AN AUTOMATED SYSTEM FOR COMPUTER AIDED PRONUNCIATION TEACHING

Steven Hiller, Edmund Rooney, John Laver and Mervyn Jack

Centre for Speech Technology Research
University of Edinburgh

ABSTRACT – This paper describes the SPELL workstation which has been designed to improve the pronunciation of foreign languages (English, French and Italian) by non–native speakers. The workstation is used presently for the automated assessment and improvement of the prosodic features of intonation and rhythm, and the segmental feature of vowel quality. The paper highlights the intonation, rhythm and vowel quality metrics used for assessing non–native speech. The results of a preliminary evaluation by language experts and teachers support the underlying phonetic analysis techniques as well as the pedagogic approach presented to the workstation user.

INTRODUCTION

The European Community's ESPRIT project SPELL (Interactive System for Spoken European Language Training) has just completed its initial two year phase (1). This feasibility study has produced demonstrator systems for teaching prosodic and segmental pronunciation features to non–native students of French, English and Italian. One of the technical innovations behind SPELL is the departure from the practice used in some existing automated teaching systems of requiring exact acoustic copying. Instead, techniques of normalization and segmentation are used to achieve comparisons at a level of perceptual and ultimately phonetic equivalence. The first three sections of this paper cover the prosodic features of intonation and rhythm as well as the segmental feature of vowel quality. For each pronunciation feature, details are given for the similarity metric used, the signal processing required by the metric and the interface presented to the user. The final section discusses the results of a preliminary evaluation of the SPELL modules by language experts and teachers.

PROSODIC ANALYSIS

Prosodic features operate over stretches of speech longer than the single segment, and here include intonation and rhythm. Intonation can be defined as the manipulation of pitch for linguistic, paralinguistic and pragmatic purposes at a level above that of the word. The rhythm of an utterance is given by the patterning in time of the 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.

Intonation teaching module

In the SPELL system, a practical phonetic approach to the description and analysis of intonation has been adopted (see Hiller et. al, 1991). 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. Central to this analysis is the relationship between the fundamental frequency (F0) contour and its associated segmental sequence.

For each intonational function in each language, a single contour has been chosen as the model for teaching. Each contour is characterized by a set of *pitch anchor points* and a corresponding set of *pitch tunnels*. Pitch anchor points specify the segmental locations of each significant pitch event within an utterance. Pitch tunnels describe the tolerances for a path taken by an intonation contour between two pitch anchor points. Figure 1 displays an example of the intonation similarity metric required for a typical intonation contour used in simple French declarative statements. This contour (the solid line) consists of a rise–fall pattern, and the pitch anchor points (the rectangles) indicate that the pitch levels go from mid to high to low. The location of the turning point (the middle rectangle) is determined by rule: it occurs on the final syllable of the first lexical word or within the first five syllables of the utterance if this is sooner. The width and height of each anchor point indicate the variability allowed in terms of pitch height and segmental location. This variability at each anchor point produces a pitch tunnel (the dashed lines) through which a F0 contour must pass if it is to be judged acceptable by the system. Feedback can be given on each component of the contour between any two anchor points, making this approach suitable for both whole–contour and componential treatments of intonation.

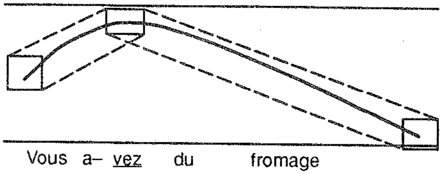
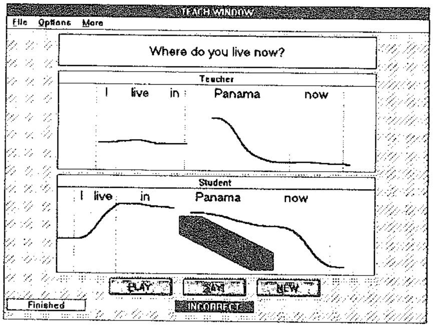


Figure 1. The intonation similarity metric as applied to the pitch contour associated with French statement intonation. Dashed lines represent the tolerances of an acceptable pitch tunnel built around the three anchor points.

The processing of an input utterance for intonation has two strands: the derivation of a heavily smoothed, normalized fundamental frequency contour and the extraction of the segmental sequence. The F0 contour is extracted from the low–pass–filtered speech waveform by an adapted version of a super–resolution pitch determination algorithm (Medan et. al, 1991). The contour is heavily smoothed by a non–linear smoother (Rabiner et. al, 1975), normalized by the mean and standard deviation of the speaker's F0, interpolated to fill in gaps and smoothed again. The segmentation is obtained using a Hidden Markov Model (HMM) technique (see McInnes et. al 1992); 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 errors predictable from the student's mother tongue, to ensure an accurate segmentation.

Figure 2 shows a typical example of the user interface for the SPELL intonation teaching module, in this case for a student studying English statement intonation. The general interface conforms to the common



Where do you live now?

Figure 2. An example of the user interface for the SPELL intonation teaching module. In this display, the target language is English and the tonic (nuclear) portion of the statement intonation is being practiced.

user access conventions of the Microsoft® Windows® 3.0 graphical environment. The three main windows are (from top to bottom) the *question window*, *teacher window* and *student window*. The *teacher window* displays the target intonation to be achieved by the student, including the teacher's smoothed F0 contour, the time–aligned orthographic representation of the utterance and the vertical word boundary lines. In this example, the student is practicing the tonic fall in the word "Panama" which has been embedded between a pre–tonic and a post–tonic (these two parts of the utterance are not being attended to by the student). The *question window* is used as a means of providing a communicative context for the student to ensure that the correct emphasis is used in the response. The *student window* displays the results of analyzing the student's utterance using all the features found in the teacher window. In this example, the student has used the wrong intonation at the tonic, and the error is highlighted by the appearance of the correct pitch tunnel at this point. A small text window at the bottom of the interface displays a CORRECT or INCORRECT message, and a system voice announces "Well done" or "Try again" accordingly. Three main buttons are provided for the student to control the intonation teaching module. The PLAY button allows the student to listen to the teacher's model. The use of the SAY button records and analyzes an attempt at the target utterance. The NEW button permits the student to choose another utterance for practice.

Rhythm teaching module

A major component of rhythm in English, French and Italian is the relationship between strong and weak syllables. In English and Italian, the frequent occurrence of strong syllables, contrasting markedly with intervening weak syllables, produces a characteristic rhythmic 'beat'. French lacks this apparently regular beat, and the distinction between strong and weak syllables is not as marked. In all three languages, syllable strength is marked acoustically by modulation of one or more of the parameters of fundamental frequency, intensity, duration and vowel quality (i.e. formants). Significant improvements in the rhythmic quality achieved by learners of French, English and Italian may be possible by concentrating on a small set of acoustic parameters. Learners of English should be encouraged to produce weak syllables with reduced duration and centralized vowel quality. On the other hand, learners of French must avoid any reduction in duration or centralization of vowel quality. An intermediate situation exists for students of Italian; they should aim to contrast duration but keep vowel qualities uncentralized. The remaining acoustic correlates of syllable strength (i.e. F0 and intensity) have not been considered since these features are used in a broadly similar manner in all three languages.

The parameters of duration and vowel quality (formants) are derived indirectly in the SPELL system, using the HMM segmenter described 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 a prescribed 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 3. The lessons in English rhythm concentrate initially on the acquisition by the student of a small set of weak forms of function words (*for, the, to, some* etc.), as a way of achieving the contrast between strong and weak syllables. The phrase model for a phrase such as *back to back* (Figure 3) allows several realizations of the word *to*. The choice of /u:/ (full citation–form) would indicate that the student had made that syllable too *strong* since the duration was too long and the vowel quality was not centralized. 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. The analysis also recognizes two levels of intermediate strength. One allows for use of the citation–form vowel quality but with an appropriately short duration, while the other allows for the correct central vowel quality with a prolonged duration.

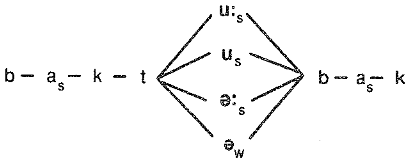
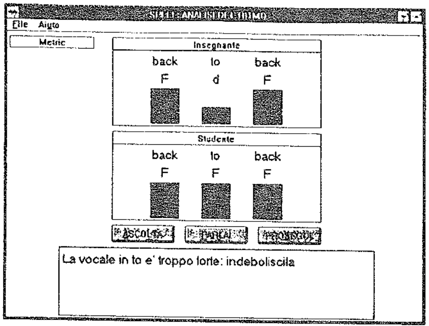


Figure 3. Segmenter phrase model for the phrase *back to back* showing the use of alternative APUs to determine the rhythmic status of a syllable. Subscript $_s$ indicates a strong syllable nucleus, subscript $_w$ a weak one.

Figure 4 shows a typical example of the user interface for the rhythm teaching module, in this case for an Italian student studying English statement intonation. There are three main windows in this interface: the *teacher window* (Insegnante), *student window* (Studente) and *diagnosis window*. The *teacher window* displays the target rhythm pattern to be achieved by the student (in this example, the student is mainly concerned with achieving a weak syllable for the function word *to* in the phrase *back to back*). The target display includes the orthographic representation of the short phrase, labels indicating the target rhythmic status (F = Forte = strong; d = debole = weak) and graphic blocks which also indicate the target rhythmic status (tall = strong; short = weak). The *student window* displays the results of analyzing the student's utterance using all the features found in the *teacher window*. Note that an incorrect syllable strength has



La vocale in to e' troppo lorte: indeboliscila

Figure 4. An example of the user interface for the rhythm teaching module for an Italian student learning English.

been used for the function word *to*. In this case, the label and graphic block indicate that a strong syllable was produced by the student. The *diagnosis window* at the bottom displays a message in Italian indicating that the vowel in the word *to* was too strong and that the student should weaken it. Three buttons are provided for the student to control the rhythm teaching module. The PLAY button (ASCOLTA) allows the student to listen to the teacher's model. The use of the SAY button (PARLA) records and analyzes an attempt at the target utterance. The NEW button (PROSEGUI) permits the student to choose another utterance for practice.

VOWEL ANALYSIS

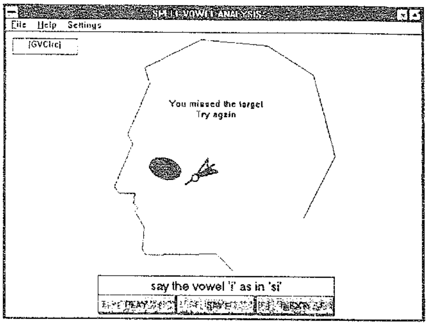
The SPELL project is also concerned with the teaching of segmental features. It was decided to focus research and development efforts on vowels (rather than consonants) for the initial phase of the project. Within this class, a SPELL teaching module has been produced for teaching monophthong vowels in the three target languages. Further work is required to develop analysis techniques for French nasal vowels, French front rounded vowels and the diphthongs of English and Italian. The issue of co–articulation has been effectively ignored in this work by analyzing vowels produced within a standard context (e.g. the English vowels are produced within the traditional *h–V–d* frame to produce *heed, head, had*, etc.).

Vowel teaching module

The main purpose of the vowel similarity metric is to check if a given student's vowel token falls within a vowel space derived from a target set of vowels produced by a group of native speakers. In order to derive the target vowel spaces, vowel data were collected from 11 speakers per sex per language, with each speaker producing 3 repetitions per vowel target. The formant data used to create the target vowel spaces were derived as described in the next paragraph. The data were checked for obvious formant tracking errors and the corresponding samples were deleted from the database. Extreme outliers were also identified and excluded from the data. The final target vowel spaces were then derived using an elliptical representation in a 2–dimensional space; targets were calculated using spreads of 1, 1.5 and 2 standard deviations either side of the mean values. The final decision of the metric determines if the input formants from the student's vowel token fall within the equivalent target vowel space.

The acoustic vowel analysis provides a representation of a vowel token in terms of the normalized formant parameters which are input to the vowel similarity metric. A given speech token is submitted to the HMM segmenter (using speaker–independent APUs) to isolate the vowel from any surrounding speech or silence. The isolated vowel is then analyzed to produce estimates of the first three formant frequencies (using a modified McCandless algorithm, 1974) and F0 for its entire duration. The most stable region of the vowel is located by a "steady–state" finder algorithm (Van Bergem, 1991). The four acoustic features are then averaged across the stable vowel region. The resultant averaged parameters are used to normalize the formants by transforming to a Bark scale and then calculating the formant differences F1–F0, F2–F1 and F3–F2 (Syrdal and Gopal, 1986).

Figure 5 shows a typical example of the user interface for the vowel teaching module for a male English student studying isolated French vowels. The main vowel display is placed above a small prompt window



You missed the target
Try again

say the vowel 'i' as in 'si'

Figure 5. An example of the user interface for the vowel teaching module for an English student learning French.

and three user buttons. The main display shows: 1) an outline of a side view of a male head to provide a reference for the articulatory position of vowel targets, 2) an elliptical vowel target for the French vowel /i/ (located as high–front relative to the figure of the head), 3) a dart whose tip indicates the position of the normalized vowel formants derived from the student's attempt (slightly low and too far back) and 4) a feedback message which, in this case, indicates that the student has missed the target. The relative position of the target and the dart may be used by the students in a "biofeedback loop" to correct their pronunciation.

PRELIMINARY EVALUATION BY LANGUAGE TEACHERS AND EXPERTS

The SPELL project has successfully constructed working systems for the analysis of non–native speech. A preliminary evaluation was completed to determine if these systems would be useful in the actual remediation of non–native pronunciations. It was decided, in the first instance, to seek the opinions of language experts and teachers since the SPELL workstation is not yet complete enough for full student field trials. System evaluations were completed for each SPELL language at an appropriate native language site. These language professionals were given full demonstrations of the SPELL systems and permitted free hands–on use of the workstation. Each system evaluator completed a questionnaire and provided a report on his/her experience during the evaluation.

The rhythm and vowel teaching modules were demonstrated in isolation as described in the sections above. The intonation teaching module was presented as a part of larger system which incorporated a complete set of intonation courseware for a single contour. This larger demonstration provided the evaluators with an example of the courseware which can be produced using the pronunciation teaching paradigm **DELTA** developed within the project (see Hiller et. al 1991). This paradigm has four phases: ***Demonstrate*** in which audible demonstrations with various utterances are used to highlight the pronunciation features of interest; ***Evaluate Listening*** where small tests are completed by the student to evaluate his/her ability to perceive those features; ***Teach*** in which students practice the pronunciation features of interest, with quantitative feedback for the student and directions for modifying inadequate performance; ***Assess*** in which a formal evaluation of the student's ability to pronounce the features is made.

Evaluation results

The results of the evaluation were compiled and a number of general observations and recommendations were found for the SPELL systems. The evaluators generally agreed that the system as it stands at present is not ready to be used as an autonomous teaching system, though all evaluators were impressed by the potential of the system in future computer–aided language teaching applications. This was to be expected since the demonstration was presented as a "work in progress" system and not as a final product. Improvements were suggested in several areas: feedback to the student, the DELTA paradigm, the complexity of the language and the coverage of the courseware.

The feedback to the student was felt to be generally appropriate, but it was suggested that the system could benefit from improvements in the SPELL teaching modules. In the intonation teaching module, more detailed diagnostic feedback was considered desirable (this type of error analysis could be difficult since recognition techniques would be required to describe particularly aberrant F0 movements). In the rhythm teaching module, several improvements to the display were suggested, to enhance the contrast between the representation of strong and weak syllables. It was felt that the dimensions of the block representing the weak syllable should be reduced in width as well as in height, to give the greatest possible contrast in size, and to reduce any tendency the students might have to interpret the height difference as one of intensity. The vowel teaching display might benefit, it was felt, from some indication of the adjustments required to bring the student's production nearer to the target vowel articulation. Some relatively simple additions were suggested, such as an indication of the lip position (spread or rounded) and the degree of jaw opening required for a given vowel.

The principle of the **DELTA** paradigm was felt to be useful and to contain most of the elements required for teaching. The evaluators suggested that the intonation courseware was too rigid and that a greater degree of control should be allowed to the student (e.g. facilities to review earlier sections of a lesson and the option of skipping some sections if desired). It was also suggested that the **DELTA** paradigm should include an intermediate stage between the *Teach* and *Assess* modules which would allow some interim self–assessment using the materials already practised at the *Teach* stage.

Some simplification of the terminology of the courseware and interface was felt to be necessary, particularly for use by beginners. An on–line "Help" facility should be provided, to allow students to obtain information at any stage of the teaching paradigm. Extending the coverage of the courseware modules was felt to be a fairly high priority in all areas. In particular, coverage of diphthongs in English, and nasal and front rounded vowels in French, were issues which should be addressed as soon as possible. In addition, the courseware for French and Italian vowels should move away from the use of vowels in isolation , to more natural tasks using vowels in a variety of phonetic contexts.

CONCLUSION

The work of the SPELL project has demonstrated the feasibility of developing speech technology tools for the remediation of non–native speech. In on–going work on the project, the teaching modules will be expanded and more complete courseware will be developed using the **DELTA** paradigm. The vowel analysis will be expanded to include diphthongs and the French nasal and front rounded vowels. Teaching modules will be developed for consonants in the three languages. Finally, all the SPELL teaching modules will be evaluated in extensive student field trials.

NOTE

(1) This project was supported by the European Community's ESPRIT program, Contract No. 5192.

REFERENCES

Hiller, S., Rooney, E., Laver, J., di Benedetto, M.–G. and Lefèvre, J.–P. (1991) "Macro and Micro features for Automated Pronunciation Improvement in the SPELL System", Proc. ESPRIT '91, 378–392.

McCandless, S.S. (1974) "An Algorithm for Automatic Formant Extraction using Linear Prediction Spectra", IEEE Trans. Signal Processing ASSP–22, 135–141.

McInnes, F.R., Carraro, F., Hiller, S.M. and Rooney, E.J. (1992) "Evaluation and Optimisation of a Segmenter for a PC–based Pronunciation Teaching System", Proc. Institute of Acoustics, in press.

Medan, Y., Yair, E. and Chazan, D. (1991) "Super Resolution Pitch Determination of Speech Signals", IEEE Trans. Signal Processing ASSP–39, 40–48.

Rabiner, L.R., Sambur, M.R. and Schmidt, C.E. (1975) "Applications of Nonlinear Smoothing Algorithm to Speech Processing", IEEE Trans. Signal Processing ASSP–23, 552–557.

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. 68, 1465–1475.

Van Bergem, D.R. (1991) "The Influence of Sentence Accent, Word Stress and Word Class on the Quality of Vowels", Proc. Eurospeech 91, 1455–1458.