Chapter 4

Duration measures

The acoustic parameters used in the analysis of prosodic aspects of speech must be extracted from a speech waveform. There are four acoustic parameters which are correlated with prosodic phenomena — duration, energy, vowel quality and fundamental frequency. This Chapter and Chapters 5, 6 & 7 concentrate on the extraction of each of these acoustic parameters from a speech signal.

The domain of phonetic units whose duration are to be determined as optimal correlates of stress and normalisation techniques applied to them are investigated in this Chapter. The phonetic units investigated are related to syllable structure. The normalisation techniques aim to compensate for variations in duration which are due to non-prosodic aspects of speech. The underlying principle of these investigations is to normalise the acoustic parameters for non-prosodic aspects of speech such that the processed acoustic parameters can then be combined to form a prosodic description of speech.

4.1 Database description

Two databases of hand labelled phonetically balanced sentences are used in the experimental investigations of duration measures (presented in this Chapter) and of energy measures and vowel quality measures (presented in Chapters 5 & 6 respectively). The first database (referred to as the *training data*) consists of 200 sentences read by a male speaker of British English (South-eastern dialect) with a non-pathological voice. The read sentences are recorded in an anechoic studio with a close-talking microphone and a

16-bit analogue-to-digital converter sampling at 20kHz. The second database (referred to as the *test data*) consists of 460 sentences read by the same speaker under the same conditions. The speaker was unaware during recording that the utterances would be used in a study of prosodic analysis.

The training data sentences are a set of 200 'phonetically rich' sentences designed so as to provide almost total coverage of permissible demi-syllables in English (Laver *et al.*, 1988). The test data sentences are a set of 460 'phonetically compact' sentences designed so as to provide as complete a coverage of phoneme pairs as possible (Lamel *et al.*, 1986). These 460 sentences are the same as the anglicised TIMIT sentences used in SCRIBE (Spoken Corpus Recordings In British English). All the speech data is phonetically transcribed by phoneticians using the criteria proposed by Laver *et al.* (1989).

Phones are grouped into syllabic units automatically using the algorithm described in Section 8.1 and the perceived prominence of each syllable is transcribed by hand. The number of degrees of stress that constitute an adequate description of speech is disputed amongst linguists (Section 2.2.1). The practical approach taken here is to describe syllables as either prominent or non-prominent. If they are prominent, they may or they may not be associated with pitch accents. Pitch accented prominent syllables are further classed into nuclear and non-nuclear accents. Thus, syllables perceived as bearing sentential stress (Section 2.1.1) are transcribed either as nuclear accented "n" (primary stress), as non-nuclear accented "a" (secondary stress) or as unaccented but stressed "s" (tertiary stress); otherwise they are transcribed as unstressed "u"[1] (Bagshaw & Williams, 1992). Nuclear and non-nuclear accented syllables are collectively referred to as pitch accented "PA" in this thesis. Pitch accented and stressed (but unaccented) syllables are collectively referred to as prominent "P". In contrast to prominent syllables, unstressed syllables are also referred to as non-prominent "NP".

The transcription of the perceived prominence of each syllable in the training and test data is performed by the author (who is also a native speaker of British English). There are inevitably instances when the discrete categorisation of a syllable as prominent or non-prominent is ambiguous. Hence, the transcription of the data cannot be regarded as definitive. There is therefore some degree of error in the comparisons of prominence

[1] See Section 2.2.4 for a definition of these terms.

levels labelled by hand and labelled by the automatic procedures discussed below. In the prosodic labelling of the Lancaster/IBM spoken English corpus, two phoneticians working independently to transcribe syllables as either prominent "P" or non-prominent "NP" are reported to achieve 90.8% agreement (Pickering *et al.*, 1994). The transcribers are reported to achieve 83.1% agreement in categorising syllables as either pitch accented "PA", stressed "s" or unstressed "u".

4.2 Phonetic units

The term *duration* refers to the length of a particular constituent of speech. In the context of prosodic analysis, the constituent of speech is usually a type of phonetic or phonological unit. Some level of abstraction from the acoustic waveform has, therefore, already been applied to derive the unit and so duration is not strictly an acoustic parameter. However, duration is widely regarded as a fundamental acoustic correlate of prosodic phenomena (see Section 3.1). The units whose durations best correlate with sentential stress and normalisation techniques which can be applied to them, are investigated here. The aim of the investigation is to determine a duration feature whose distributions for each prominence level have the greatest separability.

The units whose durations are investigated are:

- U_1, the syllable nucleus.
- U_2, syllable nucleus and coda (the syllable *rhyme*).
- U_3, the syllable onset and nucleus (which is referred to hereinafter as the syllable *lhyme* — a 'left-hand' equivalent to the rhyme).
- U_4, the entire syllable (onset, nucleus and coda).
- U_5, a nucleus-to-nucleus unit.

All the phones within a syllable preceding the nucleus form the onset, and all those succeeding the nucleus form the coda. The nucleus-to-nucleus unit starts at the beginning of a syllable nucleus and ends either at the beginning of the nucleus of a following syllable, or, if it is followed by a pause, at the end of the syllable coda. The structures of these

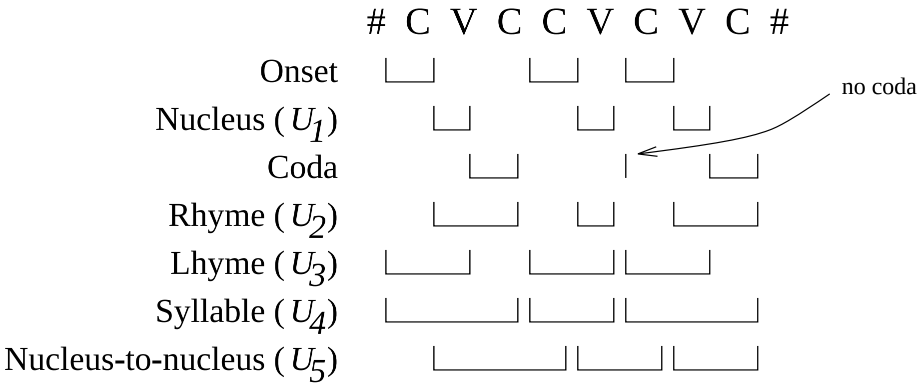

Figure 4.1: Syllable structure

units are illustrated in Figure 4.1. The clustering of phones into syllable-type units is performed automatically using the algorithm described in Section 8.1. The syllable nuclei are also identified by this algorithm.

4.3 Normalisation techniques

The duration of a unit is influenced by many parameters other than just its relative prominence. A normalisation technique is required to compensate for any variations in the duration of a unit that result from such parameters. Several approaches to duration normalisation are investigated.

Klatt (1979) describes a rule based system to estimate phonemic segment durations for synthesised speech in American English. The application of a set of rules modifies the "inherent" durations (specified in a look-up table) of each segment according to its phonemic context, its degree of stress and its proximity to syntactic junctures (word, phrase and clause boundaries). The prosodic analysis of an utterance involves determining the contribution made by stress to the duration of a segment. This may be achieved by employing Klatt's rules to predict the duration of a segment in an unstressed context. The difference between the predicted duration and the actual duration of a segment is a theoretical estimate of the contribution made by stress to the segment duration. The phonetic context of each segment and the syntactic structure of the utterance need to be known in order to predict the duration of a segment in this way. However, the syntactic structure of an utterance is usually not readily available from the speech waveform (without reliable speech recognition and syntactic parsing). Furthermore, there is no simple way of automatically determining either the factors by which durations are modified under different conditions or the inherent segment durations for a given speaker. This approach is therefore not used in the prosodic analysis system proposed in this thesis.

Fant (1988) proposes the use of duration indices for units in a study of sentential stress in Swedish. A duration index S_i is derived from the measured duration of a unit T_{unit} consisting of i phones, and two reference durations, $\overline{T_u}(i)$ and $\overline{T_s}(i)$. $\overline{T_u}(i)$ is the average duration of non-prominent units which consist of i phones, and $\overline{T_s}(i)$ is the average duration of prominent units which consist of i phones. The reference durations are empirically determined from the training data and are calculated separately for each number of phones i that can exist within a unit.

$$S_i = 1 + \frac{T_{unit} - \overline{T_u}(i)}{\overline{T_s}(i) - \overline{T_u}(i)} \qquad (4.1)$$

Campbell & Isard (1991) note that the distributions of segment durations differ by phone-type in terms of both their mean and population standard deviation values. The individual phone durations are normalised by applying a z_{mean}-transform which expresses the phone duration T_{phone} in terms of the mean duration for that phone-type $\mu_T(phone\_type)$ and standard deviation for that phone-type $\sigma_T(phone\_type)$. The values of $\mu_T(phone\_type)$ and $\sigma_T(phone\_type)$ are empirically determined from the training data.

$$z_{mean} = \frac{T_{phone} - \mu_T(phone\_type)}{\sigma_T(phone\_type)} \qquad (4.2)$$

The z_{mean} duration measure for a segment will be either positive, if the segment duration is longer than the mean duration for the given phone-type, or negative, if the segment duration is shorter than the mean. Segment durations are thus offset relative to the mean durations of their phone-type and scaled relative to the spread of their distributions. As noted by Wightman *et al.* (1992), a positive z_{mean} duration measure does *not* necessarily mean that a segment is lengthened (due to syllable prominence and/or prosodic phrasing). This is because the mean duration $\mu_T(phone\_type)$ is determined from all occurrences of a particular phone, including unlengthened ones. There is generally an inequality in the number of lengthened and unlengthened phones. Thus, if there are more lengthened segments than unlengthened segments then there will be a tendency for the mean duration to be somewhat larger than the minimum duration of lengthened phones. Therefore, lengthened segments may have a negative z_{mean} duration measure. Conversely, if there are more unlengthened segments than lengthened segments then there will be a tendency for the mean duration to be somewhat smaller than the maximum duration of unlengthened phones. Therefore, unlengthened segments may have a positive z_{mean} duration measure.

The normalisation procedure which Fant (1988) proposes does not take into account the variance of segment duration as a function of phone-type, and the Campbell & Isard (1991) normalisation method does not take into account the variance of segment duration as a function of the relative prominence of a unit or the number of phones in a unit.

It is proposed in this thesis that the z-transform used by Campbell & Isard can be modified so that it also takes the relative prominence of a unit into account. To do this, first assume that the duration of a phone in a prominent unit is greater than the duration of the same phone-type in a non-prominent unit. Thus, in a distribution of durations for a given phone-type containing p percent non-prominent phones, the lower p percent of durations are assumed to correspond to the durations of the non-prominent phones. Instead of offsetting the segment durations relative to the mean durations of their phone-type, the durations are expressed relative to the p'th-percentile duration for that phone-type $\varphi_T^p(phone\_type)$. This gives rise to a $z_{percentile}$-transform,

$$z_{percentile} = \frac{T_{phone} - \varphi_T^p(phone\_type)}{\sigma_T(phone\_type)} \tag{4.3}$$

If the assumption is correct, then the $z_{percentile}$ duration measure is positive for all prominent phones and negative for all others.

The prominent units in the training database are indicated by a transcription on the nucleus of the unit. It is assumed that the prominence level of a unit only affects the duration of the phones within that unit. Therefore, a non-nuclear phone inherits the prominence level of the nucleus in the same unit for the purpose of calculating its

p'th-percentile. All non-nuclear phones that exist outside the definition of the unit are assumed to be non-prominent.

An underlying assumption of the z_{mean} normalisation (Equation 4.2) is that the populations of duration measurements for phone-types are Normal (Gauss-Laplacian) distributions. These distributions are modelled by the two parameters $\mu_T(phone\_type)$ and $\sigma_T(phone\_type)$. The populations of duration measurements for phones are shown in Figures 4.2 (vowels) & 4.3 (consonants). These populations have positively skewed distributions and can, therefore, only be modelled approximately by the mean and standard deviation parameters. A more robust set of parameters is desired to model these distributions. The $z_{percentile}$ normalisation (Equation 4.3) effectively adjusts the mean parameter by considering the population of duration measurements for a phone-type as constituting two sub-populations — one population of the non-prominent phones and another population of prominent phones.

The duration of each phone, T_{phone} is determined directly from the phonetic transcription. This may be a transcription made either by hand or by some automatic procedure. The z-transformed phone durations (either z_{mean} or $z_{percentile}$ duration measures) derived from these are listed in chronological order to form a duration contour for each utterance. A duration contour can be smoothed using a non-linear smoother (Section 7.3.1) with a window length of 3 phones. Such smoothing aims to iron out boundary placement errors between each pair of phones. Boundary placement errors in auto-segmented data are superimposed on time quantisation errors which can be as much as 5ms. In order to test whether smoothing the duration contour is or is not beneficial to the prosodic analysis, the z-transform normalisation techniques are investigated both without smoothing (S_0) and with smoothing (S_1). Note, however, that the training data and test data used in this investigation are transcribed by hand. Quantisation errors are therefore negligible.

Some phones occur infrequently in the training data. There is therefore the possibility that the sample of phones in the training data is unrepresentative of the universal set of phones. In order to improve the robustness of the data, it may be possible to classify phone-types on a broad phonetic basis rather than on a fine phonetic basis. The duration distributions for phones classified on a fine phonetic basis are shown in Figures 4.2 (vowels) & 4.3 (consonants). Phones may be grouped into broad phonetic classes on the

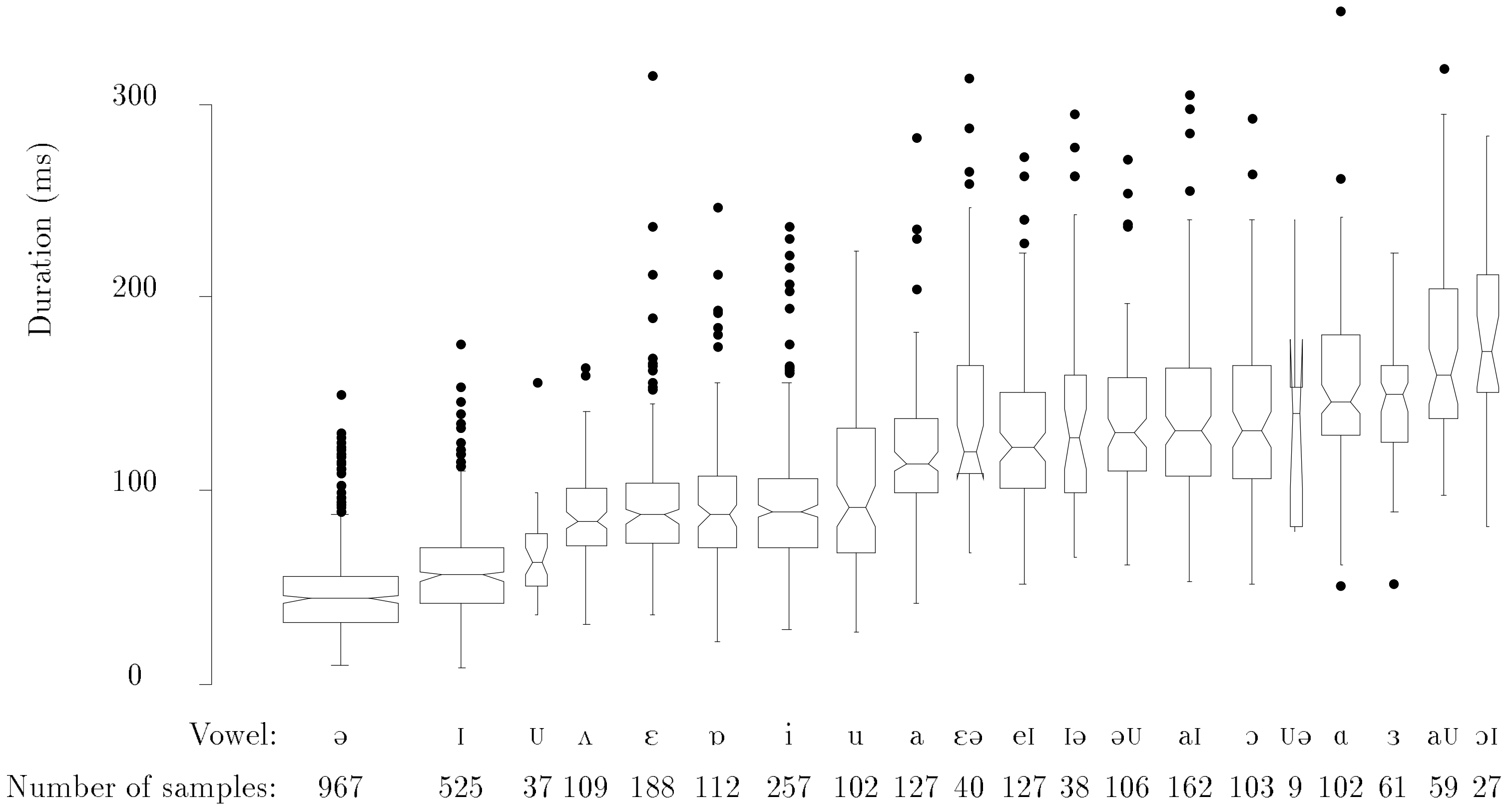

Figure 4.2: Vowel durations

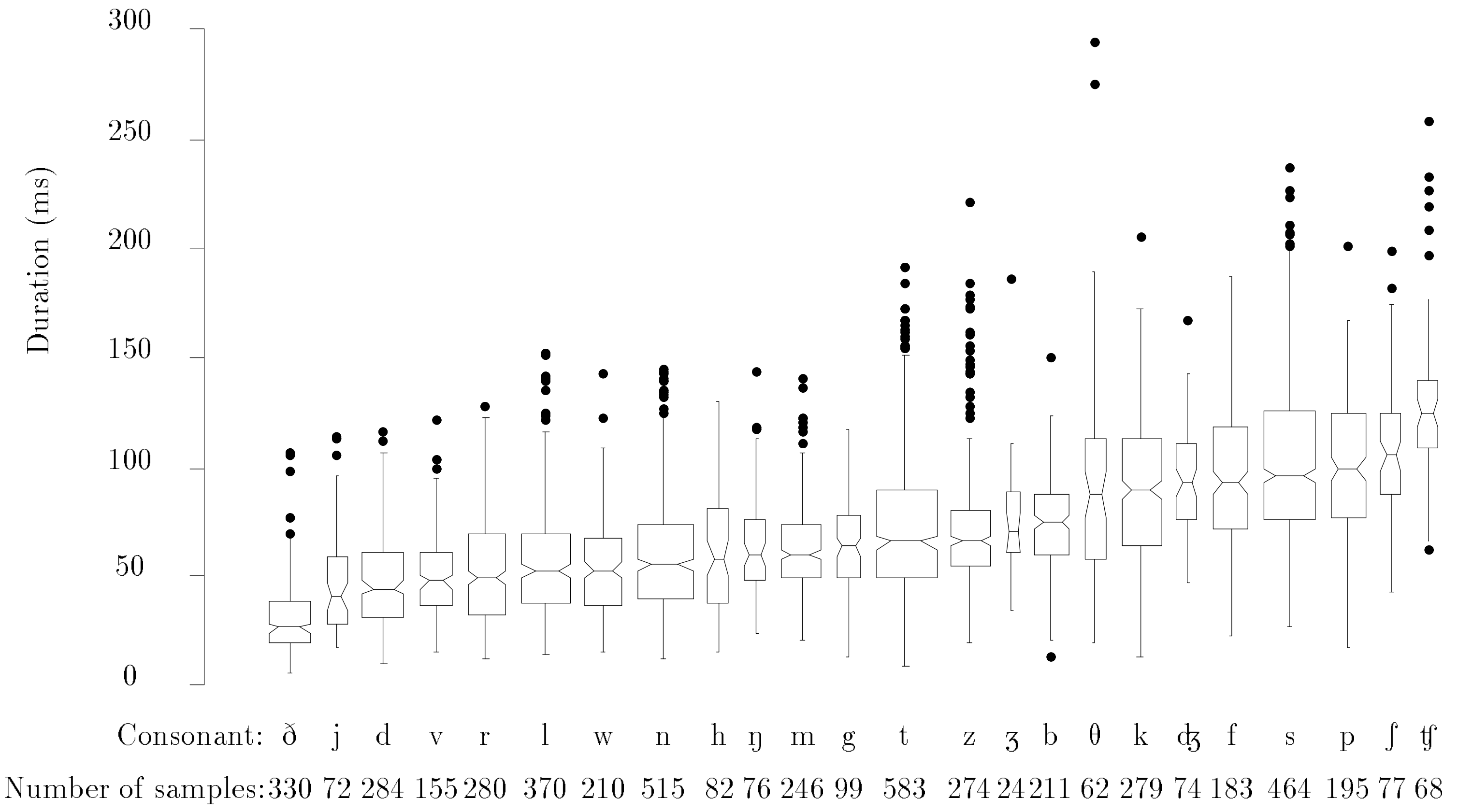

Figure 4.3: Consonant durations

Broad phonetic class	Fine phonetic classes
reduced monophthong	/ə/
short monophthong	/ɪ ʊ ʌ ɛ ɒ a/
long monophthong	/i u ɔ ɑ ɜ/
diphthong	/eɪ ɔɪ aɪ əʊ aʊ ɪə ʊə ɛə/
sonorant	/l r w j n m ŋ/
voiced obstruent	/ð d v g z ʒ b ʤ/
unvoiced obstruent	/θ t f k s ʃ p ʧ h/

Table 4.1: Duration: Broad and fine phonetic classes

basis of phonetic principles (Ladefoged, 1982) and similarities in their observed duration distributions, as indicated in Table 4.1. The effect of using either a fine (P_0) or broad (P_1) phonetic classification on the z-transform normalisation techniques, and subsequently the effect on the prosodic analysis, is investigated.

Each type of unit $U_{1,\ldots,5}$ contains a number of phones and each phone duration can be normalised by a number of different techniques. One of the following measures can be used to represent the duration of a unit in an utterance:

- M_0, unnormalised unit duration, T_{unit}.
- M_1, maximum z_{mean} duration measure of the phones in a unit.
- M_2, maximum $z_{percentile}$ duration measure of the phones in a unit.
- M_3, sum z_{mean} duration measures of all phones in a unit (same as M_1 for unit U_1, since unit U_1 only ever contains one phone).
- M_4, sum $z_{percentile}$ duration measures of all phones in a unit (same as M_2 for unit U_1).

None of the unit duration measures $M_{0,\ldots,4}$ normalise for the number of phones in the unit, as proposed by Fant. Two methods of normalising for the number of phones in the unit are investigated. The first of these methods F_1 is identical to the duration index in Equation 4.1. Therefore, normalisation method F_1 takes into account the variance of unit duration measure as a function of the relative prominence of a unit, as well as taking into account the number of phones in a unit. This may be redundant for the unit duration measures $M_{2,4}$ based on the $z_{percentile}$-transform. A second normalisation method F_2 is therefore proposed which aims only to compensate for the number of phones in a unit.

- F_0, no normalisation for the number of phones in a unit.

- F_1, duration index for a unit, S_i.

$$S_i \;=\; 1 + \frac{M - \overline{M_u}(i)}{\overline{M_s}(i) - \overline{M_u}(i)} \tag{4.4}$$

  where $M \in M_{0,\ldots,4}$, i represents the number of phones in the unit, $\overline{M_u}(i)$ is the mean value of M for all non-prominent units containing i phones, and $\overline{M_s}(i)$ is the mean value of M for all prominent units containing i phones.

- F_2, duration index for a unit, S_i' (equivalent to F_0 for unit-type U_1, since unit U_1 only ever contains one phone).

$$S_i' \;=\; M - \overline{M}(i) \tag{4.5}$$

  where $M \in M_{0,\ldots,4}$ and $\overline{M}(i)$ is the mean value of M for all units containing i phones.

The mean unit duration measures $\overline{M_u}(i)$, $\overline{M_s}(i)$ and $\overline{M}(i)$ are determined from the training data. If there are only a few examples of the units containing certain numbers of phones then the sample set can be unrepresentative of the data. Values for the mean unit duration measures are therefore accepted only if there are five or more examples of stressed units and five or more examples of unstressed units containing a particular number of phones, and if $\overline{M_u}(i) < \overline{M_s}(i)$ for that number of phones.

If, in analysing an utterance, a unit is encountered which contains a number of phones for which $\overline{M_u}(i)$, $\overline{M_s}(i)$ and $\overline{M}(i)$ are unknown, then values for these statistics are estimated by applying polynomial interpolation and extrapolation to those which are known. If $\overline{M_u}(i) \geq \overline{M_s}(i)$ for extrapolated values of i, then the normalisation technique breaks down.

In summary, the investigation examines five different types of unit $U_{1,\ldots,5}$. There are five unit duration measures $M_{0,\ldots,4}$ of which only three apply to unit-type U_1. Four of the unit duration measures $M_{1,\ldots,4}$ are related to the z-transform which can be based on either broad or fine phonetic classes — $P_{0,1}$ — and may or may not involve non-linear smoothing — $S_{0,1}$. Each of the five unit duration measures $M_{0,\ldots,4}$ for unit-types $U_{2,\ldots,5}$ may or may not be processed (in one of two ways) to normalise for the number of phones in the unit — $F_{0,1,2}$. The combinations of different units and normalisation techniques give rise to 213 possible duration features in all, any one of which is an optimal duration correlate of sentential stress.

4.4 Evaluation of duration features

The aim of the investigation is to determine which of these 213 duration features is the most effective measure in distinguishing between non-prominent and prominent syllables. The investigation uses a technique which classifies each syllable as either non-prominent or prominent given a single duration feature. The technique only uses the value of the duration feature for the syllable which is to be classified.

Each syllable is assigned to a category (either as a non-prominent syllable or as a prominent syllable) by a unidimensional Bayesian classifier[2] (see Appendix B) which uses the magnitude of a duration feature as its input parameter. In essence, this technique involves training models which consist of the mean feature vector and the covariance matrix for each category to which a token (in this case, a syllable) can be assigned, and the *a priori* probability of a token existing in a given category. A weighted variance-normalised distance measure (the quadratic discriminant score) is calculated between the feature vector for a given token and the centroid of each trained model. The token is assigned to the category which results in the shortest distance. A by-product of the Bayesian classifier is its *entropy score*, which is the theoretical number of bits of additional information required by the classifier to derive the category of a token (the prominence of a syllable) without error. In comparing the performance of the Bayesian classifier given different duration features, a lower entropy score indicates a better duration feature.

The classification of syllables, using just one duration measure at a time, is applied

[2] The Bayesian classifier is sometimes referred to as Quadratic Discriminant Analysis in the literature.

to a database in an open test. The training data is used to calculate the models for the Bayesian classifier and to determine the mean, percentile and standard deviations of the phone durations. The test data is not used in any of the training procedures.

The entropy scores are determined for the unidimensional Bayesian classifier using each of the 213 duration features as the only input parameter. These scores are shown in Figures 4.4, 4.5, 4.6, 4.7 & 4.8. All entropy scores are shown in each Figure. Each column shows the entropy scores as a function of one variable for each permissible combination of the other four variables. The columns are ranked such that the best combination lies to the left-hand side and the worst combination lies to the right-hand side. For example, each column of Figure 4.4 shows the entropy scores as a function of unit-type $U_{1,\ldots,5}$ for each permissible combination of the variables $M_{0,\ldots,4}$, $P_{0,1}$, $S_{0,1}$ and $F_{0,1,2}$. The combination of phonetic unit and normalisation techniques corresponding to the minimum entropy score (and hence the best combination of those examined) is the syllable lhyme U_3 with sum $z_{percentile}$ duration measures M_4 trained on broad phonetic classes P_1 without smoothing of the phone-level duration contour S_0 and with normalisation of the number of phones in the syllable lhyme using technique F_1.

Figure 4.4 ranks the phonetic units in order of decreasing correlation with sentential stress as the syllable lhyme U_3, the syllable nucleus U_1, the entire syllable U_4, the syllable rhyme U_2 and finally the nucleus-to-nucleus unit U_5. The correlation with syllable prominence therefore decreases with an increase in the number of phones to the right of the syllable nucleus relative to the number of phones to the left of the nucleus. Factors other than prominence cause the duration of phones to the right of the nucleus to vary more than the duration of phones to the left of the nucleus. This result supports rules for the synthesis of segment durations proposed by Klatt (1979). Klatt's rules shorten the duration of consonants in a non-word-initial position (phones in the syllable coda) and lengthen the duration of the syllable nucleus and any following consonants (phones in the rhyme) when they are in a pre-pausal position. It is claimed by Campbell (1990) that duration lengthening due to prominence affects the entire syllable U_4 whereas lengthening due to prosodic boundaries affects the syllable rhyme U_2. Wightman & Ostendorf (1991; 1992) use the average z_{mean} duration measure of phones in the rhyme of the final syllable in a word as a feature for the automatic assignment of break indices. The break indices

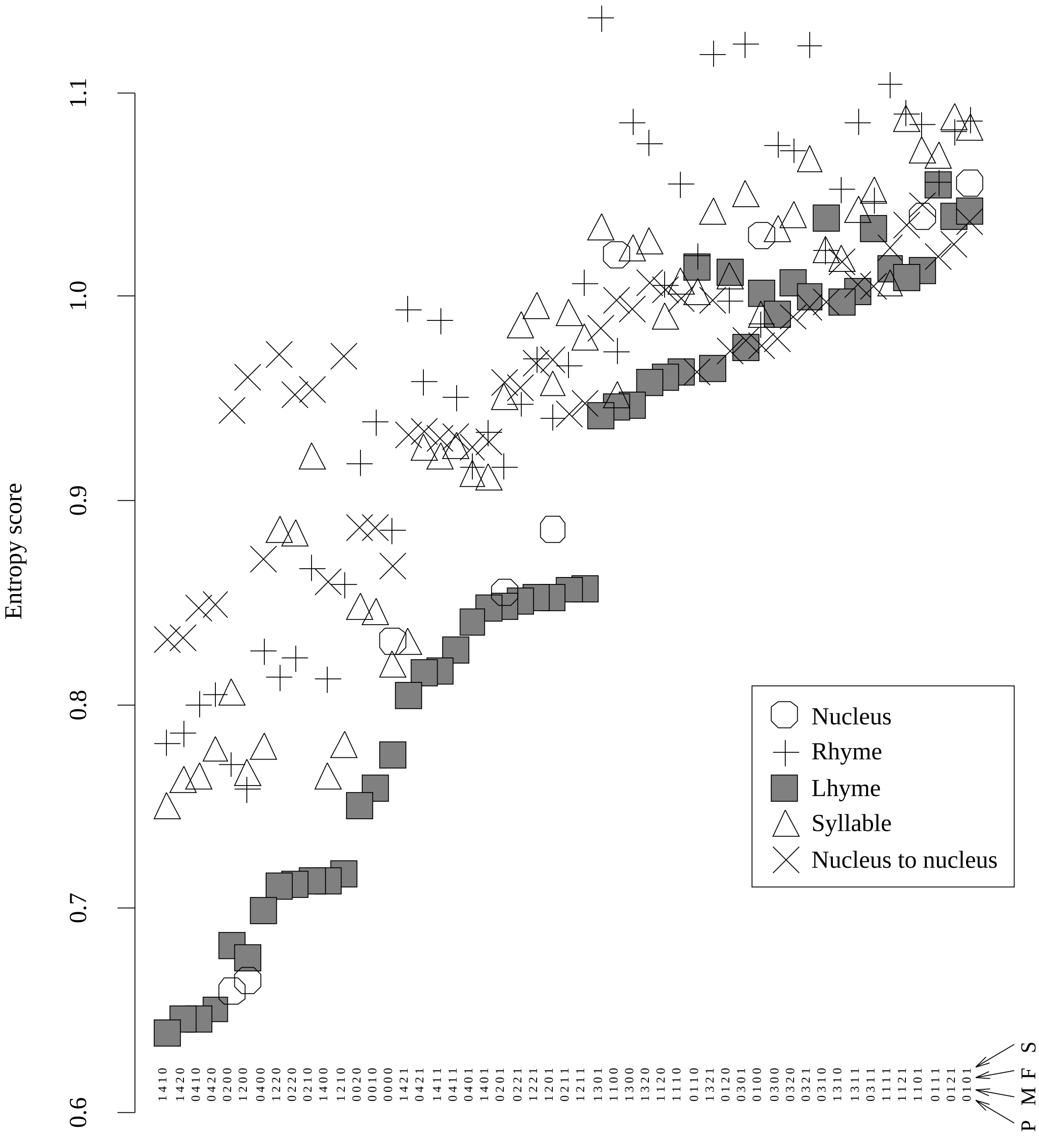

Figure 4.4: Duration feature evaluation: Entropy score ($U_{1,\ldots,5}$)

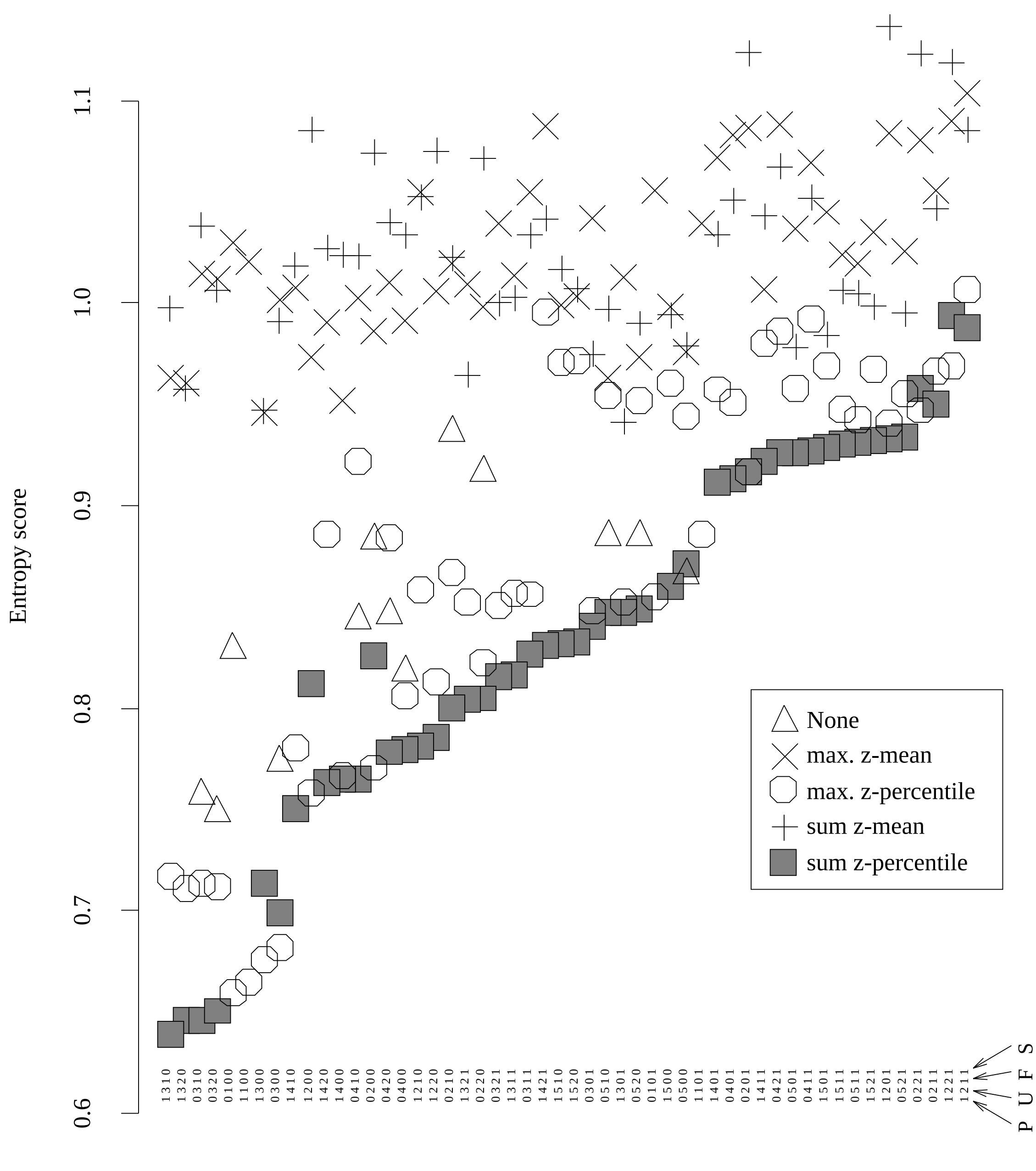

Figure 4.5: Duration feature evaluation: Entropy score ($M_{0,\ldots,4}$)

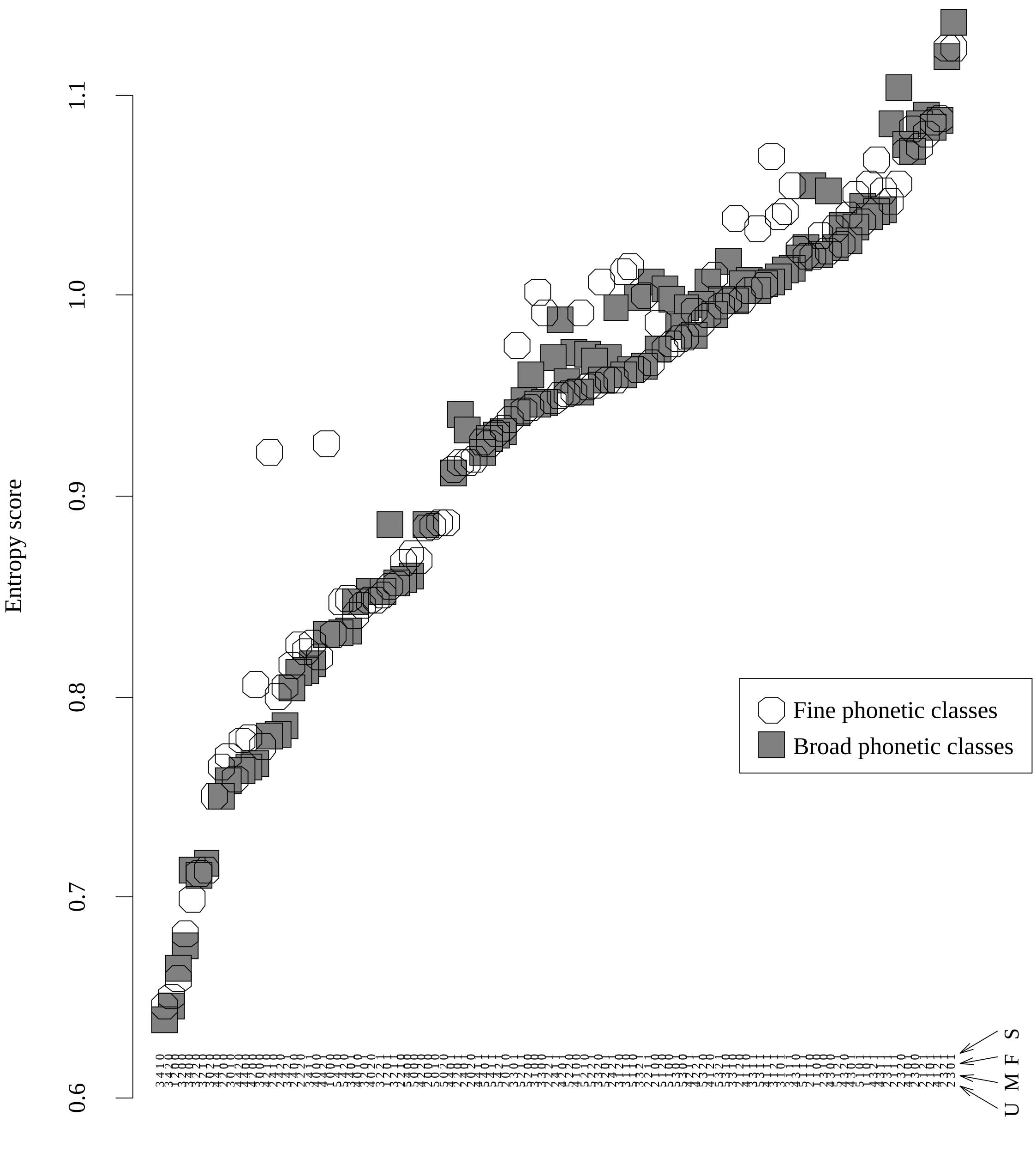

Figure 4.6: Duration feature evaluation: Entropy score ($P_{0,1}$)

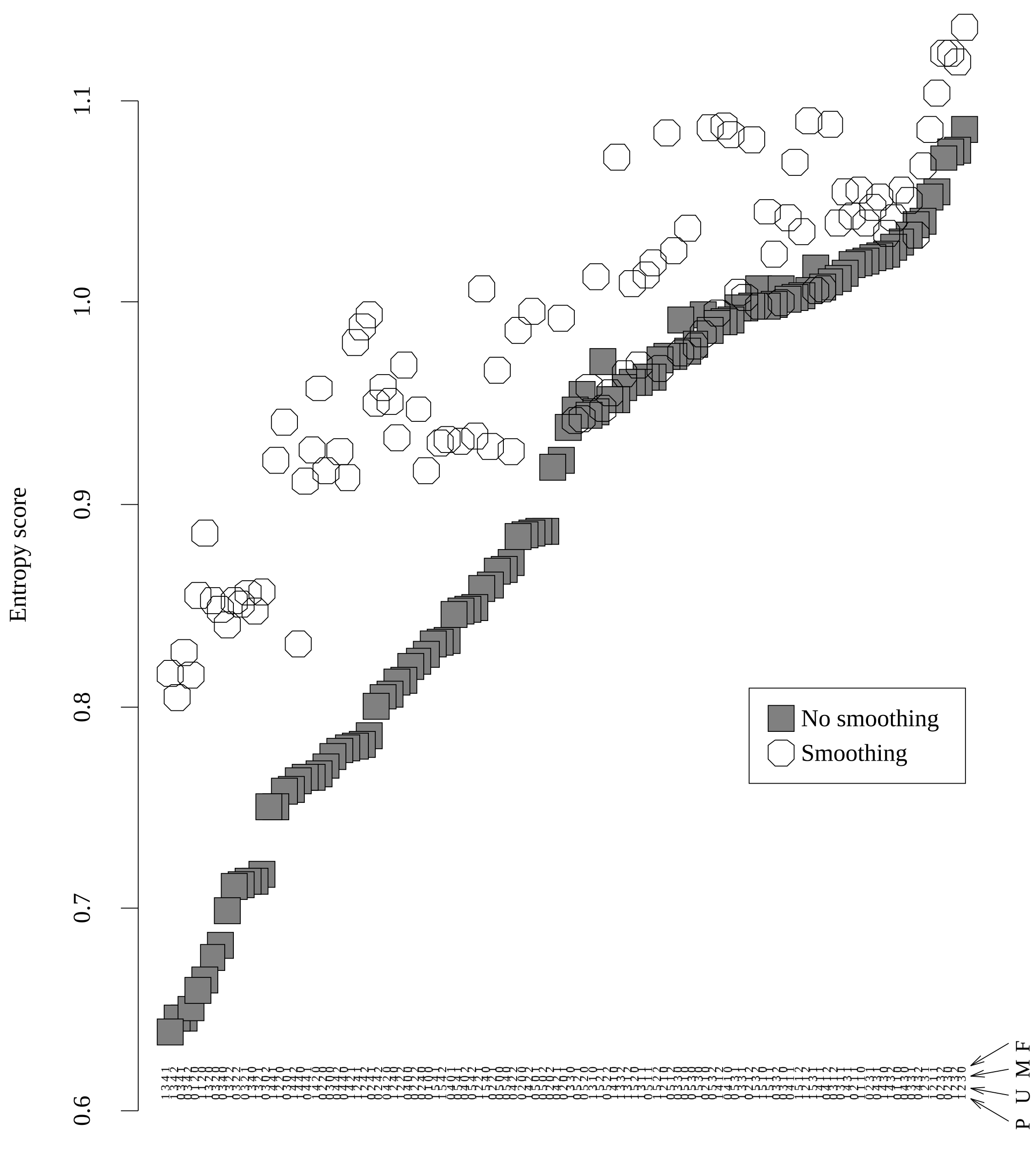

Figure 4.7: Duration feature evaluation: Entropy score ($S_{0,1}$)

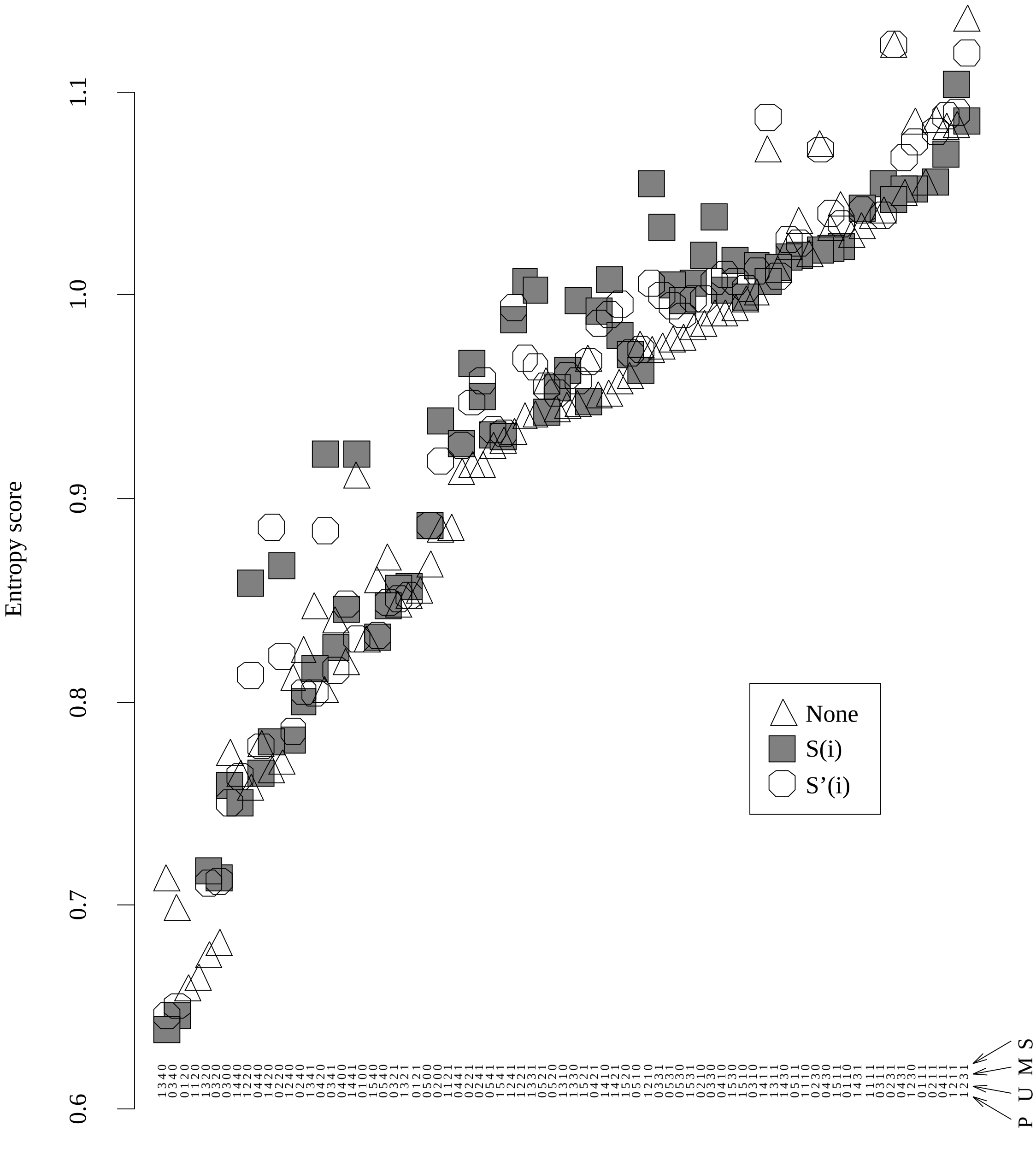

Figure 4.8: Duration feature evaluation: Entropy score ($F_{0,1,2}$)

are used to represent the strength of boundaries between neighbouring words and are thus related to the prosodic structure of an utterance.

Figure 4.5 shows a clear division between z_{mean} based normalisations (represented by plusses and crosses) and $z_{percentile}$ based normalisations (represented by octagons and shaded squares). The $z_{percentile}$ duration measures yield the preferable lower entropy scores. Lower entropy scores are obtained by using the sum of z-transformed phone durations in a unit than are obtained by using the maximum z-transformed phone duration. The z_{mean} duration measures give consistently higher entropy scores than using no normalisation for phone-type.

There is little difference in the entropy scores shown in Figure 4.6 for z-transformation statistics trained on fine and broad phonetic classes. Normalisation for broad phonetic classes results in a slightly lower entropy score than normalisation for fine phonetic classes. There is therefore only a small possibility that the training data is unrepresentative of the universal set of phones classified on a fine phonetic basis (with respect to duration analysis).

Smoothing of the duration contour causes an increase in entropy, as illustrated in Figure 4.7. The smoothing process therefore eliminates information required in the prosodic analysis rather than compensating for phone boundary placement errors in a way which could aid the analysis. The application of such smoothing is therefore not beneficial to the prosodic analysis.

Figure 4.8 reveals that normalisation for the number of phones in a unit is dependent upon the duration measure and phonetic unit-type used. No normalisation for the number of phones F_0 gives a lower entropy score for many of the combinations of duration measure and phonetic unit-type. However, in the case of the sum $z_{percentile}$ duration measure M_4 for the lhyme U_3 (shown in the left-most column) the entropy score is significantly lower when normalisation for the number of phones in the lhyme is incorporated. Normalisation method F_1 takes into account the variance of the lhyme duration as a function of the relative prominence of the syllable, as well as taking into account the number of phones in the lhyme. A degree of redundancy is expected for the $z_{percentile}$-based duration measures $M_{2,4}$ since the $z_{percentile}$-transform also takes into account the variance of the lhyme duration as a function of the relative prominence of the syllable. Normalisation method F_2 only compensates for the number of phones in the lhyme. Figure 4.8 shows that the difference in entropy score between methods F_1 and F_2 is small (but statistically significant), with method F_1 resulting in the lower entropy score. A difference in entropy exists because the assumptions made about the nature of the variation in the lhyme duration due to the prominence of a syllable differ in the F_1 normalisation method and the $z_{percentile}$-transform. Method F_1 assumes that the mean duration measure for prominent syllables is greater than the mean duration measure for non-prominent syllables. The $z_{percentile}$-transform assumes that all of the lower duration measures correspond to non-prominent syllables and that all of the higher duration measures correspond to prominent syllables. Making both assumptions yields a lower entropy score.

4.5 Optimisation of a relative duration feature

The application of a Bayesian classifier assumes that the prominence of a syllable can be determined from the magnitude of a duration feature in isolation from its neighbours. The variation of a duration feature relative to neighbouring syllables is not included as an input parameter to the classifier. This exclusion cannot be beneficial to the Bayesian classifier, given that a syllable's prominence is relative to its neighbours by definition. A possible solution to this problem is to include the duration measures for neighbouring syllables as input parameters to the Bayesian classifier. This might improve the Bayesian classifier's performance (in terms of reducing the entropy score) when given the task of classifying a syllable as either non-prominent or prominent. However, this approach is not investigated here because an emphasis is placed on employing non-statistical methods such that the underlying principles used to locate prominent syllables can be easily identified.

A simple "peak-picking" technique is proposed here which classifies the syllable as either non-prominent or prominent depending on the magnitude of the duration feature relative to its nearest neighbours. A syllable is classified as prominent if its duration feature exceeds that of both its neighbours (end-points being inherently lower) and if such a local maximum in the duration contour has a magnitude greater than some specified threshold $K_{duration}$. Syllables satisfying this criterion are labelled "*sd*". A syllable is also classified as prominent if it corresponds with the maximum value of the duration feature in the contour (independently of the threshold). Syllables corresponding to such maxima are labelled "Sd". All other syllables are labelled "ud".

The duration feature used corresponds to the syllable lhyme U_3 with sum $z_{percentile}$ durations M_4 trained on broad phonetic classes P_1 without smoothing of the phone-level duration contour S_0 and with normalisation of the number of phones in the syllable lhyme using technique F_1 — ie. the duration feature with the minimum entropy score in the above investigation. Theoretically, this duration feature has a mean value of 2.0 for prominent syllables and a mean value of 1.0 for non-prominent syllables. $K_{duration}$ should therefore be optimal at 1.5. The performance of the peak-picking method is evaluated by examining the percentage of syllables which are assigned to the correct category[3] (prominent or non-prominent). The optimal value of $K_{duration}$ (ie. the value corresponding to the greatest percentage of syllables assigned to the correct prominence category) is sought in a closed test. The data used to train the $z_{percentile}$ statistics and the mean duration measures used in the F_1 normalisation method is also used in optimising $K_{duration}$. A plot of the percentage of correct categorisation against $K_{duration}$ is shown in Figure 4.9. The optimal value of $K_{duration}$ is 1.5, as predicted.

An entropy score is not produced by the peak-picking method. Its performance is therefore compared with the performance of the Bayesian classifier by examining the percentage of syllables which are assigned to the correct prominence category in an open test. The levels of agreement between the prominence category of a syllable as transcribed by hand and as assigned automatically are shown in Tables 4.2 & 4.3 for the peak-picking algorithm and for the Bayesian classifier respectively. The simple peak-picking algorithm has a number of benefits over the Bayesian classifier. The peak-picking algorithm yields an improvement in syllable prominence classification over the Bayesian classifier, requires less computation (and hence it is faster) than the Bayesian classifier, and does not requiring a training stage.

[3]It is assumed that the "correct" prominence category is that which is transcribed by hand, despite the possibility of human error.

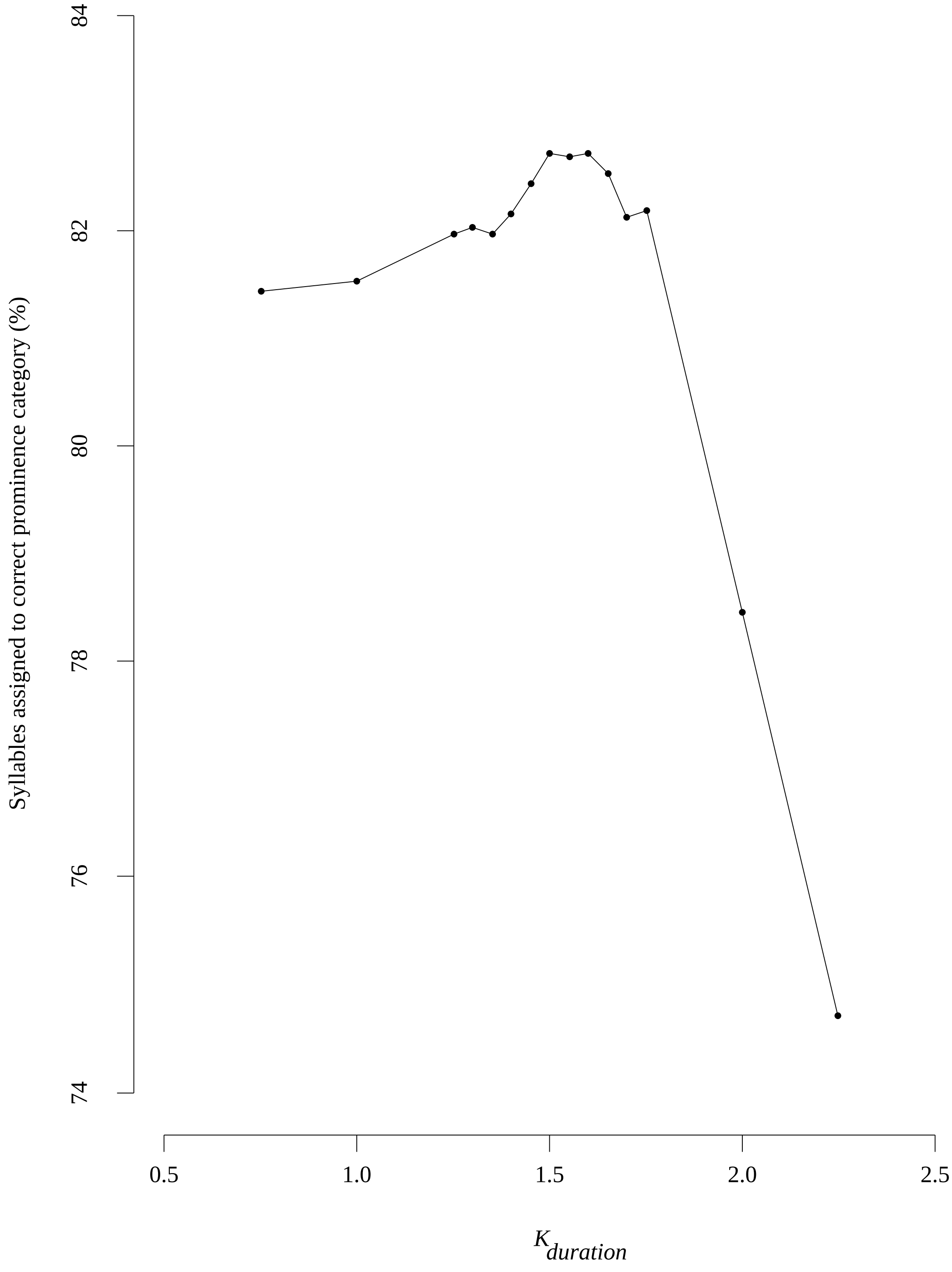

Figure 4.9: Optimisation of $K_{duration}$ threshold

		Peak-picking algorithm label			
		Sd	sd	ud	total
Hand Label	P	367 (6.9%)	1027 (19.3%)	704 (13.2%)	2098 (39.4%)
	NP	93 (1.7%)	386 (7.3%)	2742 (51.6%)	3221 (60.6%)
	total	460 (8.6%)	1413 (26.6%)	3446 (64.8%)	5319 (100.0%)

Correct classification rate = 4136/5319 (77.8%)

Table 4.2: Duration: Confusion matrix for peak-picking algorithm
(Sd — syllable with maximum duration of contour; sd — syllable corresponding to a local peak in the duration contour greater than $K_{duration}$; ud — unstressed syllable on the basis of duration; P — prominent syllable; NP — non-prominent syllable)

		Bayesian classifier label		
		P	NP	total
Hand Label	P	1565 (29.4%)	533 (10.0%)	2098 (39.4%)
	NP	723 (13.6%)	2498 (47.0%)	3221 (60.6%)
	total	2288 (43.0%)	3031 (57.0%)	5319 (100.0%)

Correct classification rate = 4063/5319 (76.4%)

Table 4.3: Duration: Confusion matrix for Bayesian classification
(P — prominent syllable; NP — non-prominent syllable)

4.6 Summary

The investigation described above aimed to determine a duration feature whose distributions for prominent and non-prominent syllables have the greatest separability. The entropy score obtained by a Bayesian classifier was used as a robust measure of this separability for 213 different duration features. The optimal duration feature $D_{feature}$ of those investigated is given by:

$$D_{feature} = 1 + \frac{M - \overline{M_u}(i)}{\overline{M_s}(i) - \overline{M_u}(i)} \tag{4.6}$$

where,

$$M = \sum_{phone \in lhyme} \frac{T_{phone} - \varphi_T^p(broad\_phone\_class)}{\sigma_T(broad\_phone\_class)} \tag{4.7}$$

and where i represents the number of phones in the lhyme, $\overline{M_u}(i)$ is the mean value of M for all non-prominent lhymes containing i phones, $\overline{M_s}(i)$ is the mean value of M for all prominent lhymes containing i phones in the training data, T_{phone} is the duration of a phone, $\varphi_T^p(broad\_phone\_class)$ and $\sigma_T(broad\_phone\_class)$ are the p'th percentile and the standard deviation of phone durations grouped on a broad phonetic basis respectively, and p is the percentage of non-prominent phones in the training data.

A simple peak-picking algorithm applied to this duration feature is shown to give an improvement of syllable prominence classification over the Bayesian classifier, while requiring less computation and not requiring an additional training stage. In an open test, 77.8% of syllables are given the same prominence status by the peak-picking algorithm as when transcribed by hand. This level of agreement is dependent upon the discrete categorisation of syllable prominence perceived by a human transcriber, and as such, is regarded in this thesis as a comparative measure only.

The duration features investigated include the normalisation of variations in duration due to phone-type, syllable structure, syllable length in terms of number of phones, and the prominence level of syllables within the training data. However, these are not the only factors which influence the duration of a segment. The normalisation of segmental context effects and pre-pausal lengthening are not included in the investigation (although it is argued that the lhyme is not affected by pre-pausal lengthening to as great an extent as other syllable components). Further research is required to investigate the possible advantages that can be gained by incorporating models of contextual influences on segment duration (Bartkova & Sorin, 1987; Crystal & House, 1988; van Santen, 1993) into the normalisation of non-prosodic effects. The duration of speech segments is also affected by variations in the rate of articulation from utterance to utterance. The inclusion of compensating for articulation rate differences (Wightman *et al.*, 1992) is also a subject of further investigation.

Chapter 5

Energy measures

The formation of a frame-level low-band energy contour is described in Section 5.1. The contour is used as an acoustic representation of the perceived variations of intensity in an utterance. The frame-level low-band energy contour is reduced to a phone-level low-band energy contour so that each phone in an utterance is associated with a single measure of intensity. The low-band energy of phones in a variety of phonetic units (Section 4.2) and normalisation techniques are investigated in this Chapter with respect to forming an energy feature correlated with sentential stress. The methodology used in the investigation of duration features is also used in this Chapter.

5.1 Low-band energy contour

The energy contour for a speech waveform is calculated from frames of data (at 5ms intervals) such that energy values are synchronised with the data used in *FØ* extraction (see Chapter 7) and with the formant frequency calculations used in the vowel quality analysis (see Chapter 6). Each frame is passed through a Blackman-Harris window (Harris, 1978) in order to filter out discontinuities (and hence high frequency artefacts) at the boundaries of the analysis frame. The frequency bins of an amplitude spectrum (calculated by a 512-point FFT for 20ms frames of data sampled at 20kHz) corresponding to the range 50Hz–2kHz are accumulated. The resultant energy values are expressed in decibels with respect to the maximum frame energy in the utterance to form an utterance-normalised low-band energy contour. Expressing the energy values in these terms normalises for utterance-to-utterance variations in energy which are dependent upon recording conditions such as the extent of signal pre-amplification and the distance of the microphone from the voice source. The contour is processed by a three-frame median filter and five-frame Hann window smoother (Rabiner *et al.*, 1975) in order to remove small perturbations which arise during frames of speech with low fundamental frequency (typically less than two fundamental periods per analysis frame).

A phone-level low-band energy contour is generated from the (frame-level) utterance-normalised energy contour. Each phone in the phonetic transcription of an utterance is associated with the energy value of one of the frames within the phone. If one or more peaks (local maxima flanked by lower values) in the energy contour occur within the phone then the energy value at the highest peak is associated with the phone. If such a peak does not exist then either a single valley in the energy contour occurs within the phone or the energy contour continuously rises or falls during the phone. If a single valley exists then the energy value at the valley is associated with the phone. Otherwise, the phone is associated with the energy value at its mid-point.

An example of a phone-level low-band energy contour and its corresponding (frame-level) utterance-normalised energy contour and speech waveform are illustrated in Figure 5.1. The dots on the frame-level energy contour indicate the frame energy peak/valley/mid-point values associated with each of the phones. The phone-level contour follows the same general trends as the frame-level contour. By associating a phone with the frame energy value at the point of a local maximum flanked by lower values during the phone rather than with the maximum frame energy value in the phone, the formation of the phone-level low-band energy contour becomes more robust to variations in the placement of phone boundaries. This is illustrated in the cases of the second [f] and the [v] in Figure 5.1. The peak/valley/mid-point value associated with a phone is also robust to phonetic-context. For example, the value of the valley in the [t] is less changeable to variations in the peak energies of the neighbouring vowels than the maximum frame energy of the phone. Although the peak/valley/mid-point value is robust to phone boundary placement and phonetic context, it is not immune to potential errors. If the boundary between the [i] and the [l] is placed beyond the local valley, or if the left context of the [l] is a low low-band energy phone such as an /f/, then the [l] would be

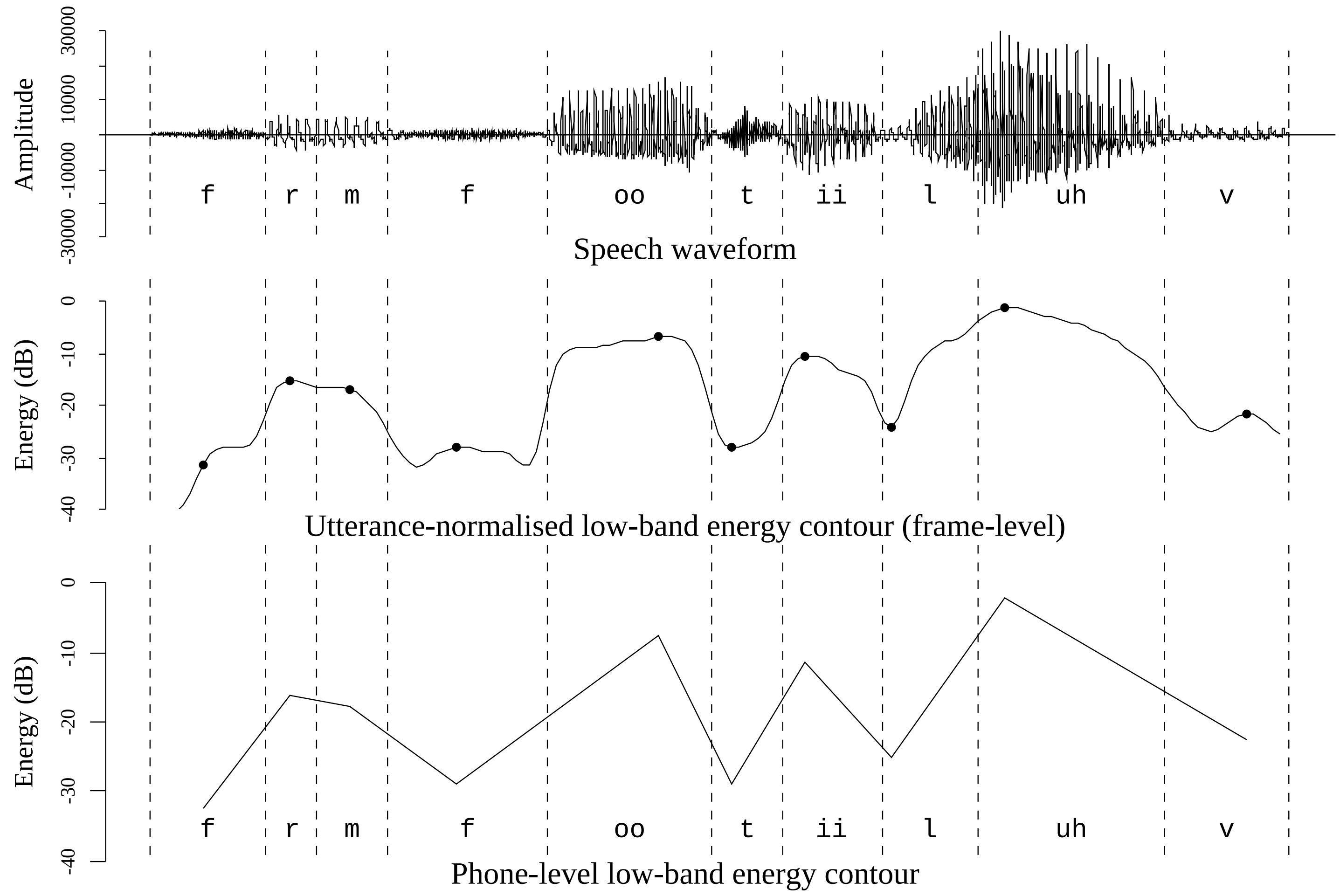

Figure 5.1: Formation of a phone-level low-band energy contour*

associated with the higher mid-point frame energy value rather than the frame energy value at the local valley.

5.2 Energy features

The low-band energy of each phone in an utterance is calculated using the procedure described in Section 5.1. The phonetic units whose low-band energy best correlate with sentential stress and normalisation techniques which can be applied to them, are studied here. As in the investigation of duration (Chapter 4), the aim is to determine an energy feature whose distributions for each prominence level have the greatest separability.

The units whose energies are investigated are the same as those used in the duration analysis, $U_{1,\ldots,5}$ (Section 4.2). The energy of a unit, like the duration of a unit, is influenced by many parameters other than just its relative prominence. Normalisation techniques are adopted from the duration analysis (Section 4.3) to compensate for any variations in the energy of a unit that result from such parameters.

One of the following measures can be used to represent the low-band energy of a unit in an utterance:

- M_0, unnormalised unit-level low-band energy, E_{unit}. The unit-level low-band energy is generated from the (frame-level) utterance-normalised energy contour by the same peak/valley/mid-point procedure used to determine the phone-level low-band energy.

- M_1, maximum z_{mean} low-band energy measure of the phones in a unit.

$$z_{mean} = \frac{E_{phone} - \mu_E(phone\_type)}{\sigma_E(phone\_type)} \tag{5.1}$$

  where $\mu_E(phone\_type)$ and $\sigma_E(phone\_type)$ represent the mean and standard deviation phone-level low-band energy respectively for each phone-type.

- M_2, maximum $z_{percentile}$ low-band energy measure of the phones in a unit.

$$z_{percentile} = \frac{E_{phone} - \varphi_E^p(phone)\_type}{\sigma_E(phone\_type)} \tag{5.2}$$

Broad phonetic class	Fine phonetic class
closed vowel	/ɪ i ʊ u/
half-closed vowel	/ʊə ɪə eɪ əʊ ɔɪ/
central vowel	/ə/
half-open vowel	/ɔ aɪ ɛ ɛə aʊ ʌ ɜ/
open vowel	/ɒ a ɑ/
liquid	/l r/
glide	/w j/
nasal	/n m ŋ/
voiced obstruent	/b ʒ v g ʤ ð d z/
unvoiced obstruent	/s θ t ʧ f ʃ p k h/

Table 5.1: Low-band energy: Broad and fine phonetic classes

where $\varphi_E^p(phone\_type)$ is the p'th-percentile phone-level low-band energy for each phone-type, and p is the percentage of a phone-type which are non-prominent.

- M_3, sum z_{mean} low-band energy measures of all phones in a unit (same as M_1 for unit U_1, since unit U_1 only ever contains one phone).
- M_4, sum $z_{percentile}$ low-band energy measures of all phones in a unit (same as M_2 for unit U_1).

As in the duration analysis, phone-types can be classified on either a fine or a broad phonetic basis for the z-transform normalisation of phone-level low-band energies. The low-band energy distributions for phones classified on a fine phonetic basis are shown in Figures 5.2 (vowels) & 5.3 (consonants). Phones may be grouped into broad phonetic classes on the basis of phonetic principles (Ladefoged, 1982) and similarities in their observed low-band energy distributions, as indicated in Table 5.1.

Each of the unit low-band energy measures can be normalised for the number of phones in the unit by the techniques $F_{0,1,2}$ described in Section 4.3.

In summary, the investigation examines five different types of unit $U_{1,\ldots,5}$. There are five unit low-band energy measures $M_{0,\ldots,4}$ of which only three apply to unit-type U_1. Four of the unit low-band energy measures $M_{1,\ldots,4}$ are related to the z-transform which can be based on either broad or fine phonetic classes — $P_{0,1}$. No smoothing is

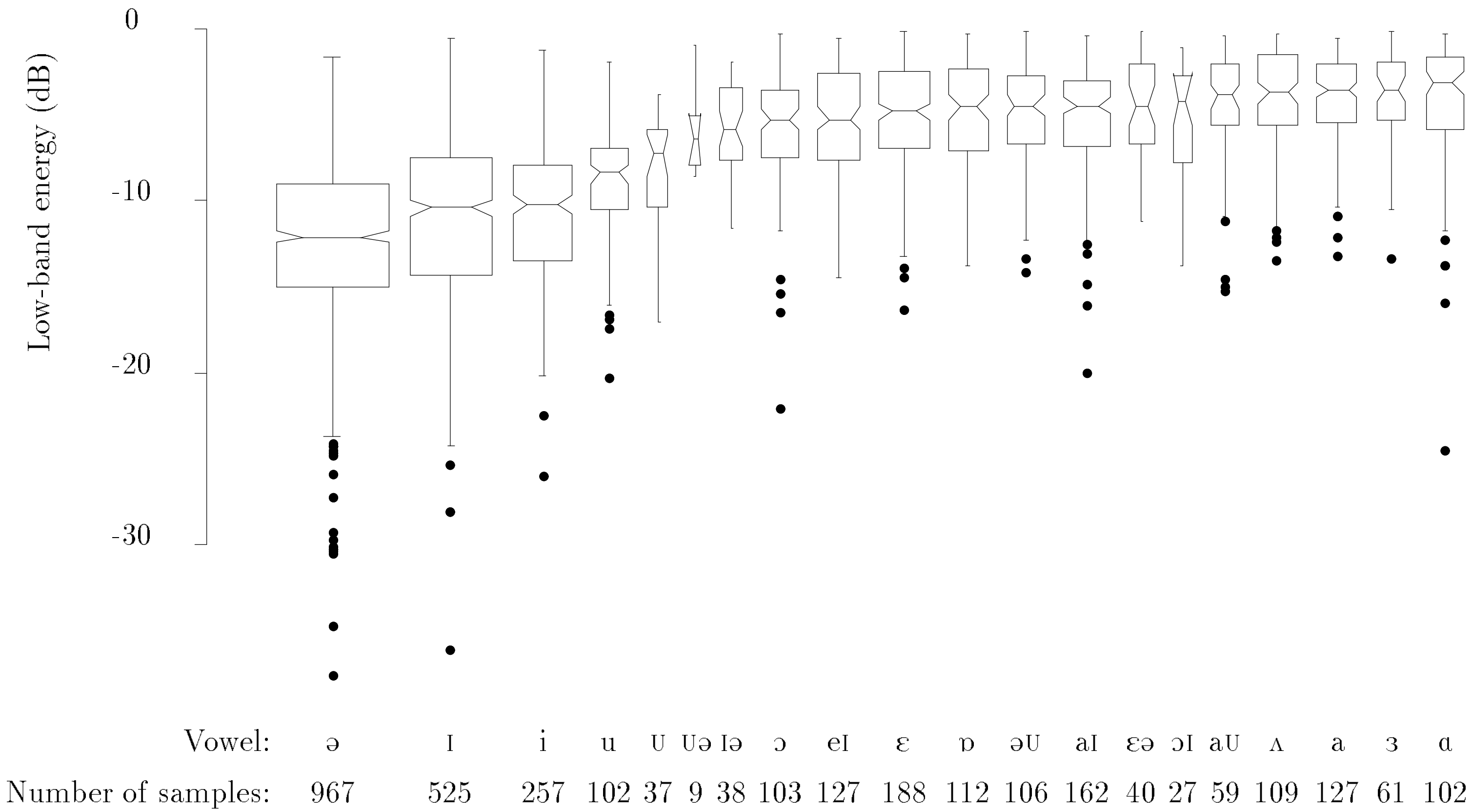

Figure 5.2: Vowel low-band energies

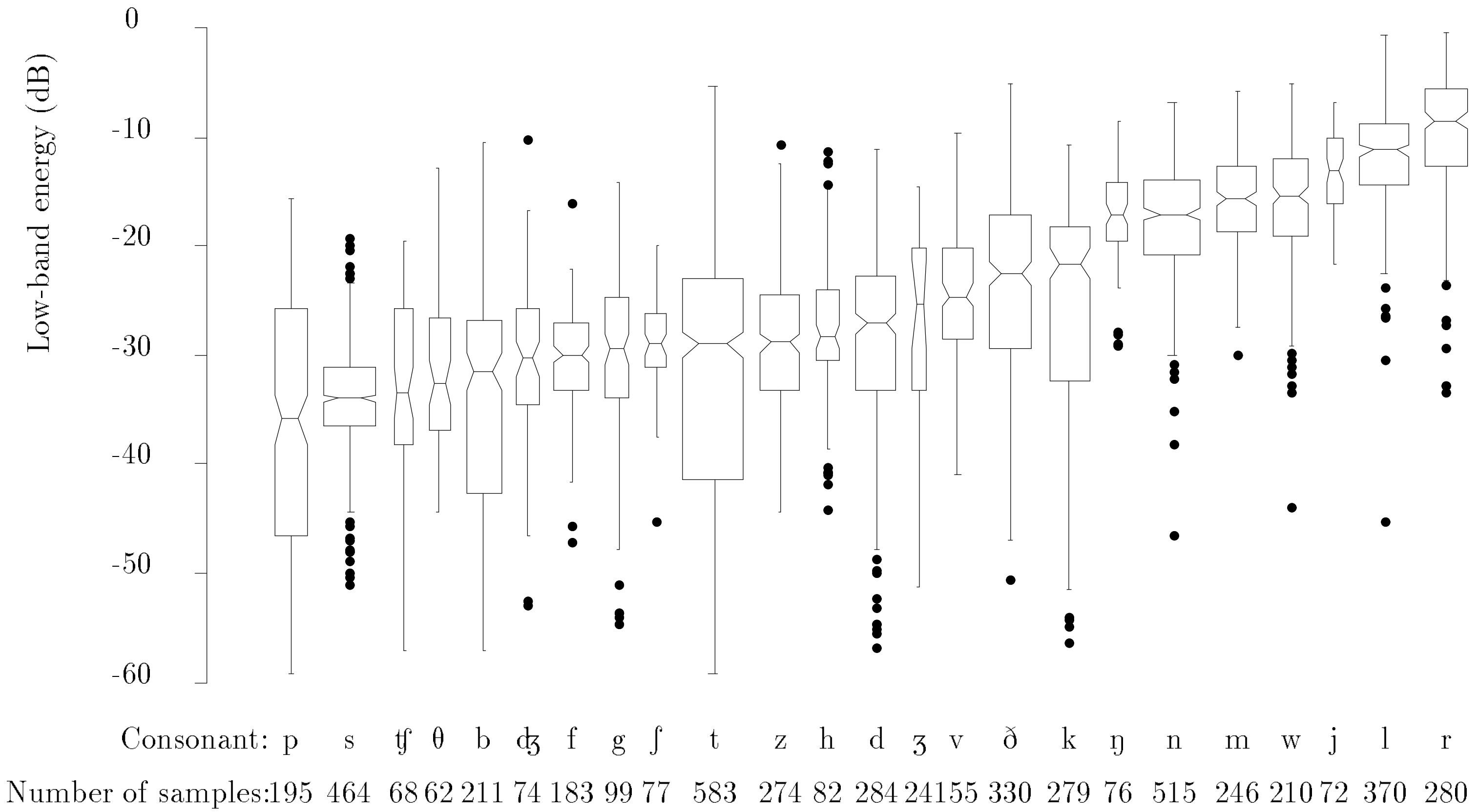

Figure 5.3: Consonant low-band energies

applied to the phone-level low-band energy contour. (Smoothing is applied to the frame-level low-band energy contour.) Each of the five unit low-band energy measures $M_{0,\ldots,4}$ for unit-types $U_{2,\ldots,5}$ may or may not be processed (in one of two ways) to normalise for the number of phones in the unit — $F_{0,1,2}$. The combinations of different units and normalisation techniques give rise to 113 possible energy features in all, any one of which is an optimal energy correlate of sentential stress.

5.3 Evaluation of energy features

The aim of the investigation is to determine which of these 113 energy features is the most effective measure in distinguishing between non-prominent and prominent syllables.

The entropy scores are determined for the unidimensional Bayesian classifier using each of the 113 energy features as the sole input parameter. These scores are shown in Figures 5.4, 5.5, 5.6 & 5.7. All entropy scores are shown in each Figure. Each column shows the entropy scores as a function of one variable for each permissible combination of the other three variables. The columns are ranked such that the best combination lies to the left-hand side and the worst combination lies to the right-hand side. For example, each column of Figure 5.4 shows the entropy scores as a function of unit-type $U_{1,\ldots,5}$ for each permissible combination of the variables $M_{0,\ldots,4}$, $P_{0,1}$ and $F_{0,1,2}$. The combination of phonetic unit and normalisation techniques corresponding to the minimum entropy score (and hence the best combination of those examined with respect to energy) is the syllable nucleus U_1 with $z_{percentile}$ low-band energy measures M_2 trained on fine phonetic classes P_0. Normalisation for the number of phones in a unit $F_{1,2}$ are not applicable to the syllable nucleus.

Figure 5.4 ranks the phonetic units in order of decreasing correlation with sentential stress as the syllable nucleus U_1, the syllable rhyme U_2, the entire syllable U_4, the syllable lhyme U_3 and finally the nucleus-to-nucleus unit U_5. By definition, the syllable nucleus is the peak of sonority in a syllable. Sentential stress is therefore correlated with an energy feature based on the peak of sonority in a syllable.

As is observed in the investigation of duration features, Figure 5.5 shows a clear division between z_{mean} based normalisations and $z_{percentile}$ based normalisations, in this case, with respect to energy. Lower entropy scores are yielded by $z_{percentile}$ low-band

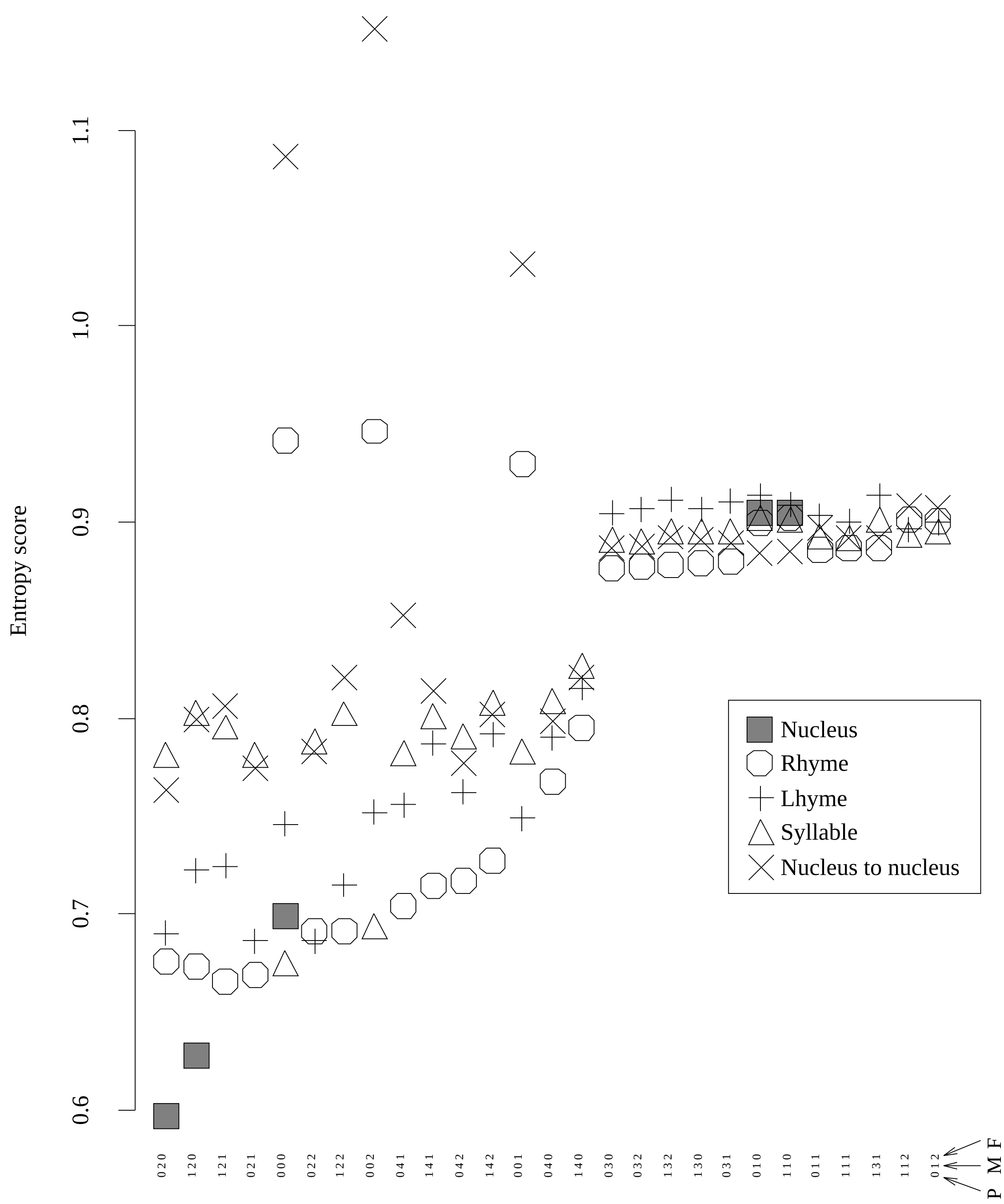

Figure 5.4: Energy feature evaluation: Entropy score ($U_{1,\ldots,5}$)

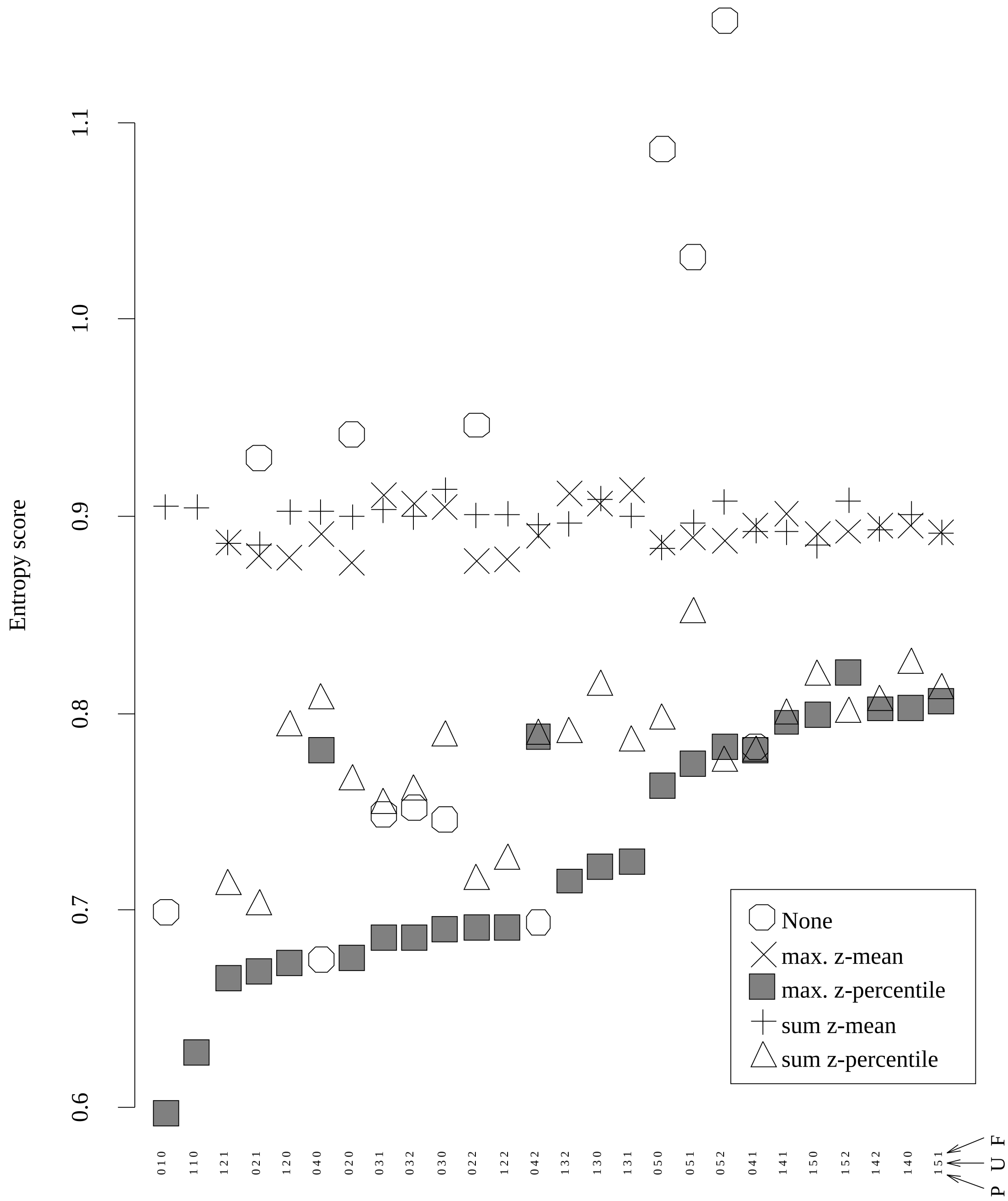

Figure 5.5: Energy feature evaluation: Entropy score ($M_{0,\ldots,4}$)

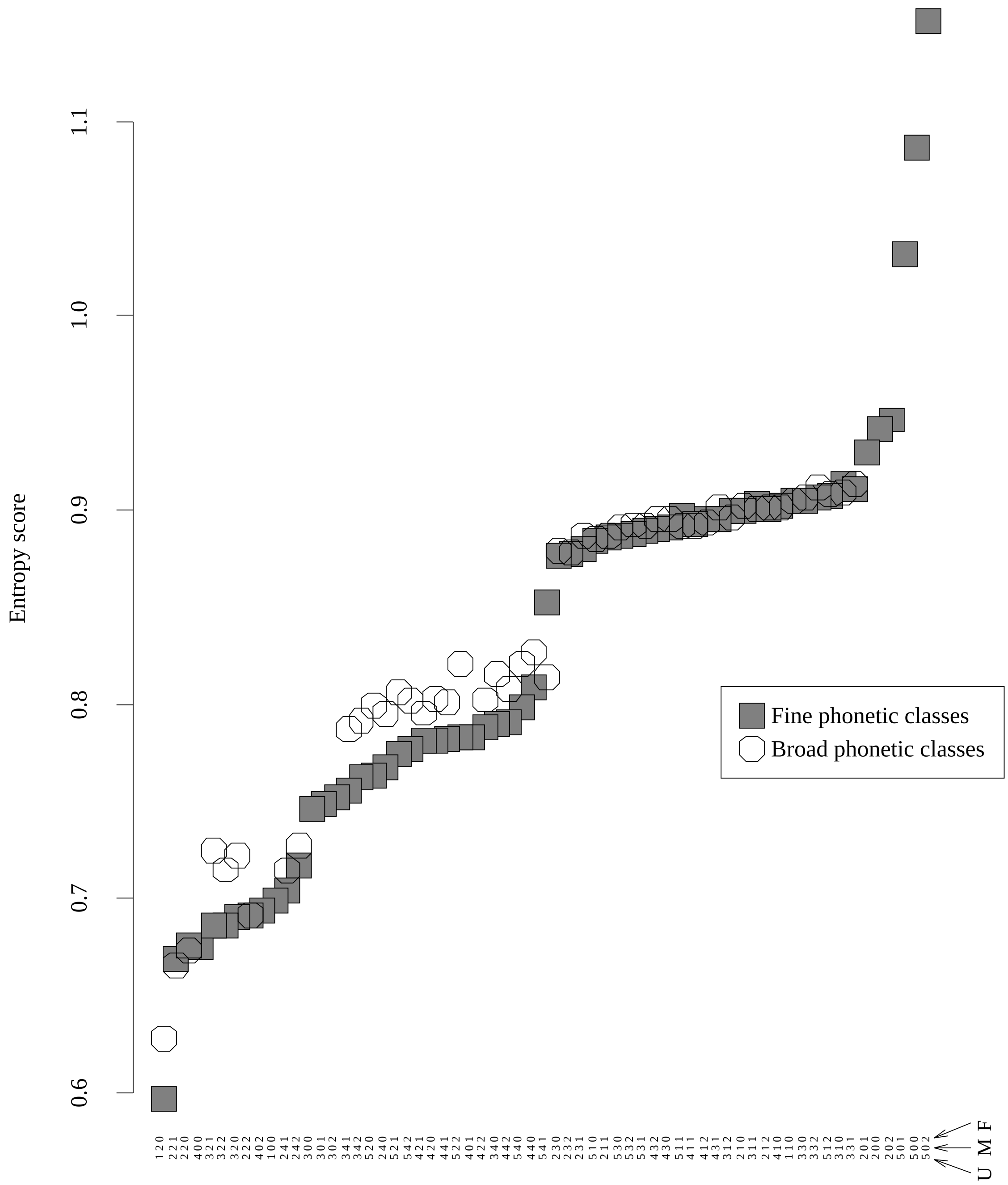

Figure 5.6: Energy feature evaluation: Entropy score ($P_{0,1}$)

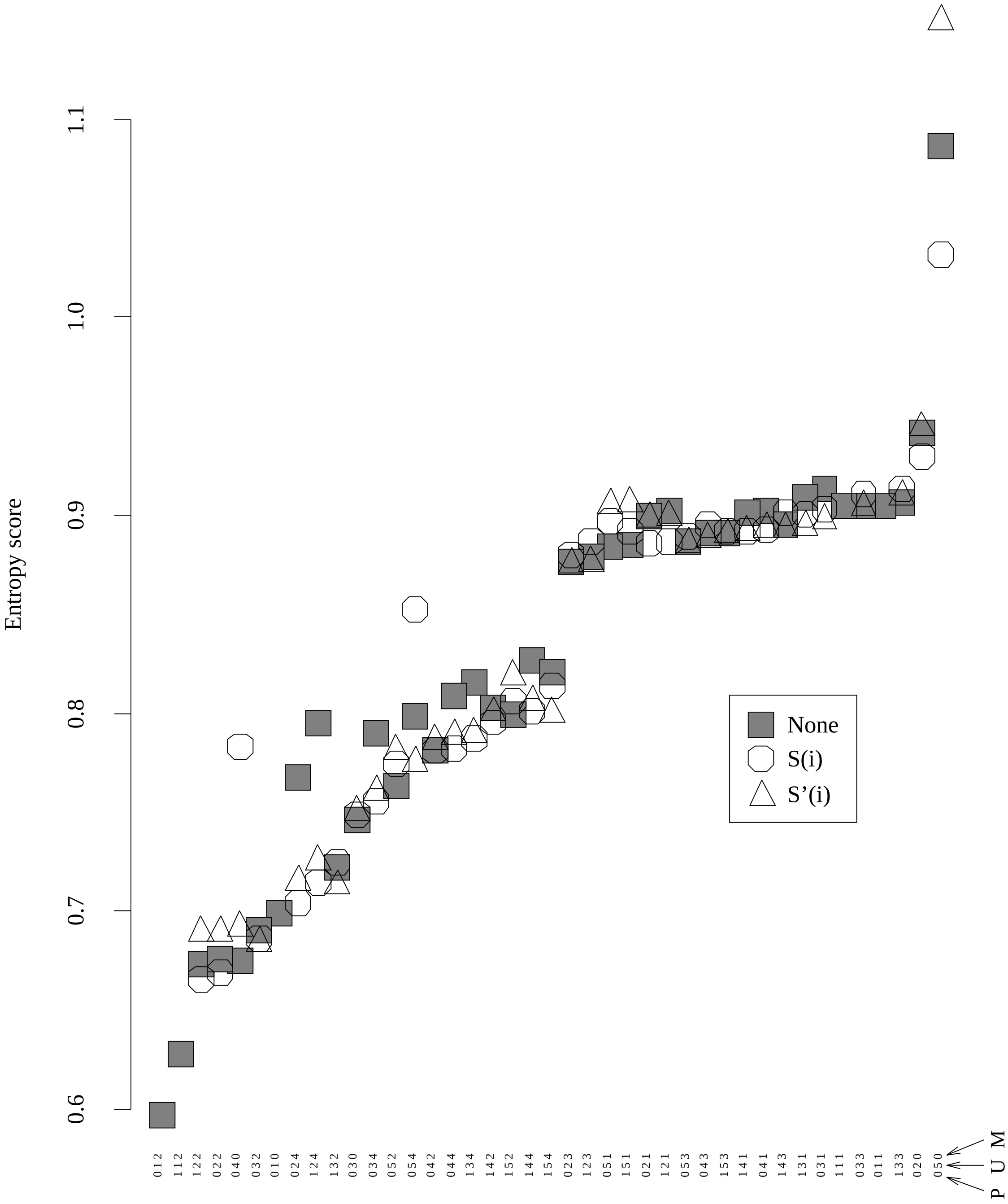

Figure 5.7: Energy feature evaluation: Entropy score ($F_{0,1,2}$)

energy and duration measures. An improvement in performance is generally obtained by using the maximum $z_{percentile}$ low-band energy measure M_2 rather than using the sum $z_{percentile}$ low-band energy measure M_4 of phones in a unit. Therefore this also provides evidence that sentential stress is correlated with an energy feature based on the peak of sonority in a syllable.

There is a significant difference in the entropy scores shown in Figure 5.6 between most $z_{percentile}$ based normalisations with statistics trained on fine phonetic classes and $z_{percentile}$ based normalisations trained on broad phonetic classes. Relatively little difference in the entropy scores is noticeable between the z_{mean} based normalisations trained on fine phonetic classes and broad phonetic classes. Since the optimal energy feature uses a $z_{percentile}$-transform and since the use of fine phonetic classes yields a lower entropy score, it is preferable to normalise for phone-types on a fine phonetic basis rather than on a broad phonetic basis. Two possible conclusions emerge from this with respect to the suitability of the training data — either the data provides a sample of the universal set of phone-types classified on a fine phonetic basis which represents the distribution of energy in such phone-types; or the broad phonetic classes shown in Table 5.1 form suboptimal groups.

Figure 5.7 reveals that normalisation for the number of phones in a unit has little effect on the performance of the Bayesian classifier for a given phonetic unit and low-band energy measure. Given that the optimal unit-type for the energy measure is the nucleus, no normalisation technique for the number of phones in a unit is applicable.

A final observation to make is in the comparison of the entropy score results for the duration and energy features. It is clear that the relative importance of duration and energy as cues to syllable prominence is dependent upon the choice of duration and energy features. No conclusions can therefore be drawn about the relative importance of these features in human perception of stress. Conclusions can only be drawn with respect to relative importance of the acoustic features investigated here for computer analysis of stress.

5.4 Optimisation of a relative energy feature

The peak-picking technique proposed in Section 4.5 is used to classify each syllable as either non-prominent or prominent depending on the magnitude of the low-band energy feature relative to its nearest neighbours. A syllable is classified as prominent if its low-band energy feature exceeds that of both its neighbours (end-points being inherently lower) and if such a local maximum in the low-band energy contour has a magnitude greater than some specified threshold K_{energy}. Syllables satisfying this criterion are labelled "*se*". A syllable is also classified as prominent if it corresponds with the maximum value of the low-band energy feature in the contour (independently of the threshold). Syllables corresponding to such maxima are labelled "*Se*". All other syllables are labelled "*ue*".

The low-band energy feature used corresponds to the syllable nucleus U_1 with $z_{percentile}$ low-band energy measures M_2 trained on fine phonetic classes P_0. Theoretically, this low-band energy feature is positive for prominent syllables and negative for non-prominent syllables. K_{energy} should therefore be optimal at 0.0. A plot of K_{energy} against the percentage of syllables which are assigned to the correct prominence category by the peak-picking technique (in a closed test) is shown in Figure 5.8. The optimal value of K_{energy} is -0.75.

There are a number of potential reasons as to why the optimal value of K_{energy} is not as expected. One possible reason is that the assumption made by the $z_{percentile}$ normalisation technique (that all of the lower low-band energy measures correspond to non-prominent syllables and that all of the higher low-band energy measures correspond to prominent syllables) is not valid. There are a number of prominent syllables for which the $z_{percentile}$ low-band energy measure is negative, and there are a number of non-prominent syllables for which the $z_{percentile}$ low-band energy measure is positive. By definition, there are the same number of prominent syllables with a negative $z_{percentile}$ low-band energy measure as there are non-prominent syllables with a positive $z_{percentile}$ low-band energy measure. It is empirically determined that reducing the threshold K_{energy} increases the number of non-prominent syllables with a $z_{percentile}$ low-band energy measure greater than the threshold more than it decreases the number of prominent syllables with a $z_{percentile}$ low-band energy measure less than the threshold.

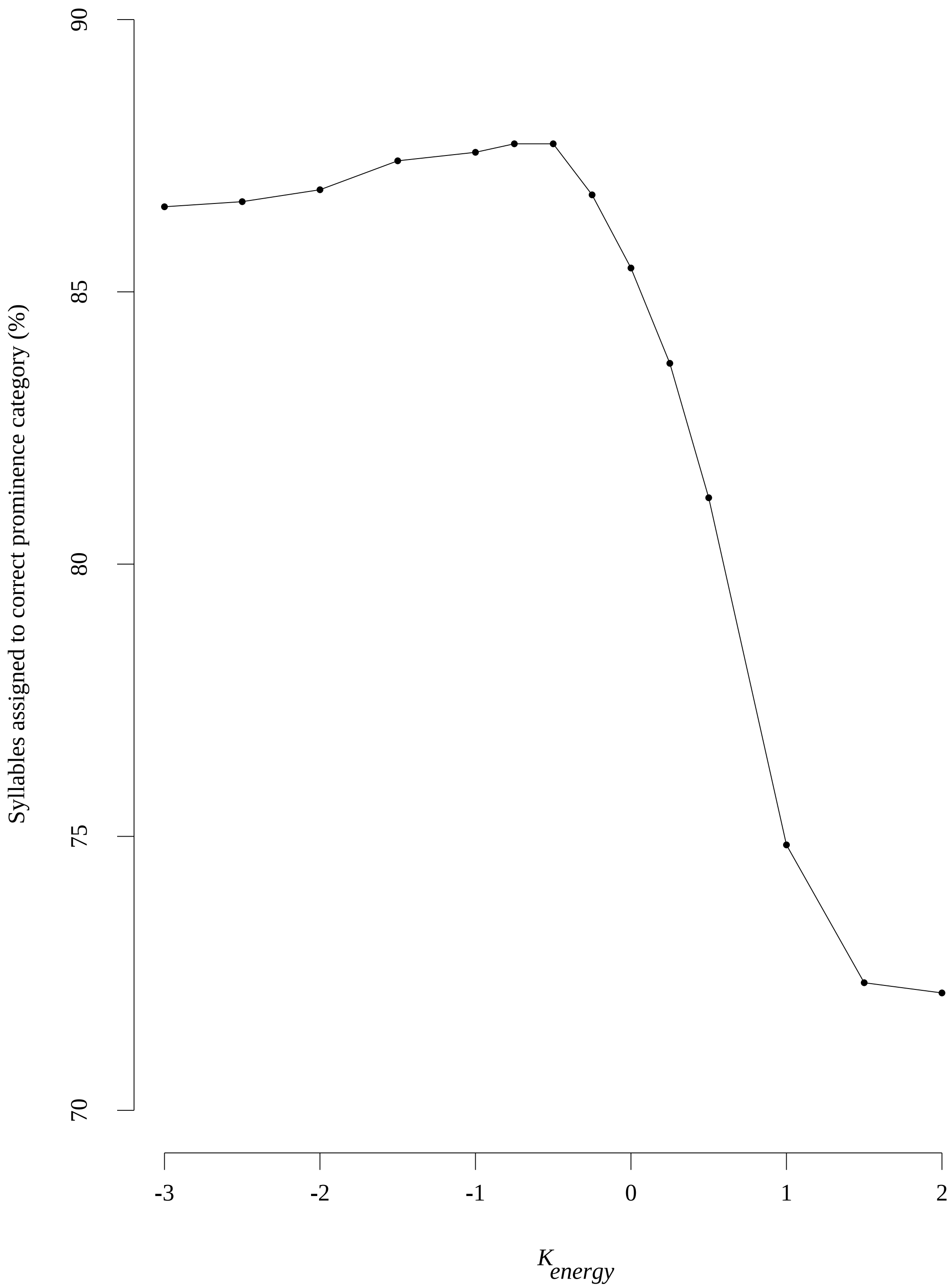

Figure 5.8: Optimisation of K_{energy} threshold

If it is assumed that a syllable is categorised as prominent only if its corresponding $z_{percentile}$ low-band energy measure is greater than K_{energy}, then a greater number of syllables should be incorrectly categorised (as either prominent or non-prominent) by reducing K_{energy}. This is contrary to the evidence provided in Figure 5.8. It is therefore unlikely that the optimal value of K_{energy} is less than expected because the assumption made by the $z_{percentile}$ normalisation technique is not strictly valid.

The above argument assumes that a syllable is categorised as prominent only if its corresponding $z_{percentile}$ low-band energy measure is greater than K_{energy}. The peak-picking technique, however, enforces the additional constraint that the syllable must correspond to a local maximum in the low-band energy contour in order to be categorised as prominent. This is a consequential factor in the optimal value of K_{energy} being lower than expected. There are more prominent syllables (22.3%) in the training data which correspond to a local maximum in the low-band energy contour than non-prominent syllables (8.0%). Optimisation of the threshold K_{energy} minimises the number of these (22.3% + 8.0%) syllables which are assigned to the incorrect prominence category. Reducing K_{energy} from 0.0 to -0.75 increases the number of prominent syllables labelled as "*se*" more than it decreases the number of non-prominent syllables labelled as "*ue*". In other words, the optimisation of K_{energy} is dependent upon the distributions of the $z_{percentile}$ low-band energy measure which correspond to local maximum in the energy contour for prominent and non-prominent syllables.

The optimisation of the threshold K_{energy} is also dependent upon the error measure used to evaluate the performance of the peak-picking technique. Here, the total-error rate is used because it is equally important to classify a given syllable correctly as non-prominent as it is to classify a given syllable correctly as prominent. If the classification criterion was to ensure that the same percentage of prominent syllables are correctly classified as non-prominent syllables, then an equal-error rate would be a more appropriate measure. A more robust measure (such as the entropy score used for the Bayesian classifier) for the performance of the peak-picking technique would produce a different optimal value for K_{energy}.

An entropy score cannot be produced by the peak-picking method. Its performance is therefore compared with the performance of the Bayesian classifier by examining the

		Peak-picking algorithm label			
		Se	se	ue	total
Hand Label	P	422 (7.9%)	1101 (20.7%)	575 (10.8%)	2098 (39.4%)
	NP	38 (0.7%)	236 (4.4%)	2947 (55.4%)	3221 (60.6%)
	total	460 (8.6%)	1337 (25.1%)	3522 (66.2%)	5319 (100.0%)

Correct classification rate = 4470/5319 (84.0%)

Table 5.2: Energy: Confusion matrix for peak-picking algorithm
(Se — syllable with maximum energy of contour; se — syllable corresponding to a local peak in the energy contour greater than K_{energy}; ue — unstressed syllable on the basis of energy; P — prominent syllable; NP — non-prominent syllable)

		Bayesian classifier label		
		P	NP	total
Hand Label	P	1650 (31.0%)	448 (8.4%)	2098 (39.4%)
	NP	516 (9.7%)	2705 (50.9%)	3221 (60.6%)
	total	2166 (40.7%)	3153 (59.3%)	5319 (100.0%)

Correct classification rate = 4355/5319 (81.9%)

Table 5.3: Energy: Confusion matrix for Bayesian classification
(P — prominent syllable; NP — non-prominent syllable)

percentage of syllables which are assigned to the correct prominence category in an open test. The levels of agreement between the prominence category of a syllable as transcribed by hand and as assigned automatically are shown in Tables 5.2 & 5.3 for the peak-picking algorithm and for the Bayesian classifier respectively. As well as giving an improvement of syllable prominence classification, the simple peak-picking algorithm has the additional benefits over the Bayesian classifier in requiring less computation (and hence it is faster) and not requiring a training stage.

5.5 Summary

The investigation described above aimed to determine a low-band energy feature whose distributions for prominent and non-prominent syllables have the greatest separability. The entropy score obtained by a Bayesian classifier was used as a robust measure of this separability for 113 different low-band energy features. The optimal low-band energy feature $E_{feature}$ of those investigated is given by:

$$E_{feature} = \frac{E_{nucleus} - \varphi_E^p(fine\_phone\_class)}{\sigma_E(fine\_phone\_class)} \tag{5.3}$$

where $E_{nucleus}$ is the low-band energy of a syllable nucleus, $\varphi_E^p(fine\_phone\_class)$ and $\sigma_E(fine\_phone\_class)$ are the p'th percentile and the standard deviation of phone-level low-band energies for phones classed on a fine phonetic basis respectively, and p is the percentage of non-prominent phones in the training data.

A simple peak-picking algorithm applied to this low-band energy feature is shown to give an improvement of syllable prominence classification over the Bayesian classifier, while requiring less computation and not requiring an additional training stage. In an open test, 84.0% of syllables are given the same prominence status by the peak-picking algorithm as when transcribed by hand.

The transcription (by hand) of the perceived prominence of each syllable in the training and test data cannot be regarded as definitive because there are instances when the discrete categorisation of a syllable as prominent or non-prominent is ambiguous. There is therefore some degree of error in the comparisons of prominence levels labelled by hand and labelled by the automatic procedures discussed above. The level of agreement between the prominence label assigned by the peak-picking algorithm and as transcribed by hand, is therefore regarded as a comparative measure only. By comparison to the duration feature $D_{feature}$ (Equation 4.6) with a 77.8% agreement level, it can be concluded that the low-band energy feature $E_{feature}$ (Equation 5.3) has a higher correlation with sentential stress for the speech of the one native English talker of the training and test data. It is important to emphasis that it cannot be concluded that the energy feature $E_{feature}$ necessarily has a higher correlation with sentential stress than the duration feature $D_{feature}$ for all speakers, or that energy has a greater importance than duration as a cue in the human perception of sentential stress.