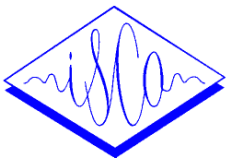

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

PROSODIC FEATURES FOR AUTOMATED PRONUNCIATION IMPROVEMENT IN THE SPELL SYSTEM[1]

Edmund Rooney, Steven M. Hiller, John Laver and Mervyn Jack

Centre for Speech Technology Research, University of Edinburgh
80, South Bridge, Edinburgh EH1 1HN, Scotland

ABSTRACT

This presentation describes the analysis of the prosodic features of intonation and rhythm within the SPELL system, a workstation for the automated assessment and improvement of English, French and Italian pronunciation by non–native speakers. For each language, a limited range of phonologically distinctive intonation contours has been chosen. These contours are characterized using a system of *pitch anchor points* and *pitch trajectories*. A similarity metric evaluates the acceptability of a student's intonation using a smoothed fundamental frequency contour and an automatic segmentation of the student's utterance derived by a Hidden Markov Model (HMM) technique. The analysis of rhythm concentrates on the control of salience relationships within an utterance (the contrast between *weak* and *strong* syllables) using the parameters of vowel quality and duration only. Judgements are expressed in terms of the *weak–strong* syllable contrast, obtained indirectly using the HMM segmenter with phrase models which allow for errors on these two parameters.

1 INTRODUCTION

The first phase of SPELL (Interactive System for Spoken European Language Training) is a two year ESPRIT project which began in September 1990. Its main aim is the development of tools to be used in the automated assessment and improvement of non–native pronunciation. This is a feasibility study involving British English, French and Italian which will lead to an initial demonstrator system. The technical objectives of the project are to develop methods for analyzing the characteristics of speech produced by non–native speakers, to develop metrics for identifying differences between a non–native speaker's pronunciation and a model offered by the system, and to provide user–friendly feedback which will help to improve pronunciation.

The SPELL workstation will be an autonomous teaching system for use by foreign language speakers without sophisticated linguistic or phonetic knowledge. Visual displays and audio feedback will help students to master relevant concepts without requiring expert knowledge. A minimal set of fully–defined courseware has been developed in key areas for the demonstrator system, using the DELTA paradigm devised on the project (see [6]). The immediate aim of this courseware is an improvement in the intelligibility of the student, since for the majority of students this is a more practical objective than the acquisition of fully native pronunciation [7], [8].

The *prosodic* features component of the SPELL project covers the parameters of intonation and rhythm. A unifying analytical approach for intonation is developed for the three target languages, making use of the concepts of *pitch anchor points* and *pitch trajectories*. A more limited approach to rhythm is being attempted, by manipulation of the acoustic features of vowel quality and segmental duration.

The *segmental* features component covers aspects of vowel and consonant articulation. Work so far has concentrated on vowels, using a distinctive feature approach to characterize non–native vowel production (see [1] in these proceedings).

The main technical innovation behind SPELL is the departure from the practice used in some existing automatic systems of requiring exact acoustic copying. Instead, techniques of normalization and segmentation are used to achieve comparisons at a level of perceptual and ultimately phonological equivalence.

2 PROSODIC FEATURES

Prosodic features are those which operate over stretches of speech longer than the single segment or phoneme, and here include intonation and rhythm. **Intonation** is generally defined as the manipulation of pitch for linguistic and paralinguistic purposes at a level above that of the segment. The **rhythm** of an utterance is given by the patterning in time of the segments, syllables and stresses.

The acquisition of prosodic features is important in language learning, since incorrect prosody can hinder communication even more than segmental errors. Mistakes in intonation, for example, may give a completely false impression of a speaker's attitude, while incorrect rhythm can make it difficult for listeners to process the segmental content of the learner's speech.

3 THE ANALYSIS OF INTONATION IN THE SPELL SYSTEM

3.1 The phonology of intonation

It is not sensible to teach intonation simply by direct imitation of target utterances, since actual pitch contours can vary enormously. Moreover, the ability to mimic a given pitch contour exactly does not guarantee any generalization of pitch use for intonation within a language, since the linguistic relevance of the contour springs partly from its integration with the segmental performance and partly from its relative placement in the pitch range of the speaker concerned. What the pupil requires is the successful acquisition of a pattern or model which can be generalized to other utterances of the same type or for the same

1. This project is supported by the European Community's ESPRIT programme, under contract no. 5192.

purpose, and the ability to choose from a set of such models to convey contrasts of meaning or emphasis.

Establishing a set of models for teaching requires a phonological analysis of intonation in each language. Available analyses in the literature vary considerably in their depth of treatment and in their general approach to the problem. One major difference in treatment, for example, is the extent to which the intonation contours of a language are seen as whole tunes (e.g. [9] for French, [2] for Italian), or as sequences of smaller elements (such as Halliday's "tone" based analysis of English [5]). This makes comparisons across languages rather difficult.

However, some general principles are clearly common to all three languages. First, pragmatic linguistic functions such as statements and questions are differentiated by opposing pitch movements, falling pitch typically being associated with statements and rising pitch with certain types of question. Second, pitch movements are also related to *rhythmical* structure by the marking of accented syllables. Finally and most importantly, intonational pitch movements are *anchored* to the segmental structure of the utterance. That is, the pitch movements which give a particular contour its characteristic shape do not occur at arbitrary points in the segmental sequence, but at clearly defined locations.

This last principle is extremely important for language teaching: a system which cannot analyse students' intonation contours with reference to the segmental structure of their utterances will not be able to work with the rules required for the correct placement of intonational features. Students will then be unlikely to succeed in generalizing the contours they have learnt to new utterances.

3.2 The Analysis of Intonation

In the SPELL system, a practical phonetic approach to the description and analysis of intonation has been adopted. This approach allows the essential intonational features of all three languages to be described using a common terminology and analyzed with a single similarity metric. The relationship between the pitch contour and the segmental sequence is central to this analysis.

For each language, the discussion is being limited to the two primary intonation functions which give significant coverage for learners, namely the marking of statements/*wh*–questions (*qu*–questions in French and Italian) and the marking of polar ("yes/no") questions.

For each function in each language, a single contour has been chosen as the model for teaching. Each contour is characterised by means of a set of *pitch anchor points* and a corresponding set of *pitch trajectories*. Pitch anchor points specify the segmental locations of each significant pitch event within an utterance. Pitch trajectories describe the path taken by an intonation contour between two pitch anchor points.

Each pitch anchor point gives not only the segmental location of a pitch event but also the pitch height which characterises it (e.g. low, medium or high in the speaker's pitch range). Also specified is the amount of latitude in both normalized pitch value and time to be associated with that pitch anchor point. This allows for a certain amount of variability in acceptable contours, and compensates for any small inaccuracies in the output of the SPELL segmenter, which determines the segment and syllable boundaries (see Section 3.3 below).

French. According to Vaissière [13], French intonation in simple constructions is based on unitary pitch contours (or "tunes"). Tune 1 is used for declarative statements, *qu*–questions and inverted polar questions, and consists of a rise–fall pattern going from mid to high to low. The location of the turning point is determined by rule: it falls on the final syllable of the first lexical word or within the first five syllables of the utterance if this is sooner. Tune 2 is used for non–inverted polar questions, and consists of a shallow rise from mid–low to mid, with a rapid rise from mid to high on the very last syllable of the utterance. These contours are illustrated in Figure 1, with their associated pitch anchor points. In each case only three anchor points are needed: one on the initial segment, one on the final segment, and one marking the turning point from high to low (Tune 1) or from mid to high (Tune 2).

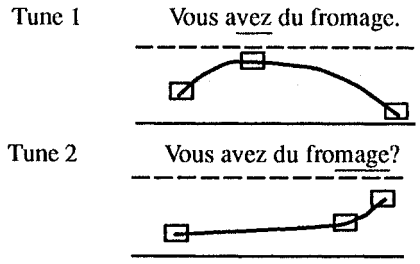

Figure 1. Schematic representations of the two primary tunes for French intonation, with the internal pitch anchor point underlined.

Italian. A tune analysis is also used by Chapallaz [2] for Italian intonation. Tune 1 is the usual intonation for statements, *qu*–questions, commands and exclamations, while Tune 2 is the typical intonation for short, introductory non–final statements and polar questions. The two tunes are generally similar in form: the contour begins mid and rises to high at the first stressed syllable; it then descends gently to mid and falls rapidly from mid to low during the final stressed syllable. Thereafter the two tunes diverge: Tune 1 stays on a low level pitch, while Tune 2 rises again to mid or mid–high. This pitch behaviour on the last syllable is all that serves to distinguish the two contour types. These contours are illustrated in Figure 2 along with their associated pitch anchor points: one on the initial and final segments, one on the first stressed syllable and two on the last stressed syllable.

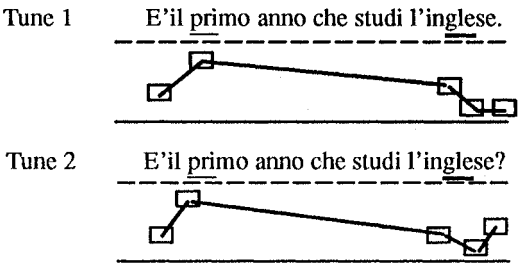

Figure 2. Schematic representations of the two primary tunes for Italian intonation, with the internal pitch anchor points underlined (multiple anchor points within a single syllable are marked by double underlining).

English. For English, the outcome of a more abstract analysis, that of Halliday [5], has been adopted. In Halliday's analysis, the intonation contour is the result of a series of choices at key locations in the "tone group": the *tonic*, which has a choice of five nuclear "tones" (that is, contours), the *pre–tonic* and the *post–tonic*. The choice of tone on the tonic syllable is what dis-

tinguishes the intonational meaning of the phrase (statement versus question, for example), while its placement in the tone group indicates the semantic focus: by default, the tonic is placed on the accented syllable of the last lexical word.

In order to simplify the teaching task, two resulting contours or "tunes" have been selected from the range of possibilities allowed by Halliday, to serve as models for English intonation. Tune 1 is to be used for declarative statements, *wh*–questions and imperatives. It consists of Halliday's primary Tone 1 – a fall from high to low on the tonic syllable – preceded by a level mid pre–tonic and followed by a level low post–tonic. English Tune 2 is to be used for polar questions, and consists of Halliday's Tone 2 – a rise from low to high over the tonic syllable – preceded by a level low pre–tonic and followed by a high rising post–tonic. These contours are illustrated in Figure 3, with their associated pitch anchor points: one on the initial segment, one on the final segment, and the remainder on the last syllable of the pre–tonic, the beginning and end of the tonic, and the initial segment of the post–tonic.

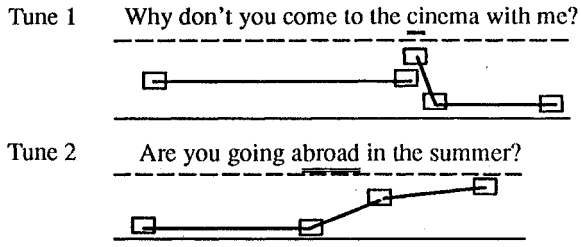

Figure 3. Schematic representations of the two primary tunes for English intonation. Double underlining is used for the tonic syllable, indicating that it has two anchor points.

3.3 Intonation signal processing

The processing of a student's utterance for intonation has two strands: the derivation of a heavily smoothed, normalized fundamental frequency contour and the extraction of the segmental sequence.

The fundamental frequency (F0) contour is extracted from the low–pass–filtered speech waveform by an adapted super–resolution pitch determination algorithm [10]. The contour is heavily smoothed, normalized by the mean and standard deviation of the speaker's F0 (established during an enrolment phase), interpolated to fill in gaps, and smoothed again.

The segmentation is obtained using a Hidden Markov Model technique similar to systems used for speech recognition; labelling of the incoming speech is constrained by a *phrase model* which gives the possible segmental content of the student's utterance in terms of a set of sub–phonemic Acoustic Phonetic Units (APUs). This model allows for a variety of alternative pronunciations, including predictable cross–language errors, to ensure a reasonably accurate segmentation.

The pitch contour and the segmentation are then evaluated using a similarity metric based on the the pitch anchor point and pitch trajectory features of the target utterance.

3.4 The intonation similarity metric

The intonation similarity metric judges the student's pitch contour, as derived by the intonation signal processing modules, against a target pitch contour for the utterance in question. There is, however, no direct comparison between the student's contour and that of the "teacher". Instead, the target is derived indirectly, using measurements from the teacher's pitch contour, knowledge of the student's pitch range and the phonetic segmentation of the student's own realization to construct a set of pitch anchor points and trajectories for the student's utterance (see Figure 4). The variability allowed at each anchor point produces a pitch "tunnel" through which the student's pitch contour must pass if it is to be judged acceptable. Feedback can be given on each component of the contour between any two anchor points, making this approach suitable for both whole–tune and componential treatments of intonation.

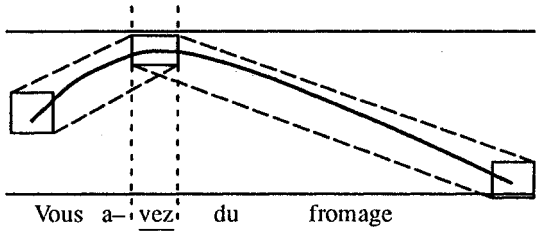

Figure 4. Application of the intonation similarity metric for French Tune 1 to a student's pitch contour (schematic). Dashed lines represent the limits of the "pitch tunnel" built around the three anchor points.

4 THE ANALYSIS OF RHYTHM

4.1 The nature of rhythm

In most European languages, including English, French and Italian, stress or salience is marked acoustically by modulation of one or more of the parameters of fundamental frequency, intensity, duration and segmental features. Control of the interaction between rhythm and stress is of great significance for foreign language learners, particularly in English and Italian, since the occurrence of stressed syllables is one of the factors which gives them their characteristic rhythm. In addition, certain stressed syllables constitute the anchor points for the pitch movements on which intonation depends. In both languages the differences between stressed and unstressed syllables are quite marked. French, in contrast, lacks the apparently regular recurrence of stress beats which characterizes the other two rhythmic systems, and the distinction between stressed and unstressed syllables is not as marked [12].

A useful approach to rhythm is proposed by Dauer [3], who suggests that languages should be considered as being more or less "stress–based" according to their tendencies on parameters such as syllable structure, the nature of stress and the use of vowel reduction (see Figure 5). Thus, English is at one extreme

Figure 5. Degrees of dependence on rhythmic stress following Dauer [3].

of the "stress–based" scale: it marks the distinction between stressed and unstressed syllables quite strongly, typically with changes in the duration of the stressed vowel and the location of a pitch movement in the intonation contour, while the quality and duration of unstressed vowels are reduced. Italian, while also stress–based in that it marks stress strongly with duration and pitch, does not centralize its unstressed vowels, and has a perceptibly different rhythm from that of English. French, which is placed at the opposite end of the stress–based scale by Dauer, minimizes any durational or qualitative difference between stressed and unstressed syllables, and the absence of

vowel reduction produces a rhythm entirely different from that of Italian and English.

Significant improvements in the rhythmic quality achieved by learners of these three languages may be possible simply by concentrating on a small set of acoustic parameters. Learners of English should be encouraged to produce weak vowels with reduced duration and centralized quality [4]. Learners of Italian should aim to contrast duration but keep vowel qualities uncentralized [2]. Finally, learners of French must avoid any reduction in duration or vowel quality [12]. The remaining acoustic correlates of stress (i.e. fundamental frequency and intensity) are not considered since these features are used in a similar manner for marking stress in all three languages.

4.2 The analysis of rhythm

The parameters of duration and vowel quality are derived indirectly in the SPELL system, using the Hidden Markov Model segmenter described in Section 3.3 above. APU models are created for a variety of realizations of a given vowel, and are included as options in the phrase model which governs the operation of the segmenter for an utterance. The rhythmic status of the syllable – *strong* or *weak* – is then inferred from the choice made by the segmenter in processing the student's utterance of a given phrase.

An example for English is given in Figure 6. The lessons in English rhythm concentrate initially on the acquisition by the student of a small set of weak (phonetically–reduced) forms of function words (*for, the, to, some* etc.), as a way of achieving the contrast between strong and weak syllables [11]. The phrase model for a phrase such as *back to back* (Figure 6) would allow realizations of the word *to* ranging from the full citation form with the vowel /u:/ to the correct weak form with schwa /ə/. The choice of /u:/ would indicate that the student had made that syllable too *strong*, and the system would advise the student accordingly. The choice of schwa, on the other hand, would indicate that the student had made the syllable *weak*, as required by the rhythmic structure of the phrase, and the student's attempt would be judged to be correct.

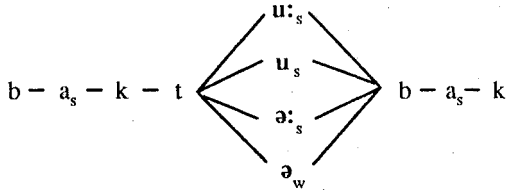

Figure 6. Segmenter phrase model for the phrase "back to back" *showing the use of alternative APUs to determine the rhythmic status of a syllable. Subscript $_s$ indicates a strong syllable nucleus, subscript $_w$ a weak one.*

5 SUMMARY

The SPELL workstation covers both prosodic and segmental aspects of foreign language pronunciation. This paper has described the analysis and teaching of the prosodic features of intonation and rhythm.

In a departure from traditional contour matching techniques, the teaching of intonation concentrates on more abstract representations which can be generalized by the student and are teachable by rule. Achieving the correct segmental alignment of prosodic features is central to this approach.

Of particular interest is the SPELL automatic segmenter, which provides the segmental sequence against which the prosodic features are aligned, and which allows the system to cope with predictable pronunciation errors.

The SPELL segmenter also provides judgements on the parameters of duration and vowel quality, which form the basis for the rhythm analysis in the three languages. These judgements are used to provide feedback in terms of the *weak–strong* syllable contrast.

Both intonation and rhythm analyses are integrated into a courseware package using the DELTA teaching paradigm. Instruction and feedback are provided by user–friendly window–based graphical and audio presentations.

A recently announced two–year extension to the project will allow the intonation and rhythm teaching courseware to be extended and the response time of the similarity metrics to be improved.

6 REFERENCES

[1] M.–G. di Benedetto, F. Carraro, S. Hiller, E. Rooney. "Vowel pronunciation assessment in the SPELL system." To appear in *Proc. ICSLP–92*, 1992.

[2] M. Chapallaz. *The Pronunciation of Italian: a Practical Introduction.* London: Bell and Hyman, 1979.

[3] R. Dauer. "Stress–timing and syllable–timing reanalyzed." *Journal of Phonetics*, 11, 51–62, 1983.

[4] D. Faber. "Teaching the rhythms of English: a new theoretical base." *International Review of Applied Linguistics* 24, 206–216, 1986.

[5] M.A.K. Halliday. "Tones of English." In W.E. Jones and J. Laver (eds.), *Phonetics in Linguistics: A Book of Readings*, 103–126, London: Longman, 1973.

[6] S. Hiller, E. Rooney, J. Laver, M.–G. di Benedetto, J.–P. Lefèvre. "Macro and micro features for automated pronunciation improvement in the SPELL system." *Proc. ESPRIT '91*, 378–392, 1991.

[7] J. Harmer. *The Practice of English Language Teaching.* London: Longman, 1983.

[8] J. Kenworthy. *Teaching English Pronunciation.* London: Longman, 1987.

[9] P. Leach. "French intonation: tone or tune?" *Journal of the International Phonetics Association*, 18, 125–139, 1988.

[10] Y. Medan, E. Yair and D. Chazan. "Super resolution pitch determination of speech signals." *IEEE Trans. Sig. Proc., ASSP–39* (1), 40–48, 1991.

[11] D.S. Taylor. "Non–native speakers and the rhythm of English." *International Review of Applied Linguistics* 19, 219–226, 1981.

[12] B. Tranel. *The Sounds of French: an Introduction.* Cambridge: Cambridge University Press, 1987.

[13] J. Vaissière (personal communication).