MACRO AND MICRO FEATURES FOR AUTOMATED PRONUNCIATION IMPROVEMENT IN THE SPELL SYSTEM[1]

Edmund Rooney*, Steven Hiller*, John Laver*, Maria–Gabriella Di Benedetto**

*CSTR, University of Edinburgh; **INFOCOM Dept., University of Rome

ABSTRACT

The analysis of macro (prosodic) and micro (segmental) features is described for a workstation designed to improve the pronunciation of English, French and Italian by non–native speakers. The SPELL workstation is intended to be a teaching device aimed at intermediate ability foreign language learners. Audio and visual aids will be used to help students improve their general intelligibility within a basic teaching paradigm called DELTA (Demonstrate, Evaluate Listening, Teach and Assess). Prosodic analysis will apply to the features of intonation, stress and rhythm. A phonological approach is used for intonation which provides a well–structured system of contrasting units that correlate with discrete linguistic functions. A more limited approach to the prosodic phonology of stress and rhythm will be taught in the SPELL system by manipulating vowel quality and segmental duration. The micro feature analysis will focus on vowel contrasts, using a distinctive feature approach to characterize non–native vowel pronunciation. Acoustic properties are sought which will be speaker–independent.

Keywords: Computer–aided Language Learning; Pronunciation; Prosody; Articulatory Phonetics; ESPRIT

1 INTRODUCTION

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 language pronunciation. This is a feasibility study involving 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 *macro features* section of the SPELL project covers the prosodic parameters of intonation, stress 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 the prosodic phonology of stress and rhythm will be taught in the SPELL system by manipulating the acoustic features of vowel quality and segmental duration. The *micro features* section focuses on the segmental class of vowels using a distinctive feature approach to characterize non–native vowel pronunciation.

The main technical innovation behind SPELL is the departure from the traditional practice of whole utterance matching used when teaching pronunciation. Instead, well–founded phonetic and phonological principles will be applied to teaching selected aspects of English, French and Italian pronunciation.

2 A GENERAL DESCRIPTION OF THE SPELL SYSTEM

2.1 Some Basic Assumptions

The SPELL workstation will be an autonomous teaching system for use by intermediate ability foreign language speakers without sophisticated linguistic or phonetic knowledge. Visual displays will help students to master relevant concepts without requiring expert knowledge, while audio aids will enable them to listen to the pronunciation of items of interest and will synthesize intermediate or exaggerated targets to attract the student's performance into the required zone of acceptability. A minimal set of fully–defined courseware will be developed in key areas for the demonstrator system. Intelligibility will be used as the criterion for improvements in pronunciation, since for the majority of students improvement in intelligibility is a more practical objective than the acquisition of fully native pronunciation (see, for example, Harmer, 1983; Madsen, 1983).

2.2 The Development of SPELL Courseware

The use of teaching courseware provides directed instruction to the student, allowing better predictions to be made about the student's performance, and enabling a proper evaluation of the system itself as a teaching aid. The typical aspects common to many foreign language teaching practices can be summarized in a paradigm called **DELTA**, which will be used to structure the courseware designed for the SPELL workstation:

Demonstrate — Audio demonstrations of various utterances are used to highlight the pronunciation features of interest.

Evaluate Listening — Small listening tests are completed by the student to evaluate his or her ability to perceive the pronunciation features of interest.

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

Teach — The pronunciation features of interest are taught, with quantitative feedback for the student and directions for modifying inadequate performances.

Assess — A formal evaluation of the student's ability to pronounce the features of interest is made.

In addition to the DELTA paradigm, a fuller assessment of proficiency can be given to the students after several lessons (e.g. using standard language proficiency tests such as the cloze test), to evaluate their general performance in using the workstation.

2.3 Design Considerations in the Development of the SPELL workstation

1. Integration of macro and micro features in language use.

A useful pronunciation–teaching system cannot rely solely on the teaching of phonemes in isolation, nor on direct imitation of target utterances. For example, 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. It is therefore more desirable to concentrate on getting the student to imitate more abstract aspects of the contour such as its overall shape and the location of any pitch features. Accurate feature analysis will depend on the location of phonetic segments within an utterance produced by a student. Therefore, an automatic SPELL segmentation program is being developed for application to the speech signal prior to feature extraction and analysis.

2. Ability to handle pronunciation errors

The SPELL system must be able to deal with a variety of pronunciation errors produced by non–native speakers (e.g. systemic, structural and realization errors). Careful construction of teaching materials can limit the types of error which might occur, but some will remain. The SPELL segmenter has therefore been designed with a segmental transition network which includes the more predictable types of error occurring between two given languages.

3. Importance of appropriate feedback

The appropriate feedback will have to be provided by the SPELL workstation for the user. Quantitative feedback may be appropriate for vowel quality contrasts, whereas diagnostic feedback with instructions for improvement is appropriate for prosodic features. It should be emphasized that the feedback to a SPELL user will not be expressed in terms of explicit linguistic or phonetic concepts.

3 THE ANALYSIS OF MACRO FEATURES IN THE SPELL SYSTEM

3.1 Definition of Macro Features

Macro or *prosodic* features are those which operate over stretches of speech longer than the single segment or phoneme, and here include intonation, stress and rhythm. **Intonation** is generally defined as the manipulation of pitch for linguistic and paralinguistic purposes at a level above that of the segment. **Stress** is the term used to refer to a number of ways in which certain syllables are made more prominent than surrounding syllables. The **rhythm** of an utterance is given by the patterning in time of the segments, syllables and stresses.

3.2 Phonological Approaches to Intonation

It is not possible to teach intonation simply by direct imitation of target utterances, since actual pitch contours can vary enormously; what the pupil requires is a pattern or model which can be generalized to other utterances of the same type or for the same purpose, and the ability to choose from a set of such models to convey contrasts of meaning or emphasis.

Comparing intonational systems amongst the three target languages in terms of their phonology is quite difficult given the varying depth of treatment and the differing approaches to the problem in the published literature. Some general principles are clearly common to all three languages. Firstly, pragmatic linguistic functions such as statements and questions are differentiated by opposing pitch movements (e.g. falling versus rising pitch). Secondly, pitch movements are related to rhythmical structure by the marking of accented syllables. Finally, intonational pitch movements are *anchored* to the segmental structure of the utterance.

The major difference in terms of phonology between English, French and Italian is the extent to which the intonation contour is treated as a structural chain with elements of choice at certain locations. In French, the choices within the contour are very limited with the whole contour being treated as a single "tune" (see, for example, Leach 1988). Italian is slightly more complex in that the contour can be subdivided into a chain but with a limited choice of elements. In English, the contour can be subdivided into a very complex chain with many choices at various locations (see, for example, Halliday 1973). A practical approach to describing and teaching intonation has been adopted to overcome these differences in phonology between the three languages, as discussed below.

For each language, the discussion will be limited to the two primary intonation functions which will provide significant coverage for learners: statements/wh–questions (qu–questions in French and Italian) and polar ("yes/no") questions.

3.3 The Analysis of Intonation

In this analysis, particular attention is drawn to two notable features which characterize intonation contours. The first feature is called a *pitch anchor point*, which specifies a segmental location within an utterance (usually a syllable) that has a significant pitch event attached to it. The second feature is a *pitch trajectory*, which describes the path taken by an intonation contour between two pitch anchor points. The use of such contour features simplifies the task of teaching intonation and allows the phonological features of all three languages to be described using a common terminology.

According to Vaissière (personal communication), French intonation is based on unitary pitch contours (or "tunes"). Tune 1 is used for declarative statements, qu–questions and inverted polar questions while Tune 2 is for non–inverted polar questions. Figure 1 displays schematic representations for the two French tunes to be taught as part of SPELL prosodic features. Both tunes have a single internal anchor point located by rule.

A tune analysis is also appropriate for Italian intonation (e.g. Chapallaz (1979). Tune 1 is the usual intonation for statements,

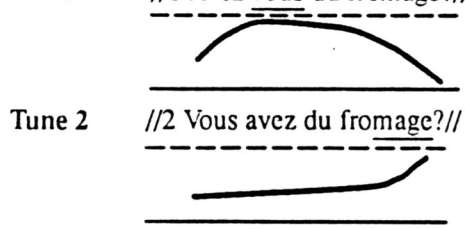

Figure 1. Schematic representations of the two primary tunes for French intonation, with the internal pitch anchor points underlined.

qu–questions, commands and exclamations. Tune 2 is the typical intonation for short, introductory non–final statements and polar questions. These two tunes are generally similar in structure and differentiated by the final movement of the contours. Figure 2 displays schematic representations for the two Italian tunes chosen for the SPELL demonstrator. Both tunes are broadly shaped by three pitch anchor points, one located on the first and two on the last stressed syllable.

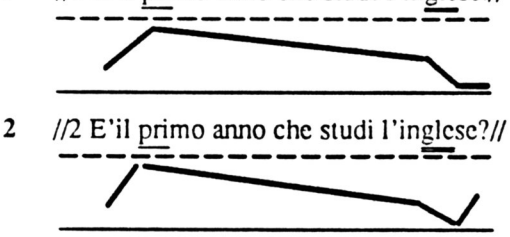

Figure 2. Schematic representations of the two primary tunes for Italian intonation, with the internal pitch anchor points underlined (multiple anchor points within a single syllable are marked by double underlining).

English exhibits the greatest complexity in its structuring of intonation, and a more abstract analysis, in terms of the choice and placement of a nuclear *tone* with associated *pre–tonic* and *post–tonic* contours has been adopted (Halliday 1973).

Halliday's primary Tones 1 and 2 have been selected since they provide a substantial coverage of intonational uses within English. Tone 1 (high fall to low) is used for declarative statements, wh–questions and imperatives, Tone 2 (rise from low to high) for polar questions and certain other attitudinal information.

In order to simplify the teaching task, one set pre–tonic contour for each tone will be taught to the student learning English; the post–tonic choice of pitch pattern is in any case completely prescribed. Figure 3 displays schematic representations for the two chosen English tones, with their associated pre–tonic and post–tonic contours.

3.4 Stress and 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 rhythmic 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

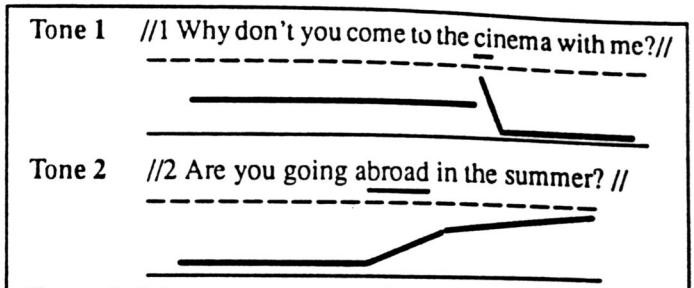

Figure 3. Schematic representations of the two primary tones and associated pre– and post–tonics chosen for English intonation. Double underlining is used for the tonic syllable, indicating that it has two anchor points.

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 (Tranel 1987).

A useful approach to rhythm is proposed by Dauer (1983). Dauer 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 4). Thus, English is at one ex-

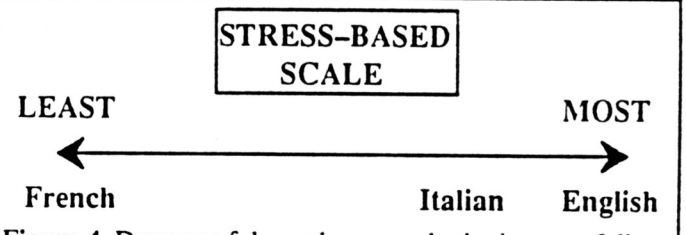

Figure 4. Degrees of dependence on rhythmic stress following Dauer (1983).

treme 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 towards the bottom 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 vowels with reduced duration and centralized quality. Learners of Italian should aim to contrast duration but keep vowel qualities uncentralized. Finally, learners of French must avoid any reduction in duration or vowel quality. The remaining acoustic correlates of stress (i.e. fundamental frequency and intensity) are not considered since these features are difficult to relate to stress and rhythm, and they are used for stress marking for all three languages.

4 THE ANALYSIS OF MICRO FEATURES IN THE SPELL SYSTEM

The *micro* or *segmental* feature analysis is focussed on the teaching of vowel contrasts for the initial demonstrator system. The first main area of research addresses the characterization of non–native vowel production in terms of the distinctive features of the target language. The second seeks an acoustic representation of vowels which is independent of the speaker, to allow between–speaker and cross–language comparisons.

4.1 Characterization of non–native vowel production.

For the initial demonstrator system, only the most common errors in non–native vowel pronunciation are being considered. For non–native speakers of English, these include the tense/lax high vowel contrasts /i ~ ɪ/ and /u ~ ʊ/, and the front–back distinction for the English low vowels /a ~ ɑ/. Major problems for non–native speakers of French are the production of nasal and front rounded vowels , and the need to avoid the tendency to diphthongize pure vowels. French speakers learning Italian have no major problems, since all Italian vowels have correspondences in the French vowel system. In the case of native English speakers, the main problem is avoiding the diphthongization of the pure vowels.

4.2 Problems of vowel representation

Two problems arise when representing vowels by means of acoustic parameters: *vowel coarticulation* within the production of one speaker and *vowel normalization* between speakers.

The *coarticulation* effect – the influence of phonetic context on the articulation of a vowel – gives rise to a range of different formant frequency values for a given vowel within the production of one speaker, and may cause the acoustic parameters of two different vowel phonemes to overlap. A given formant pattern cannot then be identified uniquely. Phonetic context must therefore be held constant when vowels are being compared.

Comparisons between the vowels of two different speakers give rise to the problem of *vowel normalization*, since the same vowel phoneme may have a different acoustic realization for two speakers owing to the differences in their vocal tract shape and dimensions. This normalization can be achieved by considering the normalized bark–scaled difference values using the first three formants and the fundamental frequency (e.g. F1–F0, F2–F1, F3–F2 as suggested by Syrdal and Gopal 1986), or by obtaining a representation of the speaker's peripheral vowels /i ,a ,u/ during a limited training phase (see e.g. Minifie 1973).

The multi–lingual basis of the SPELL project brings a third problem: the need to compare vowel articulations across languages. Two questions are under investigation: (a) whether a speaker uses similar formant frequency values for a foreign vowel which corresponds to a vowel of his or her native vowel system; and (b) whether it is possible to predict the location within a speaker's vowel space of a vowel which does not exist in his or her native language.

5 SUMMARY

The SPELL workstation covers both prosodic (*macro*) and segmental (*micro*) aspects of foreign language pronunciation. In a departure from traditional contour matching techniques, it concentrates on more abstract representations which are generalizable by the student. Particular attention is being paid to the segmental alignment of prosodic features and the ability to cope with predictable pronunciation errors.

SPELL research is now being completed in a number of areas. A multi–lingual speech database has been collected and transcribed to SAM standards. Acoustic parameter extraction programs are being constructed for the features of fundamental frequency, formant frequency and segmental duration. These features are to be analyzed using a set of SPELL metrics which are under development. Of particular interest is the SPELL phonemic segmenter which will provide the segmental sequence against which the prosodic features are aligned. Preliminary work has also begun on a user friendly interface for the SPELL workstation using state–of–the–art window–based graphical presentations.

The SPELL project is already considering possible extensions of this technology into other fields. The most direct extension is into the speech pathology area, particularly for speakers with articulation disorders. One example would be the assessment and rehabilitatory treatment of the speech of patients suffering from dysarthria. Another would be the use of a SPELL workstation in restoring a degree of intelligibility to the speech of patients who have undergone oral surgery for tumors of the lingual or pharyngeal structures.

6 REFERENCES

Chapallaz, M. (1979); *The Pronunciation of Italian: a Practical Introduction*, London: Bell and Hyman.

Dauer, R. (1983); Stress–timing and syllable–timing reanalyzed. *Journal of Phonetics*, 11, 51–62.

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

Harmer, J. (1983); *The Practice of English Language Teaching*, London: Longman.

Leach, P. (1988); French intonation: tone or tune? *Journal of the International Phonetics Association*, 18, 125–139.

Madsen, H.S. (1983); *Techniques in Testing*, Oxford: Oxford University Press.

Minifie, F.D. (1973); Speech acoustics. In Minifie, F.D., Hixon, T.J. and Williams, F. (eds.), *Normal Aspects of Speech, Hearing and Language*, 235–284, New Jersey: Prentice–Hall, Inc.

Syrdal, A.K. and Gopal, H.S. (1986); A perceptual model of vowel recognition based on the auditory representation of American English vowels. *J. Acoust. Soc. Am.*, 79, 1086–1100.

Tranel, B. (1987); *The Sounds of French: an Introduction*, Cambridge: C.U.P.