Chapter 9

Application to Computer Aided Pronunciation Teaching

This Chapter describes how the integrated prosodic analysis system outlined in Section 8.4 can be applied to computer aided pronunciation teaching. Section 9.1 gives a general description of the framework of the SPELL[1] system (Lefèvre *et al.*, 1992; Hiller *et al.*, 1993) within which the proposed automatic prosodic analysis system can operate. An example of the operation of the automatic prosodic analysis within the framework of SPELL is described in Section 9.2.

A short discussion of possible extensions to the work presented in this thesis, is presented in Section 9.3.

9.1 Framework of the SPELL system

SPELL is a feasibility study to develop a tool for teaching prosodic and segmental aspects of pronunciation to non-native students of English, French and Italian. The SPELL workstation is an autonomous teaching system which is used by a student without the need for a language teacher to organise the pronunciation tasks. The system is aimed for use by students who do not need to be taught the basics of a foreign language as well as pronunciation. A student is provided with both audio and visual aids. The audio aids enable a student to listen to the pronunciation of utterances. The utterances presented to a student may be either the ideal target pronunciation spoken by a native

[1]SPELL (Interactive System for Spoken European Language Training) is a project supported by the European Community's ESPRIT programme, under contracts No.5192 & No.7153.

talker, or a resynthesised version of the student's voice with emphasis placed on a specific feature of its pronunciation. The visual aids enable a corrective diagnosis of a student's pronunciation to be presented to the student without the student requiring a knowledge of the underlying phonological theory or phonetic concepts on which the diagnosis is based. The objective of the SPELL system is to improve a student's intelligibility rather than to promote fluency in a foreign language or to encourage near-perfect mimicry of a native speaker.

Foreign language pronunciation is taught in the SPELL system under a teaching methodology called the DELTA paradigm. This paradigm constitutes four steps:

- **D**emonstate — A student is presented with audible utterances spoken by a native talker which highlight a particular feature of pronunciation.

- **E**valuate **L**istening — The student performs a number of listening tests to ensure that the pronunciation feature which was demonstrated in the previous step can be perceived by the student. If the student fails to exhibit competence, the demonstration phase is repeated.

- **T**each — A feature of pronunciation is taught to the student. This involves presenting an exemplar utterance to the student and requesting the student to reproduce the utterance whilst concentrating on the pronunciation feature of interest. Quantitative feedback and a corrective diagnosis are provided in order to modify inadequacies in the student's reproduction of the utterance.

- **A**ssess — The student's ability to generate a particular feature of pronunciation is evaluated. This may involve requesting the student to pronounce a number of previously unheard utterances containing the pronunciation feature of interest. If the student completely fails to exhibit competence, the demonstration phase is returned to. If the student shows only some ability, the teaching phase is repeated.

In teaching prosodic aspects of a language with this methodology, an automatic prosodic analysis system is required within the teaching and assessment phases of the DELTA paradigm.

There is an enormous degree of variation in the trajectories of *FØ* contours. It is therefore not possible to teach intonation simply by encouraging a student to reproduce exact imitations of isolated contours. A student needs to be familiarised with patterns or models of *FØ* trajectories which can be generalised to utterances of the same type, and needs to acquire an ability to select a particular pattern or model to convey a specific linguistic function (Rooney *et al.*, 1992). Therefore, in the teaching of intonation within the SPELL framework, prompts are used to provide a communicative context for the student and to elicit an appropriate pattern or model of *FØ* trajectories in the student's response. In addition, the segmental structure of an utterance is associated with the *FØ* trajectories which make up an intonation pattern. This is required to ensure that pitch accents are realised at clearly defined locations in an utterance.

The intonation teaching within the SPELL framework is primarily limited to two patterns of *FØ* trajectories; one pattern is typically used in syntacticly simple (subject-verb-object or subject-verb-adverb) declarative statements and *wh*-questions, and the other pattern is used in polar (*yes/no*) questions.

9.2 Example of automatic prosodic analysis applied to the SPELL framework

In this example, an intonation pattern commonly applicable to syntacticly simple declarative statements in English is being taught to a student whose mother-tongue is Italian. One of the chief characteristics of the intonation pattern is a falling pitch that is associated with the focus of the statement.

The exemplar utterance is, "I enjoy swimming these days." This statement is prompted with the question, "What sports do you do?" in order to provide a plausible context for the student and to elicit an emphasis on the word *swimming*. In contrast, the prompt question, "Do you hate swimming these days?" may be used to elicit an emphasis on the word *enjoy*.

In the teaching phase of the DELTA paradigm, the student is provided with a visual orthographic transcript of the prompt question. This is accompanied by a recording of the question which is played to the student. An audible version of the declarative statement pronounced by a native English speaker (referred to as the *teacher*) is then presented. The student may repetitively listen to the teacher's pronunciation of the statement. An automatic prosodic analysis of the teacher's utterance is also presented to the student as a visual aid. The student attempts to reproduce the statement with the same sentential stress pattern and intonation as exhibited in the ideal target pronunciation demonstrated by the teacher. An automatic prosodic analysis of the student's utterance is performed, by the system outlined in Section 8.4, using the same acoustic parameter normalisation statistics that are used in the analysis of the teacher's utterance. This ensures that a syllable pronounced by the student is automatically transcribed as prominent only if the relative variations in the duration and energy measures are analogous to those in the speech of a native English talker. The automatic prosodic analysis of the student's utterance is presented such that the beginning and end points are aligned with the analysis of the teacher's utterance (see Figure 9.1).

In terms of the configuration theory, the target pronunciation of, "I enjoy swimming these days," includes a falling nuclear pitch movement beginning on the first syllable of the word *swimming*. In the pronunciation demonstrated by the teacher, the word *I* is stressed but unaccented (part of the prehead) and the second syllable of *enjoy* is accented (the head). The word *days* is also stressed but unaccented (part of the tail).

Although the *FØ* trajectory within the student's pronunciation of *I* is falling (in contrast to the rising trajectory in the target pronunciation) the syllable is categorised as stressed but accented. The pronunciation of this word is therefore satisfactory and the difference in the shape of the trajectories is not brought to the attention of the student during the diagnosis of the pronunciation. There is, however, an underlying error in the phonetic realisation of the diphthong /aɪ/. The automatic segmentation of the Italian student's pronunciation of *I* is represented as two monophthongs [a i]. These segmental aspects of pronunciation are taught as a separate issue within the SPELL framework.

Two errors occur in the prosodic aspects of the student's pronunciation. The student successfully places the nuclear pitch movement on the first syllable of the word *swimming*. This is automatically detected as being an accented syllable, however a rising *FØ* trajectory is associated with its realisation. The student also accentuates *days* with a continuation rise. In addition, the word *days* is associated with a segmental pronunciation error in the diphthong /eɪ/. The automatic segmentation of the Italian student's pronunciation of *days* is represented as [d ɛ i z]. The student's attention can be drawn

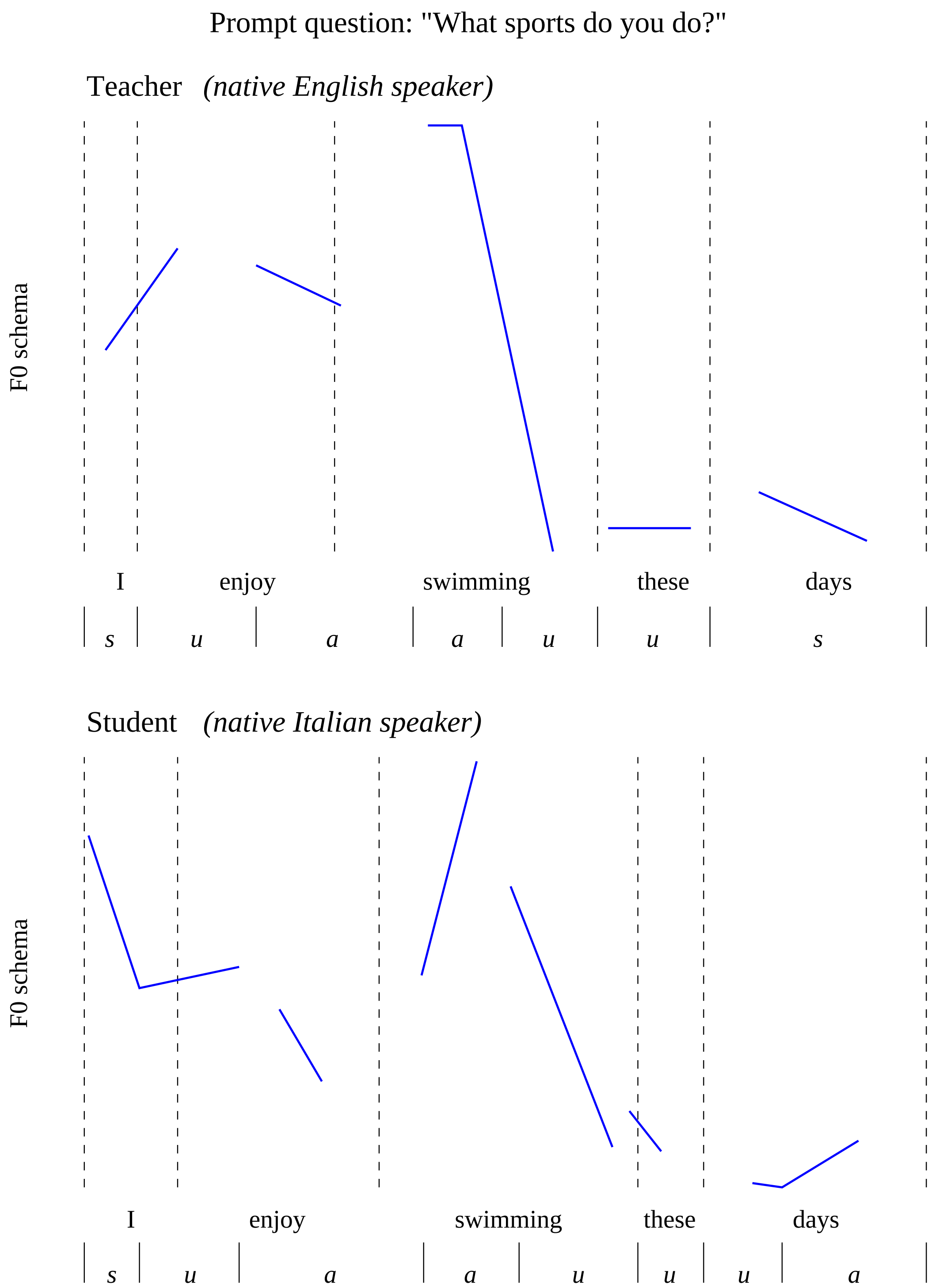

Figure 9.1: Example of prosodic analysis in pronunciation teaching

to the prosodic pronunciation errors by highlighting differences in the *FØ* schemata and the transcription of sentential stress.

The analyses of duration and energy measures are embedded in the sentential stress transcription. If a syllable is detected as being unstressed when the target pronunciation is to produce a prominent syllable, then a student may be instructed to make a specific syllable “longer” and/or “louder” in order to achieve the required prominence.

The acoustic-phonetic representation makes no references to any phonological theories. It is therefore suitable for language students who do not need a proficient knowledge of linguistic theory.

9.3 Further work

There are areas of the integrated prosodic analysis system where the underlying research may be extended. The aim of any further work would be to ensure that a foreign language learner is advised to alter the pronunciation of an utterance only in parts where it can convey a semantically contrastive function.

In the low-level analysis of the acoustic parameters, additional normalisation is required to compensate for contextual influences on the duration and energy of phones. A greater understanding of the effects of contextual assimilation on vowels may enable a vowel quality measure to be devised which is correlated with the perceived prominence of a syllable.

The unidimensional contributions of the acoustic parameters are integrated after the normalisation of non-prosodic aspects of speech. Integration of the acoustic parameters at a lower level of the analysis may enable, for example, duration measures to assist in the normalisation of energy measures or visa versa. A reduction in energy at the end of an intonational phrase may assist in normalising a duration measure for pre-pausal lengthening effects.

The abstraction of an *FØ* schema to a phonological representation (in terms of high and low tones) would enable the intonation of a student's speech to vary only in places where it does not convey a semantically contrastive function. Such abstraction must be reliable and the phonological representation must have a firm theoretical foundation. Moreover, the phonological representation must consider the mother-tongue of the foreign language learner.

Other aspects of prosodic structure need to be automatically transcribed in order to teach a foreign language learner rhythm and syllable timing differences between languages. The automatic transcription of break indices (which represent the degree of prosodic coupling between neighbouring words) would assist in the analysis of rhythm.

Finally, the automatic prosodic analysis of the speech needs to be performed as quickly as possible to prevent a user of a pronunciation teaching system from growing impatient. In the integrated prosodic analysis system presented here, there is a bottleneck in the analysis of an *FØ* contour. The least median of squares regression is time consuming even though a Monte Carlo type speed-up technique is applied. Methods of reducing the computational work-load involved in the analysis may be worthy of investigation.

Chapter 10

Summary and Conclusions

This thesis presents research towards the automatic analysis of prosodic events in utterances of English spoken by native and non-native talkers. The research concentrates on the automatic analysis of speech to generate an acoustic-phonetic representation of sentential stress and intonation. The acoustic-phonetic representation of prosody is demonstrated as a tool for computer aided pronunciation teaching.

It is argued in Chapter 2 that prosodic aspects of speech need to be explicitly taught to students who wish to communicate competently in a foreign language. In summary, the argument is that prosodic aspects of speech need to be taught to a student because prosody has functional differences between languages, and the acoustic-phonetic realisation of prosodic aspects of speech also differs between languages. Explicit teaching is required because a student's comprehension and intelligibility can be influenced by prosodic features. Inter-language differences are illustrated in exemplar languages.

In teaching the prosody of a foreign language, it is suggested that a formalised phonological theory of prosody should not be explicitly taught to a foreign language learner. This suggestion is made by taking the analogy that a formalised grammar is usually not explicitly taught to a student who needs to acquire a knowledge of the grammatical structures of a foreign language. However, phonological theories of prosody are needed to design courseware by which a student can learn stress patterns and intonation.

Ideally, the prosodic composition of the speech of a foreign language learner should be described in a manner which allows any linguistically significant deviations from a near-native pronunciation to be determined. The theories of intonational phonology aim to describe a minimal set of pitch variations which are semantically contrastive. A comparison of the prosodic composition of a student's speech and a native's speech can ideally be made by describing the prosody with phonological models. Such a comparison would allow a student's pitch to differ from that of a native speaker only at times where it does not carry any semantically contrastive function. A corrective diagnosis could then be offered to a student in order to aid the student's pronunciation.

However, it is argued in Chapter 2 that phonological representations of prosodic structure are language-specific. The prosodic composition of a student's speech can be highly influenced by mother-tongue interference and by a student's stereotypic images of prosody in the target language, thereby rendering a language-specific phonological representation of prosody inappropriate. Moreover, languages vary in the way acoustic characteristics of speech are modified to manifest prosodic aspects of speech. It is therefore proposed in Chapter 2, that the only secure means to describe prosody for foreign language teaching lies in an acoustic-phonetic representation. It is assumed that a higher level of analysis can be used to determine which of the detectable phonetic differences do have and do not have a semantically contrastive function in a specific language.

In order to derive an acoustic-phonetic representation of prosody from a speech signal, it is necessary to identify acoustic parameters which are correlated with prosodic aspects of speech. In Chapter 3, a review of research conducted to identify acoustic parameters indicates that duration, energy, fundamental frequency and vowel quality are the key correlates of prosodic features. There is relatively little consideration made in the literature to the effects of vowel quality on the perception of (lexical or sentential) stress, compared with the extensive research relating duration, energy and $F\O$ with stress.

Studies of the acoustic correlates of stress propose a variety of different features for duration and energy. Duration features include the duration of the vocalic portion of a syllable and the duration of an entire syllable. Energy features include the peak amplitude in a syllable, the amount of fall from a peak in the amplitude envelope of a syllable, the integral of the peak-to-peak amplitude over the vocalic portion of a syllable, the average low-band syllable energy, and the maximum low-band vowel energy.

Chapters 4, 5, 6 & 7 concentrate on optimising the extraction of acoustic correlates required for the prosodic analysis of English. The duration and energy measures of phonetic units are dependent upon the definition of the phonetic units. A number of different phonetic units related to syllable structure are investigated. The phonetic units investigated are the syllable nucleus, the syllable rhyme, the syllable lhyme, the entire syllable and a nucleus-to-nucleus unit. Duration and energy measures are made from these units. It is asserted that the duration and energy measures are influenced both by prosodic and by non-prosodic aspects of speech. Normalisation techniques are applied to the measures in order to compensate for a number of non-prosodic factors, including variations in the measures which are due to phone-type, syllable structure, syllable length in terms of number of phones, and the prominence level of syllables within the training data. A novel $z_{percentile}$-transform is proposed for the normalisations.

There are 213 different duration features investigated in Chapter 4. It is shown that factors other than sentential stress cause the duration of phones to the right of a syllable nucleus (but within the same syllable) to vary more than the duration of phones to the left of the nucleus. The syllable lhyme is defined as the unit which constitutes the onset and the nucleus of a syllable; hence the syllable lhyme contains the parts of a syllable which are least affected by factors other than sentential stress. The optimal duration feature, of those investigated, is related to the syllable lhyme with sum $z_{percentile}$ duration measures trained on broad phonetic classes without smoothing of the phone-level duration contour and with normalisation of the number of phones in the syllable lhyme. The optimal energy feature, out of 113 energy features investigated, is related to the syllable nucleus with $z_{percentile}$ low-band energy measures trained on fine phonetic classes.

A simple peak-picking algorithm is used to determine the contribution made by each acoustic parameter to the degree of syllable prominence. The optimal duration feature is used to label syllables as either prominent or non-prominent and yields a 77.8% agreement level with prominence labels based on human perception (in an open test). The optimal energy feature is used to label syllables as either prominent or non-prominent and yields an 84.0% agreement level with the prominence labels based on human perception.

A number of measures of vowel quality are proposed in Chapter 6 which are based on the assumptions that prominent syllables are well articulated and are less affected by contextual assimilation than non-prominent syllables, and that these properties are reflected in the nuclei of syllables. A Bayesian classifier is used to obtain a measure of how close a vowel is to its respective vowel target model. This distance measure is used to label syllables spoken by a native English speaker as either prominent or non-prominent and yields a 71.6% agreement level with prominence labels based on human perception. It is argued that the distance measure cannot be used as a correlate of stress in utterances spoken by non-native speakers of English because of vowel pronunciation errors. A number of alternative measures are investigated as features to characterise the degree of vowel stability. The average Euclidean distance of frame-to-frame log-power spectra is used as a measure of the degree of spectral change in a vowel due to contextual assimilation. This vowel stability measure is used to label syllables as either prominent or non-prominent and yields a 68.8% agreement level with prominence labels based on human perception. A higher level of agreement is obtained by using a $z_{percentile}$-transformed, normally-distributed random variable as an input parameter to a peak-picking algorithm. The application of the vowel stability measure as a correlate of stress is therefore inappropriate. The use of phonological rules to transcribe syllables as non-prominent under certain conditions, is also dismissed.

Fundamental frequency is identified as an acoustic correlate of stress, and is used in algorithms (reviewed in Section 3.5) to automatically transcribe the location of prominent syllables in isolated words and in connected speech. In addition, fundamental frequency is principally related to the intonation of an utterance. Nevertheless, many of the reviewed algorithms use a measure of *FØ* as a key acoustic correlate of stress with little or no consideration of the intonational role of *FØ*. However, *FØ* can be used as a secondary cue to the location of prominent syllables in connected speech because pitch accents are observed to fall on prominent syllables.

The fundamental frequency of a speech waveform must be determined as an initial process for prosodic analysis, since it is associated with both sentential stress and intonation. The performance of *FØ* determination algorithms (a selection of which are reviewed in Section 3.2) must be considered with respect to their application in systems of prosodic analysis, and errors arising from the malfunctions of FDAs must be prevented from propagating into the subsequent prosodic analysis of speech. A number of modifications to the super resolution *FØ* determination algorithm are proposed in

Section 7.1. The modifications reduce the occurrence of errors involved in the extraction of *FØ* such that it is optimised for prosodic analysis, relative to the selection of other algorithms designed to determine *FØ*. A novel de-step filter is proposed to post-process an *FØ* contour generated by an FDA in a way which further reduces the occurrence of discontinuity errors commonly observed in *FØ* contours.

The processing of fundamental frequency is based on the principle that an *FØ* contour has a composite structure (Section 3.4). In particular, the *FØ* contour of an utterance is affected by the talker's anatomy and physiology, the segmental content of an utterance, cycle-to-cycle jitter (*FØ* perturbation) and errors involved in its determination from the speech waveform. The piece-wise stylisation of an *FØ* contour aims to eliminate microprosodic variations. An algorithm to stylise an *FØ* contour into linear piece-wise sections with respect to the syllables of an utterance, is illustrated in Section 8.2.1. 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 *FØ* trajectories of a schema are associated with syllables and a pitch accent decision filter analyses the trajectories to provide judgements of syllable accentuation. The integration of syllable information with the stylisation and schematisation of an *FØ* contour is an innovative approach. 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 judgements of syllable accentuation provided by the pitch accented decision filter yields a 72.7% agreement level with syllables labelled as pitch accented by hand.

Syllables are used in the extraction of duration and energy features and in the processing of *FØ*. The syllabic domain also ties together the information embedded within these three acoustic parameters. Furthermore, prosodic aspects of speech are described in a syllabic domain. The automatic syllabification of speech is therefore an important part of prosodic analysis. Algorithms reviewed in Section 3.3 are devised to partition a speech signal into syllable-sized units by using measures of sonority based on low-band energy and by using spectral characteristics to locate boundaries between adjacent sonorants. The reviewed syllabification algorithms are applied to speech recognition systems. Reliable information about the segmental content of an utterance is therefore not available to them. In the application of prosodic analysis for computer aided pronunciation teaching, however, reliable segmental information is available and may be used to enhance the syllabification process.

An algorithm is proposed in Section 8.1.1 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.

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 pitch accent decision filter provides judgements of syllable accentuation. 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, in a syllabic domain. This 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.

The former algorithms for prosodic analysis reviewed in Chapter 3 have not addressed the problems of determining the fundamental frequency of speech and have not comprehensively addressed the need to normalise for non-prosodic variations in the acoustic parameters which are used in the prosodic analysis. The system of prosodic analysis described in this thesis addresses the problems related to the determination of the fundamental frequency of speech and focuses on techniques of normalising for variations in acoustic parameters which are due to non-prosodic aspects of speech.

The integrated prosodic analysis system proposed in this thesis is shown, in Chapter 9, to produce prosodic descriptions which are useful in comparing the prosodic aspects of the speech of a non-native learner of English with the speech of a native English talker.

Appendix A

Least Median of Squares Regression

A highly robust method of fitting a linear regression model to a set of observations (including spurious samples) is least median of squared residuals (LMedS) regression analysis; introduced by Rousseeuw (1984).

Consider a set of N observations, $\{i, y_i\}_{i=1}^{N}$.

For all pairs of observations, $\{P_u, P_v \mid u \in 1, \ldots, N; v \in 1, \ldots, N; P_u = (u, y_u)\}$ a linear model is computed. There are $K = {}_N\mathbf{C}_2 = \frac{N!}{2!(N-2)!}$ such pairs of observations. The linear model calculated from the k'th pair of observations is represented as,

$$z'_k(i) \;=\; a_k.i + b'_k \text{ where } \begin{cases} a_k = \dfrac{y_u - y_v}{u - v} \\ b'_k \text{ need not be calculated} \end{cases} \tag{A.1}$$

The point of intersection, b'_k, is disregarded since it is not a robust value. Let a simplified version of the model $z_k(i) = a_k.i$. The set of residuals (offset by b'_k) $\{\alpha_i = y_i - z_k(i)\}_{i=1}^{N}$ is computed and a mode-seeking algorithm is applied to it.

One method of estimating the mode (the sample value which occurs most frequently) of a distribution is named "estimating the rate of an inhomogeneous Poisson process by $(N/2)$'th waiting times," (Press *et al.*, 1988, chapter 13.3). The set $\{\alpha_i\}_{i=1}^{N}$ is initially sorted into ascending order to give $\{\hat{\alpha}_i\}_{i=1}^{N}$. The smallest window is then sought which accommodates half of the data ($N/2$ points) in the distribution of sorted offset residuals.

That is to say, δ_k is determine as,

$$\delta_k \;=\; \frac{1}{2}. \min_{i=1\rightarrow\lfloor N/2\rfloor} \left(\hat{\alpha}_{i+\lceil N/2\rceil} - \hat{\alpha}_i\right) \tag{A.2}$$

50% of the observations exist within the window and 50% exist outside the window. The evaluation of Equation A.2 is equivalent to shifting this window through the set of sorted offset residuals, $\{\hat{\alpha}_i\}_{i=1}^{N}$, to locate the area of highest density within its distribution. Let this point of highest density be located when $i = j$. A robust value of the point of intersection, b_k, is given by the mid-point of the window when $i = j$,

$$b_k \;=\; \frac{\hat{\alpha}_{j+\lfloor N/2\rfloor} + \hat{\alpha}_j}{2} \tag{A.3}$$

The magnitude of the residuals within the window is always less than δ_k. Let $\hat{\alpha}_{[i]}$ denote the 50% of sorted offset residuals which reside within the window. Thus, $\left(b_k - \hat{\alpha}_{[i]}\right)^2 < \delta_k^2$. Since $\hat{\alpha}_{[i]}$ corresponds to 50% of the data, then δ_k^2 is the median[1] of the squared residuals by definition. Hence $\delta_k = \sqrt{\mathrm{med}(b_k - \hat{\alpha}_i)^2}$.

This process of determining a_k (from Equation A.1), b_k (from Equation A.3) and δ_k (from Equation A.2) is repeated for all k, giving $\{a_k, b_k, \delta_k\}_{k=1}^{K}$. It is then determined which pair of observations gives the minimum median squared residuals.

$$k_{min} \;=\; \operatorname*{argmin}_{k=1\rightarrow K} \left(\delta_k^2\right) \tag{A.4}$$

If significant Gaussian noise is present simultaneously with numerous outliers, then the LMedS estimates, $\{a_{k_{min}}, b_{k_{min}}\}$ become less reliable. Thus, the LMedS algorithm is used for outlier detection and least (mean of) squares regression analysis is employed.

The data, modelled by $z''_{k_{min}}(i) = a_{k_{min}}.i + b_{k_{min}}$, has a standard deviation, $\hat{\sigma}$, from the model given by (Rousseeuw & Leroy, 1987),

$$\hat{\sigma} \;=\; 1.4826\left(1 + \frac{5}{N-2}\right).\delta_{k_{min}} \tag{A.5}$$

[1]The *median*, φ_{50}, of a set of samples is a number such that at least 50% of the sample values are smaller than or equal to φ_{50} and also at least 50% of those values are larger than or equal to φ_{50}. If there is more than one such number (in which case there will be an interval of them), the median is defined as the average of the numbers (midpoint of the interval).

All points for which the magnitude of the residual, $|y_i - (a_{k_{min}}.i + b_{k_{min}})|$, is less than or equal to $2.5\hat{\sigma}$ are taken to be *inliers* and all others samples are taken to be *outliers*. A linear model is then determined for all the inliers using least (mean of) squares regression analysis (see Press *et al.*, 1988, chapter 4).

The median computations are applied over the whole data set. Therefore, the LMedS estimator remains reliable if less then half of the observations are contaminated by outliers; ie. it has a 0.5 asymptotic breakdown point. The time-complexity of the algorithm, however, is high. There are $O(n^2)$ pairs and for each of them a sorting $O(n \log_2 n)$ is required. Thus, the LMedS algorithm has a time-complexity of the order $O(n^3 \log_2 n)$.

It is possible to reduce the time-complexity of the algorithm by applying a *Monte Carlo* type speed-up technique. When ${}_N\mathbf{C}_2$ is very large, a random subset of pairs of observations (m pairs) are taken, reducing the time-complexity to $O(mn \log_2 n)$. In order for the LMedS estimates to be reliable, it is necessary for at least one candidate model, z_k to carry the correct parameter values. If this condition is not met, the LMedS estimator loses its robust properties. Note, however, that in practice, more than one such candidate model is preferred, because noise will be present amongst the inliers.

Let ϵ be the fraction of data contaminated by outliers, and m be the number of randomly selected pairs. The probability of all m pairs containing at least one outlier is $P = [1-(1-\epsilon)^2]^m$ and thus $Q = 1 - P$ is the probability that at least one of the m pairs contains no outliers. For $\epsilon < 0.5$, taking $m = 3000$ pairs (${}_N\mathbf{C}_2 > 3000$ for $N > 77$), gives $P \approx 0$, $Q \approx 1$. Hence, there is a high probability of maintaining a high breakdown point when the Monte Carlo type speed-up technique is applied (Rousseeuw & Leroy, 1987).

Appendix B

Bayesian Classification

Consider a number of categories, $\{C_i \mid i \in 1, \ldots, I\}$, for a population of objects where each object is characterised by a set of p features. The aim of a classifier is to partition the p-dimensional feature space into regions, one region for each category, with the minimum probability of error.

A classification rule can be regarded as being optimal if the proportion of objects that are misclassified to any of the possible categories other than that which it belongs, is minimised by its use. Bayes' rule aims to minimise this classification error by assigning an object to a category with the highest *conditional probability*.

Bayes' rule (see, for example, Duda & Hart, 1973; James, 1985): An object is assigned to category C_i if,

$$P(C_i|\boldsymbol{x}) \; > \; P(C_j|\boldsymbol{x}) \qquad \forall\, j \neq i \tag{B.1}$$

where $\boldsymbol{x}$ is a vector of p feature measurements which characterise the object.

If the conditional probability $P(C_i|\boldsymbol{x})$ is a maximum for more than one category, the object is randomly assigned to one of the categories for which $P(C_i|\boldsymbol{x})$ is a maximum.

In practice, it is possible to estimate the probability $P(\boldsymbol{x}|C_i)$ of getting a particular set of features $\boldsymbol{x}$ given a sample of data from a particular category C_i. Bayes' theorem (see, for example, James, 1985) relates the *a posteriori* probability $P(C_i|\boldsymbol{x})$ to the probability $P(\boldsymbol{x}|C_i)$ and the *a priori* probability $P(C_i)$ of category C_i existing in the population:

$$P(C_i|\boldsymbol{x}) \; = \; \frac{P(\boldsymbol{x}|C_i).P(C_i)}{\sum_{\text{all } i} P(\boldsymbol{x}|C_i).P(C_i)} \tag{B.2}$$

Bayes' rule can be rewritten by substituting Equation B.2 into Equation B.1 as:

Assign an object to category C_i if,

$$P(\boldsymbol{x}|C_i).P(C_i) \;>\; P(\boldsymbol{x}|C_j).P(C_j) \qquad \forall\, j \neq i \tag{B.3}$$

This rule is optimal for classification but requires a large training sample set to determine $P(\boldsymbol{x}|C_i)$ $\forall i$ and the *a priori* probabilities $P(C_i)$ $\forall i$ (which may be estimated by the proportion of each category in the sample set).

Let us assume that the measurements being used for classification form a multivariate Normal distribution, thus,

$$P(\boldsymbol{x}|C_i) \;=\; \frac{1}{(2\pi)^{p/2}.|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right] \tag{B.4}$$

where,

$$\begin{aligned}\boldsymbol{\Sigma}_i &= E\left[(\boldsymbol{x}-\boldsymbol{\mu}_i).(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\right] \text{ (covariance matrix)}\\ \boldsymbol{\mu}_i &= E\left[\boldsymbol{x}_i\right] \text{ (mean vector)}\end{aligned}$$

Substituting Equation B.4 into Equation B.3, taking the natural logarithm, cancelling common terms, multiplying by -2 and reversing the inequality, gives Bayes' rule for multivariate Normal data as: Assign an object to category C_i if,

$$Q_i(\boldsymbol{x}) \;<\; Q_j(\boldsymbol{x}) \qquad \forall\, j \neq i \tag{B.5}$$

where the *quadratic discriminant score*, $Q_i(\boldsymbol{x})$, is given by,

$$Q_i(\boldsymbol{x}) \;=\; \ln|\boldsymbol{\Sigma}_i| + (\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - 2\ln(P(C_i)) \tag{B.6}$$

The above rule enables an object to be classified to a category on the basis of a vector of features, $\boldsymbol{x}$, given the mean values of the features, a covariance matrix of the features for each of the possible categories, and the *a priori* probability of an object existing in a given category (all of which can be determined from a training set). If the values of the features form a Normal distribution, the resultant classification will have a minimum probability of error.

Note that,

$$\exp\left[-\frac{Q_i(\boldsymbol{x})}{2}\right] \;\propto\; P(\boldsymbol{x}|C_i).P(C_i) \tag{B.7}$$

Thus, if by applying Bayes' rule expressed in Equation B.5, an object is assigned to category C_i, then the *a posteriori* probability that the object has been assigned to the correct category (substituting Equation B.7 into Equation B.2),

$$P(C_i|\boldsymbol{x}) \;=\; \frac{\exp\left[-\frac{Q_i(\boldsymbol{x})}{2}\right]}{\sum_{\text{all } j}\exp\left[-\frac{Q_j(\boldsymbol{x})}{2}\right]} \tag{B.8}$$

The measure adopted to evaluate the performance of the classification is the entropy (Shannon & Weaver, 1949; Pierce, 1962) associated with the task of correctly identifying the true category of the input object[1]. An entropy score can be determined from the *a priori* and the *a posteriori* probabilities of each class to give an indication of the effectiveness of the classification. The entropy score, H, is given as,

$$H \;=\; \sum_{\text{all } i}\left[P(C_i).\frac{1}{N_i}\sum_{n=1}^{N_i} -\log_2 P(C_i|\boldsymbol{x}_{i,n})\right] \tag{B.9}$$

where N_i represents the number of objects of category C_i in the sample set and $\boldsymbol{x}_{i,n}$ is the n'th feature vector characterising an object in the sample set which is known to be of category C_i. The *a posteriori* probabilities, $P(C_i|\boldsymbol{x}_{i,n})$, are calculated from Equation B.8.

If the *a priori* probabilities, $P(C_i)\ \forall i$, are estimated from the proportion of each category in the training sample set,

$$P(C_i) \;=\; \frac{N_i}{N} \tag{B.10}$$

where N is the total number of objects in the sample set.

Substituting Equation B.10 into Equation B.9, the entropy score reduces to,

$$H' \;=\; -\frac{1}{N}\sum^{N}\log_2 P(C_i|\boldsymbol{x}_i) \tag{B.11}$$

[1] An entropy measure has previously been used to evaluate the performance of a Hidden Markov Model based speech recognition system (McInnes *et al.*, 1989). The entropy score presented here has been derived with the assistance of Dr. Fergus McInnes.

Thus, the entropy score, H', is the average negative logarithm (to base 2) *a posteriori* probability of an object being assigned to the correct category. The entropy is the theoretical number of bits of additional information required by any process whose task is to derive the category of the input object without error, subsequent to the Bayesian classifier. A lower entropy score indicates a Bayesian classification of better performance.