables contains a section of *FØ* halving errors, two syllables contain extreme *FØ* perturbations due to creaky voice, and three syllables include microprosodic variations which are not successfully removed by the stylisation process. If there are two remaining sections (their classifications must differ) and if either is classified as level "–", then they too are joined in the same way. Otherwise, one section is a fall "\" and the other section is a rise "/". These sections are combined to give a single trajectory which is described as either a fall-rise "∨" or a rise-fall "∧" depending on their order, and the relative level at their mid-point is kept for reference. Thus, for the fall-rise and rise-fall descriptors, the relative heights of the onset, the mid-point and the offset of the trajectory are known.

FØ schemata are generated from sequencies of *FØ* trajectory descriptors and their internalised relative heights. An example of the processing of a speech waveform to generate an *FØ* schema is illustrated in Figure 8.3. The speech waveform for the sentence, "From forty love the score was now deuce and the crowd grew tense," is shown with its aligned phonetic transcription represented by MRPA symbols (Appendix D). The raw *FØ* contour is extracted using the super resolution *FØ* determinator (*e*SRFD) and includes *FØ* doubling and halving errors. These errors are eliminated by the de-step filter and non-linear smoothing algorithm (Section 7.3). The linear piece-wise stylisation of the post-processed *FØ* contour removes many of the microprosodic *FØ* variations. The *FØ* schema derived from this exhibits only the intonational component of a raw *FØ* contour.

Judgements of syllable accentuation are based on the *FØ* trajectory descriptors using the pitch accent decision filter developed by Hieronymus (1989) — illustrated in Figure 3.5. The decision filter examines three *FØ* trajectory descriptors for each syllable — the first syllable *FØ* trajectory descriptor which is not classified as "unknown" on the left-hand side of the syllable, the *FØ* trajectory descriptor of the syllable, and the first syllable *FØ* trajectory descriptor which is not classifier as "unknown" on the right-hand

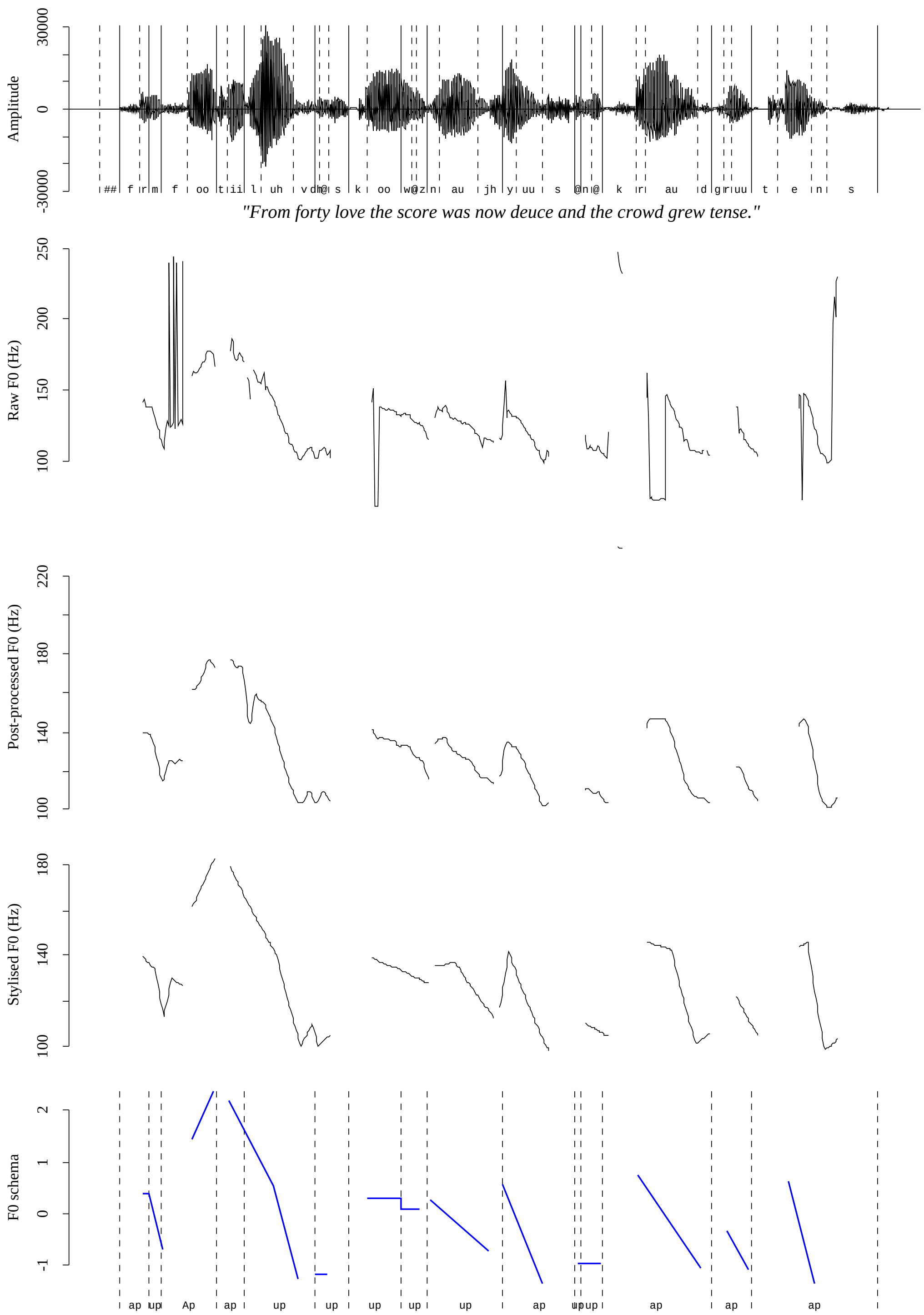

Figure 8.3: Transformation from a raw $F\emptyset$ contour to a schematic acoustic-phonetic representation of intonation*

side of the syllable. If the *FØ* trajectory of a syllable is either a fall-rise "$\vee$" or a rise-fall "$\wedge$" then the syllable is classified as being accented. Syllables with an "unknown" *FØ* trajectory are classified as being unaccented. If the *FØ* trajectory descriptor of the left-hand syllable context or the right-hand syllable context is either a fall-rise "$\vee$" or a rise-fall "$\wedge$", then only the half of the fall-rise or rise-fall closest to the syllable is used by the decision filter.

Syllables classified as accented by the decision filter are labelled as "*ap*". A syllable is also classified as accented if it corresponds with the maximum relative *FØ* height of an *FØ* schema. Syllables corresponding to such maxima are labelled "*Ap*". All other syllables are labelled "*up*". The *FØ* schema in Figure 8.3 is shown with syllable boundaries and the syllable accentuation labels produced by this method.

The levels of agreement between the categories of syllable accentuation as transcribed by hand and as assigned automatically are shown in Table 8.2. Syllables which are transcribed by hand as stressed but unaccented "*s*" or unstressed "*u*", are collectively referred to as unaccented syllables "*UA*". The syllables which are transcribed by hand as pitch accented "*PA*" do not include non-stress accents. (Refer to Section 2.2.4 for a clarification of these terms.) The pitch accent decision filter does not include the location of prominent syllables as an input parameter. Syllables transcribed as accented "*ap*" by the decision filter therefore include non-stress accents. This accounts for the large number of unaccented syllables "*UA*" which are confused as pitch accented "*ap*" by the decision filter.

8.3 Combination of acoustic parameters

The unidimensional contribution to the perceived prominence of a syllable made by features of duration and energy is investigated in Chapters 4 & 5. Sections 4.5 & 5.4 show that the duration feature $D_{feature}$ (Equation 4.6) and the energy feature $E_{feature}$ (Equation 5.3) can be used in conjunction with a simple peak-picking algorithm to label syllables as either prominent or non-prominence and respectively yield a 77.8% and an 84.0% agreement level with prominence labels based on human perception (in an open test).

Section 8.2.2 shows that the trajectories of an *FØ* schema can be used in conjunction

		Pitch accent algorithm label			
		Ap	*ap*	*up*	total
Hand Label	*PA*	359 (6.7%)	992 (18.7%)	486 (9.1%)	1837 (34.5%)
	UA	101 (1.9%)	864 (16.2%)	2517 (47.3%)	3482 (65.5%)
	total	460 (8.6%)	1856 (32.0%)	3003 (56.5%)	5319 (100.0%)

Correct classification rate = 3868/5319 (72.7%)

Table 8.2: *FØ* schema: Confusion matrix for pitch accent decision filter (*Ap* — syllable with maximum *FØ* in schema; *ap* — accented syllable according to decision filter; *up* — unaccented syllable according to decision filter; *PA* — pitch accented syllable; *UA* — unaccented syllable)

with a pitch accent decision filter to label syllables as either accented or unaccented and yield a 72.7% agreement level with syllable accentuation labels based on human perception. The accented syllables identified by the pitch accent decision filter include non-stress accents.

Syllable classification labels are generated by the peak-picking algorithm for the duration and energy features, and by the pitch accent decision filter for the trajectories of the *FØ* schema. These labels are combined to classify each syllable in an utterance as either pitch accented "*PA*", stressed but unaccented "*s*", or unstressed "*u*". These labels are assigned using a procedure adapted from the equal-weight two-out-of-three voting scheme proposed by Hieronymus (1989).

A syllable is categorised as prominent if it is associated with an "*Sd*" label (maximum $D_{feature}$), an "*Se*" label (maximum $E_{feature}$) or an "*Ap*" label (maximum in *FØ* schema). A syllable is also categories as prominent if it is associated with at least two-out-of-three of the labels "*sd*" (local maximum in $D_{feature}$ contour), "*se*" (local maximum in $E_{feature}$ contour) and "*ap*" (accented syllable according to the pitch accent decision filter). A syllable which is categorised as prominent by these rules is labelled as being pitch accented "*PA*" if it is associated with either an "*Ap*" or an "*ap*" label; otherwise it is labelled as being stressed but unaccented "*s*". All other syllables are categorised as unstressed and are labelled "*u*".

Figure 8.4 shows the confusions between perceived prominence labels (transcribed

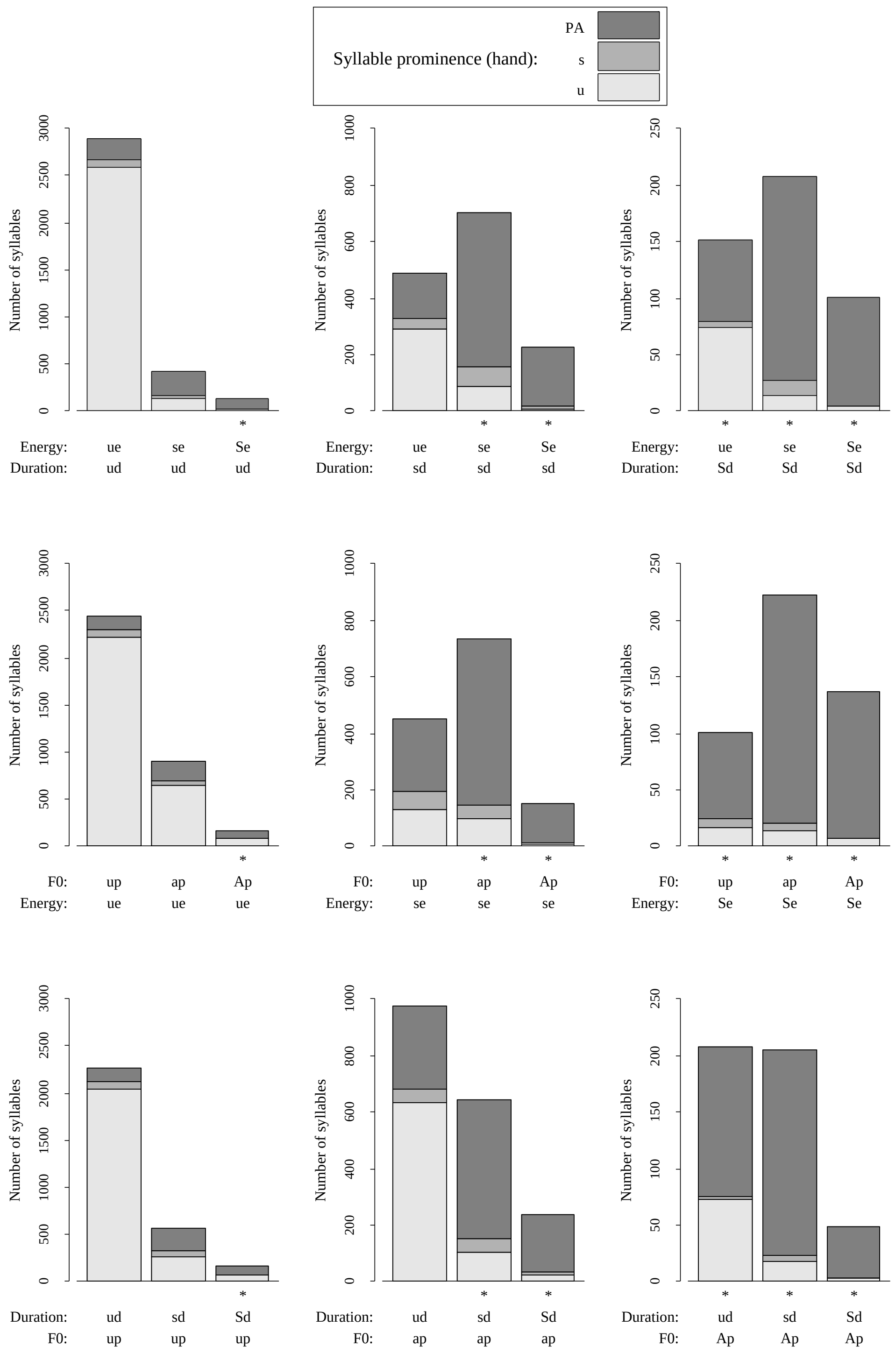

Figure 8.4: Multi-way confusions across acoustic parameters

by hand) and each possible pair of labels generated by the processing of the acoustic parameters. The combinations of labels which are marked by an asterisk are those which result in a syllable being labelled as prominent (both accented and unaccented) by the procedure described above. There are more prominent syllables than non-prominent syllables (transcribed by hand) associated with the combinations of automatically assigned labels which are marked by asterisks.

8.4 Overview of the integrated prosodic analysis system

The prosodic analysis of speech described in this thesis involves a number of intergated modules. An overview of the automatic prosodic analysis system is shown in Figure 8.5. There are three underlying channels of analysis — one channel for each acoustic parameter (energy, duration and fundamental frequency). A direct flow of information through these channels from the input speech signal to the combination procedure is interrupted by the syllabification module. The syllabic domain ties together the information embedded within the three acoustic parameters. The syllabification module incorporates information about the low-band energy contour and the segmentation of an utterance. This information is passed on in an abstracted form to the normalisation of energy and duration measures and to the processing of an *FØ* contour.

A summary of the modules which make up this system to form a procedural analysis of an utterance is as follows.

Front-end signal processing:

- Transform the speech signal from the time domain to a frequency domain representation by applying an FFT to 20ms frames of data at 5ms intervals.

- Extract the (frame-level) low-band energy contour from the frequency domain representation of the speech signal.

- Cepstral coefficients are required for the automatic segmentation of speech into acoustic-phonetic units. The cepstral analysis is performed on 20ms frames of data at 5ms intervals.

- Segment the speech waveform into acoustic-phonetic units either by hand (for the

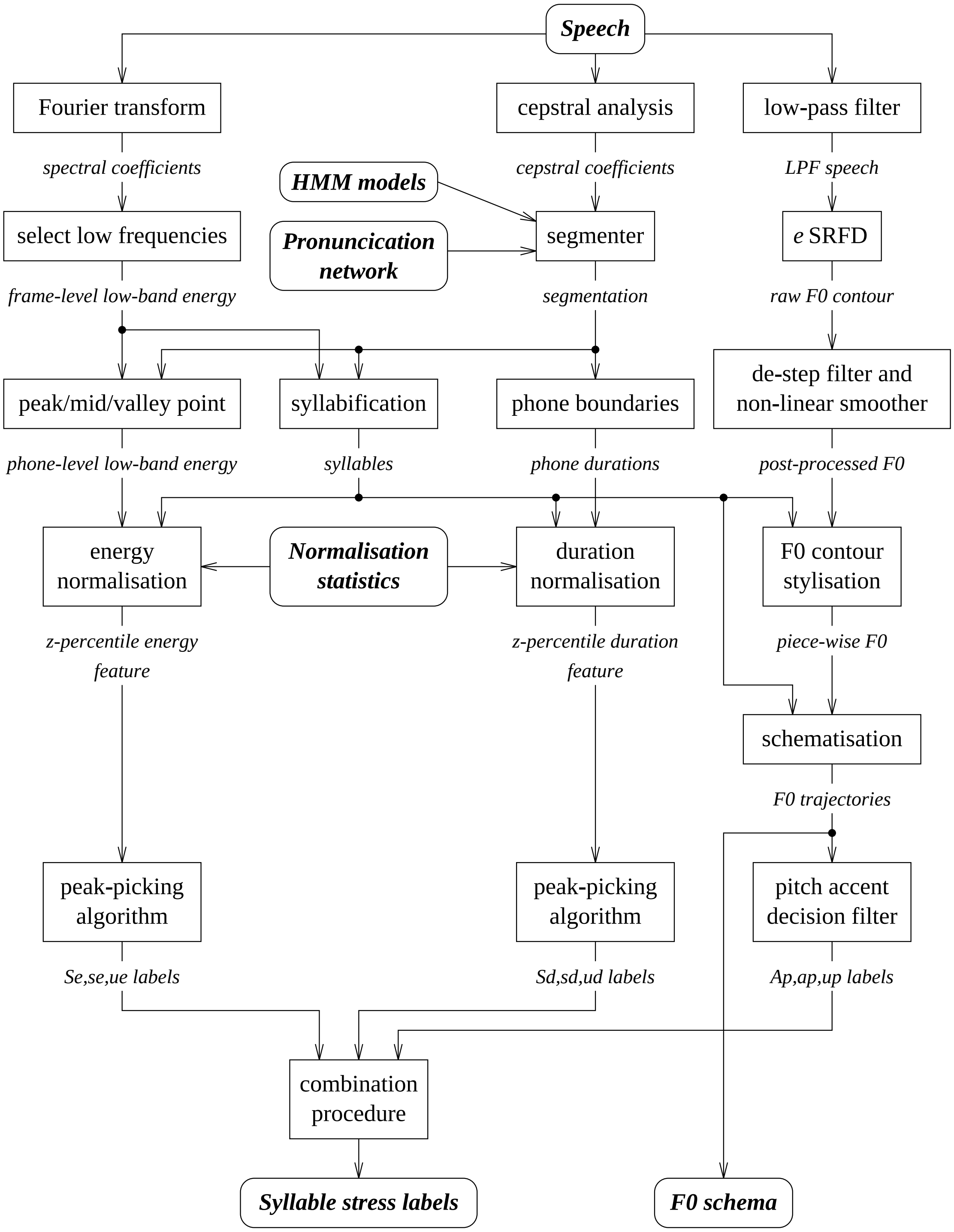

Figure 8.5: Block diagram of integrated prosodic analysis system

training data) or by an automatic segmentation system (McInnes *et al.*, 1992). This provides the phone boundary and label information of an utterance.

- Low-pass filter the speech signal using a finite impulse response (FIR) filter with a -3dB cut-off at 600Hz and rejection greater than -85dB above 700Hz.
- Extract the *FØ* contour from the low-pass filtered speech waveform using the enhanced super resolution *FØ* determination (*e*SRFD) algorithm. *FØ* values are generated at 5ms intervals.
- Apply the de-step filter and non-linear smoothing to the raw *FØ* contour in order to eliminate *FØ* doubling and halving errors and reduce *FØ* cycle-to-cycle perturbations.

Phone-level abstraction of acoustic parameters:

- Calculate the phone-level low-band energy contour from the frame-level low-band energy contour and the segmentation. A peak/valley/mid-point is selected in the energy contour for each phone.
- Obtain an acoustic-phonetic syllabification of the speech from the frame-level low-band energy contour and the segmentation. The syllabification identifies syllable nuclei and provides information about phone grouping.
- Absolute phone durations are obtained from the distance between phone boundaries in the segmentation.

Syllable-level abstraction:

- Calculate the normalised phone-level low-band energy feature $E_{feature}$ (Equation 5.3) for each syllable nucleus. The variations in $E_{feature}$ from syllable to syllable form a normalised energy contour.
- Calculate the normalised phone duration feature $D_{feature}$ (Equation 4.6) for each syllable lhyme. The variations in $D_{feature}$ from syllable to syllable form a normalised duration contour.

- Stylise the post-processed *FØ* contour into a sequence of linear piece-wise sections. The stylisation process makes use of syllable nucleus boundaries and reduces microprosodic variations in the contour.

Sentential stress and intonation representations:

- Schematise the piece-wise stylised *FØ* contour to identify *FØ* trajectories related to pitch accents. The *FØ* schema provides an acoustic-phonetic representation of the utterance intonation.
- Apply the pitch accent decision filter to the *FØ* schema in order to locate accented syllables.
- Apply the peak-picking algorithm to the normalised energy contour. This represents the unidimensional contribution made by energy to the prominence of a syllable.
- Apply the peak-picking algorithm to the normalised duration contour. This represents the unidimensional contribution made by duration to the prominence of a syllable.
- Finally, combine the syllable accentuation, energy and duration labels to categorise each syllable as either pitch accented "*PA*", stressed but unaccented "*s*", or unstressed "*u*".

8.5 Performance evaluation of prosodic analysis algorithms

The syllable prominence labels in the training and test data described in Section 4.1 are transcribed on the basis of human perception. This transcription, however, cannot be regarded as definitive. In order to give an objective evaluation of the integrated prosodic analysis system proposed in this thesis (referred to as the *PCB-system*), the performance of the system is compared with that of two algorithms formerly proposed in the literature, relative to the transcription based on human perception. The two former algorithms which are evaluated, are the algorithm based on a Bayesian classifier (Waibel, 1988) (referred to as the *AW-system*) and the algorithm which uses knowledge-based rules to automatically transcribe syllable prominence levels in connected speech

(Hieronymus, 1989; Hieronymus & Williams, 1991) (referred to as the *JLH-system*). The AW-system and the JLH-system are reviewed in Section 3.5.

Although the AW-system and the JLH-system have been evaluated by their respective developers, it is not possible to draw comparisons of their reported performances. Each evaluation depends on the form of the prosodic transcription which is taken as a reference and on the nature of the test speech and the data used to train the algorithms. In the evaluations presented here, the algorithms are both trained and tested using the data described in Section 4.1. A direct comparison can therefore be made between the performance evaluations of the prosodic analysis system proposed in this thesis and of the two former analysis techniques. The statistical significance of the differences in performance is determined.

8.5.1 Integrated prosodic analysis (PCB-system)

The normalisation statistics used to calculate the duration and energy features are determined from the training data. The statistics are determined from two inspections of the training data.

The first inspection determines the percentage p of non-prominent phones, the distributions of phone durations in which phones are grouped on a broad phonetic basis (Table 4.1), and the distributions of phone-level low-band energies in which phones are grouped on a fine phonetic basis. The p'th percentile φ^p and the standard deviation σ are calculated for each of the distributions. These values are shown in Tables 8.3 (duration) & 8.4 (energy).

The second inspection of the training data determines the distributions of the sum $z_{percentile}$ duration measures of all phones in the lhyme (Equation 4.7) of prominent and non-prominent syllables, for each unique number of phones that are found in a syllable lhyme. The calculation of the $z_{percentile}$ duration measures uses the statistics derived from the first inspection of the training data. The mean value is calculated for each of the distributions. These values are shown in the lower part of Table 8.3.

The integrated prosodic analysis system is used to assign sentential stress categories to the syllables in the test data (an open test). These categories are compared with those transcribed by hand (see Table 8.5). The transcriptions are in agreement for 79.1% of

Broad phonetic class	$\varphi_T^p(broad\_phone\_class)$ (ms)	$\sigma_T(broad\_phone\_class)$ (ms)
reduced monophthong	126.19	19.90
short monophthong	78.40	35.89
long monophthong	109.90	47.65
diphthong	126.79	48.58
sonorant	73.53	24.58
voiced obstruent	91.15	28.85
unvoiced obstruent	117.40	38.37

No. phones in lhyme, i	$\overline{M_s}(i)$	$\overline{M_u}(i)$
1	0.865	-1.483
2	0.321	-3.312
3	-1.043	-4.061
4	-3.172	-6.715

Table 8.3: Duration feature normalisation statistics

the syllables.

8.5.2 Bayesian classifier (AW-system)

Each syllable (derived using the acoustic-phonetic syllabification algorithm presented in Section 8.1.1) is characterised by four features:

- The integral of the low-band energy over the syllable nucleus (*ptpint*).
- The duration of the syllable nucleus (*sondur*).
- The maximum *FØ* in the entire syllable (*F0max*).
- The average Euclidean distance of frame-to-frame log-power spectra over the central half of the syllable nucleus (*spchave*).

A Bayesian classifier (Appendix B) is trained using the training data to model three categories of sentential stress — pitch accented "*PA*", stressed but unaccented "*s*", and unstressed "*u*". The Bayesian classifier is used to assign sentential stress categories to syllables in the test data (an open test). These categories are compared with those

Fine phonetic class	$\varphi_E^p(fine\_phone\_class)$ (dB)	$\sigma_E(fine\_phone\_class)$ (dB)
[ɪ]	-5.885	5.160
[i]	-7.768	3.898
[ʊ]	-7.195	3.374
[u]	-9.406	3.479
[ʊə]	-7.429	2.844
[ɪə]	-7.010	2.829
[eɪ]	-5.757	3.136
[əʊ]	-5.075	3.064
[ɔɪ]	-7.500	3.419
[ə]	-1.633	4.862
[ɔ]	-6.814	3.380
[aɪ]	-4.026	3.250
[ɛ]	-6.029	3.232
[ɛə]	-5.320	2.708
[aʊ]	-4.474	3.455
[ʌ]	-4.673	3.063
[ɜ]	-4.722	2.799
[ɒ]	-4.869	3.284
[a]	-5.171	2.644
[ɑ]	-4.788	3.908
[l]	-0.615	4.794
[r]	-0.390	5.738
[w]	-5.098	6.173
[j]	-6.702	3.780
[n]	-6.700	4.865
[m]	-5.801	4.450
[ŋ]	-8.446	4.290
[b]	-10.375	10.271
[ʒ]	-14.526	9.137
[v]	-9.645	6.205
[g]	-14.094	8.161
[ʤ]	-10.095	7.426
[ð]	-5.174	8.643
[d]	-11.093	8.686
[z]	-10.612	6.581
[s]	-19.290	4.813
[θ]	-12.988	6.254
[t]	-5.275	11.458
[ʧ]	-19.554	9.021
[f]	-16.002	4.676
[ʃ]	-19.851	4.494
[p]	-15.663	11.113
[k]	-10.607	10.747
[h]	-11.507	7.013

Table 8.4: Energy feature normalisation statistics

		PCB-system label			
		PA	s	u	total
Hand Label	PA	1264 (23.8%)	293 (5.5%)	280 (5.3%)	1837 (34.5%)
	s	78 (1.5%)	55 (1.0%)	128 (2.4%)	261 (4.9%)
	u	234 (4.4%)	100 (1.9%)	2887 (54.3%)	3221 (60.6%)
	total	1576 (29.6%)	448 (8.4%)	3295 (61.9%)	5319 (100.0%)

Correct classification rate = 4206/5319 (79.1%)

Table 8.5: PCB-system: Confusion matrix of prosodic transcription
(PA — pitch accented; s — stressed but unaccented; u — unstressed)

		AW-system label			
		PA	s	u	total
Hand Label	PA	938 (17.6%)	0 (0.0%)	899 (16.9%)	1837 (34.5%)
	s	99 (1.9%)	0 (0.0%)	162 (3.0%)	261 (4.9%)
	u	384 (7.2%)	0 (0.0%)	2837 (53.3%)	3221 (60.6%)
	total	1421 (26.7%)	0 (0.0%)	3898 (73.3%)	5319 (100.0%)

Correct classification rate = 3775/5319 (71.0%)
(Entropy score = 0.956)

Table 8.6: AW-system: Confusion matrix of prosodic transcription
(PA — pitch accented; s — stressed but unaccented; u — unstressed)

transcribed by hand (see Table 8.6). The transcriptions are in agreement for 71.0% of the syllables. No syllables are labelled as stressed but unaccented "s" by the Bayesian classifier, and syllables labelled as prominent ("PA" or "s") by hand are labelled as non-prominent ("u") more often than they are labelled as prominent by the Bayesian classifier.

8.5.3 Knowledge-based rules approach (JLH-system)

The duration and energy thresholds used in the knowledge-based rules approach proposed by Hieronymus (1989; 1991) must first be determined from the training data. To do this, the vowels are classified on a broad phonetic basis as either reduced [ə], short [ɪ, ɨ, ɛ, a, ʌ, ʊ, ɒ, æ], long [i, ɜ, u, ɔ, ɑ, ʉ, ɚ] or a diphthong [eɪ, aɪ, ɔɪ, əʊ, aʊ, ɪə, ɛə, ʊə]. For each vowel-type, the distribution of durations is calculated with a fixed adjustment factor (0.6) being applied to all pre-pausal vowels. The percentage $p(vowel\_type)$ of vowels transcribed by hand as unstressed is determined for each vowel-type. The duration threshold for each vowel-type is given by the $p(vowel\_type)$'th percentile of the corresponding distribution. A distribution of maximum intra-vowel low-band energies is calculated across all vowels. The energy threshold, above which the intensity is regarded as contributing to the prominence of a vowel, is given by the p_{all}'th percentile of the all-vowel-types energy distribution, where p_{all} is the percentage of all vowels which are hand transcribed as unstressed.

Although the knowledge-based rules approach proposed by Hieronymus makes no distinction between reduced and short vowels in assigning a level of prominence, the reduced vowels are separated from the category of short vowels in determining the duration thresholds. This is done because all the reduced vowels in the training data are transcribed as unstressed and would offset the duration threshold if included.

The distributions of pre-pausal compensated durations and maximum intra-vowel low-band energies for the training data are shown in Figure 8.6. The duration and energy thresholds derived from these distributions are indicated by the dotted lines. The -20.0dB energy threshold is used by the knowledge-based rules approach to annotate all low intensity vowels as unstressed, regardless of other acoustic parameters.

The knowledge-based rules approach is used to assign sentential stress categories to the vowels in the test data (an open test). These categories are compared with those transcribed by hand (see Table 8.7). The transcriptions are in agreement for 72.6% of the syllables.

8.5.4 Comparison

It is important to determine if the differences in performance of these algorithms is statistically significant before any claims can be made about their relative efficacy. The McNemar test (Gillick & Cox, 1989) is used to decide whether or not the difference in the classification error rates between any two algorithms is statistically significant. The null hypothesis H_0 to test is that the true (but unknown) error rates of two algorithms

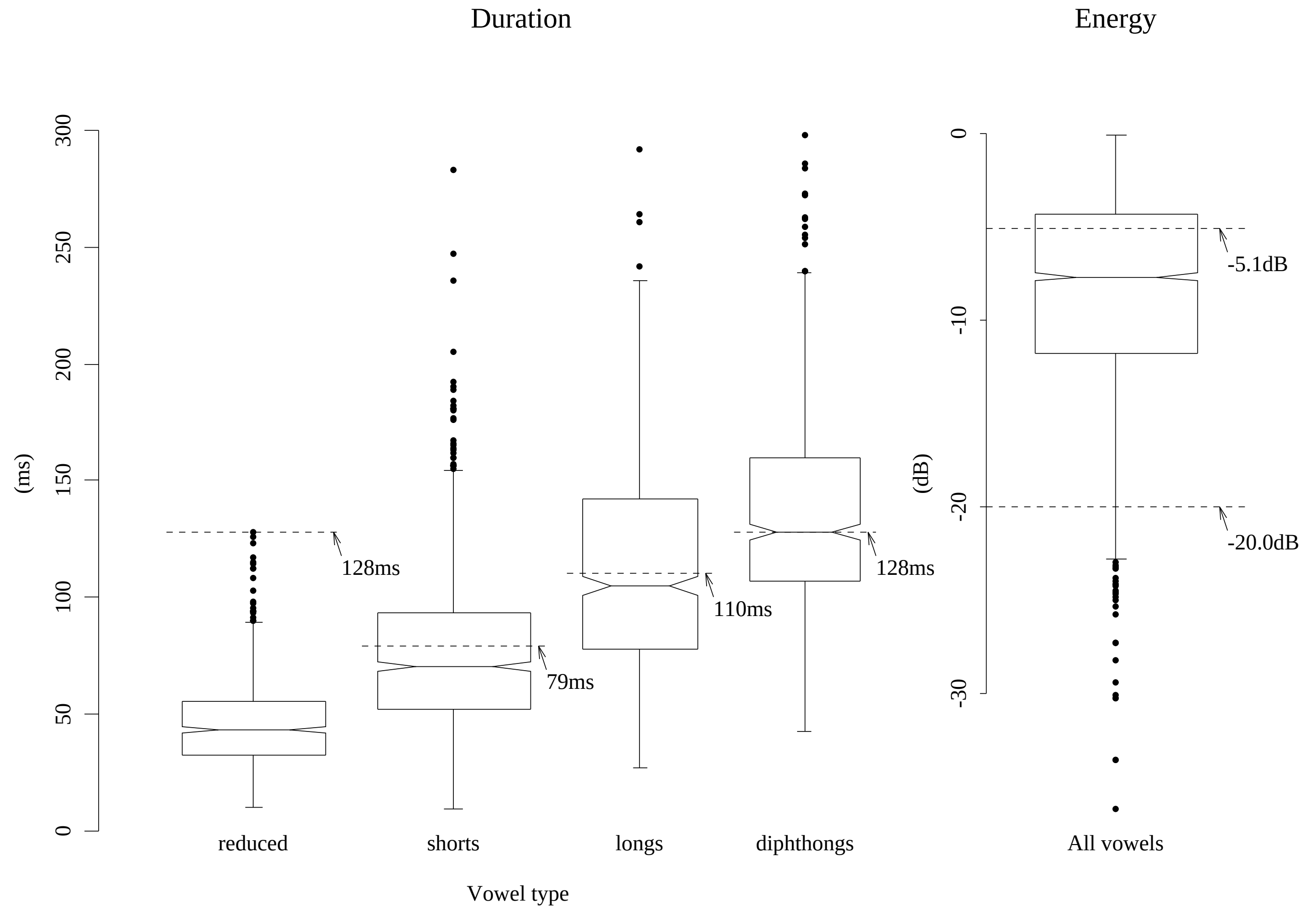

Figure 8.6: Parameter distributions for Hieronymus algorithm

		JLH-system label			
		PA	s	u	total
	PA	1267 (23.8%)	328 (6.2%)	242 (4.5%)	1837 (34.5%)
Hand Label	s	107 (2.0%)	58 (1.1%)	96 (1.8%)	261 (4.9%)
	u	461 (8.7%)	223 (4.2%)	2537 (47.7%)	3221 (60.6%)
	total	1835 (34.5%)	609 (11.4%)	2875 (54.1%)	5319 (100.0%)

Correct classification rate = 3862/5319 (72.6%)

Table 8.7: JLH-system: Confusion matrix of prosodic transcription
(PA — pitch accented; s — stressed but unaccented; u — unstressed)

are the same.

A binary decision is made as to whether an algorithm classifies a syllable correctly or incorrectly. A two-by-two cross-performance matrix is used to indicate the number of syllables which are correctly classified by two different algorithms, the number of syllables which are incorrectly classified by both algorithms, and the number of syllables which are correctly classified by one of the algorithms but incorrectly classified by the other algorithm (or visa versa). The cross-performance matrices for each two-way selection of the algorithms (the PCB-system, the AW-system and the JLH-system) are shown in Table 8.8.

The null hypothesis H_0 is tested by applying a two-tailed test to an observation drawn from a Binomial distribution. The probability P of the observation is calculated from a cross-performance matrix using the method described by Gillick & Cox (1989).

In comparing the error rates of the PCB-system and the AW-system, and the error rates of the PCB-system and the JLH-system, P is negligible. The null hypothesis is therefore rejected and there is evidence of a genuine difference. However, in comparing the error rates of the AW-system and the JLH-system, $P = 0.028$. The null hypothesis cannot by rejected at the 2% significance level.

The PCB-system and the JLH-system use the same pitch accent decision filter. There are fewer stressed but unaccented and unstressed syllables which are confused as being pitch accented by the PCB-system than by the JLH-system. (Compare the left-hand column of Tables 8.5 & 8.7.) This reduction in errors is due to the stylisation and

		AW-system	
		Correct	Incorrect
PCB-system	Correct	3307	899
	Incorrect	468	645

		JLH-system	
		Correct	Incorrect
PCB-system	Correct	3404	802
	Incorrect	458	655

		AW-system	
		Correct	Incorrect
JLH-system	Correct	3056	806
	Incorrect	719	738

Table 8.8: Cross-performance matrices of algorithm pairs

schematisation of the *FØ* contour in the PCB-system. The application of the extensive processing of the *FØ* contour appears to satisfy its goal of preventing microprosodic variations from being confused as pitch accents. There is, however, a slight increase in the number of pitch accented syllables which are confused as being unstressed by the PCB-system in comparison with the JLH-system. Some pitch accent related *FØ* variations may also be removed by the stylisation and schematisation processes.

There are fewer prominent syllables which are confused as being non-prominent syllables by the PCB-system than by the AW-system. There are also fewer non-prominent syllables which are confused as being prominent syllables by the PCB-system than by either the AW-system or the JLH-system. This marked reduction in errors is due to the normalisation of the acoustic parameters for non-prosodic aspects of speech in the PCB-system.

Methods of normalising acoustic parameters for non-prosodic aspects of speech, and the techniques of processing *FØ* contours to isolate the intonational component are proposed in this thesis. It is concluded from the above comparison that these techniques provide a statistically significant improvement in the performance of automatic prosodic analysis systems.

8.6 Conclusions

An automatic prosodic analysis system is described which integrates the unidimensional contributions made by features of duration, energy and fundamental frequency. The system produces an acoustic-phonetic representation of sentential stress patterns and intonation within a syllabic domain.

An algorithm is proposed to group phones automatically into syllables using acoustic-phonetic parameters. The syllabification algorithm is unique in its use of both a low-band energy contour and the segmentation (phone boundary and label information) of an utterance. Each syllable generated by this algorithm is associated with one local maximum in the low-band energy contour per group of potential syllable nuclei (vowels and syllabic consonants). It is assumed that such syllables can only ever be perceived as single prominent units in connected speech. This automatic syllabification from acoustic-phonetic parameters is shown to have a large level of agreement (95.2%) with a syllabification based on abstract phonological rules. The automatic syllabification algorithm has advantages over the syllabification based on phonological rules in that it is robust to errors in the segmentation of an utterance and that it groups phones according to the manner in which the speaker produces the utterance rather than according to some predicted manner. This is particularly important for the syllabification of an utterance produced by a non-native speaker, where the actual pronunciation may deviate from a pronunciation which may be predictable from phonological rules only for a native speaker.

Syllables are used in the extraction of duration and energy features and in the processing of fundamental frequency. The syllabic domain also ties together the information embedded within these three acoustic parameters.

The processing of fundamental frequency is based on the principle that an *FØ* contour has a composite structure (see Section 3.4). The piece-wise stylisation of an *FØ* contour aimed at eliminating microprosodic variations is illustrated. The piece-wise sections are abstracted to form *FØ* trajectory descriptors which are independent of speaker-dependent effects. This leads to a schematic representation of an *FØ* contour which provides an acoustic-phonetic representation of the intonation of an utterance. The analysis of an *FØ* schema by a pitch accent decision filter provides judgements of syllable accentuation. The processing of raw *FØ* contours to form *FØ* schemata reduces the number of unaccented syllables which are erroneously classified as being pitch accented by the pitch accent decision filter.

The unidimensional contributions to the perceived prominence of a syllable made by features of duration and energy are captured by a peak-picking algorithm. The information captured by the peak-picking algorithm and the syllable accentuation information are combined to produce a description of the sentential stress patterns of an utterance.

The intergrated prosodic analysis system yields a 79.1% agreement level with sentential stress categories transcribed by hand. This is shown to be a statistically significant improvement over the agreement levels yielded by two former algorithms. This level of agreement is comparable with the 83.1% agreement between transcriptions of syllable prominences made independently by two phoneticians in the Lancaster/IBM spoken English corpus (Pickering *et al.*, 1994).