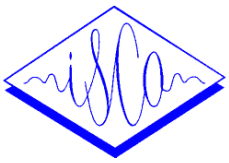

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

SPELL: AN AUTOMATED SYSTEM FOR COMPUTER–AIDED PRONUNCIATION TEACHING

Steven Hiller*, Edmund Rooney*, Jean–Paul Lefèvre** and Mervyn Jack*

* *Centre for Speech Technology Research, University of Edinburgh, Edinburgh, Scotland*
** *Agora Conseil, Sassenage, France*

ABSTRACT

This paper describes the application of speech technology in a workstation to improve the pronunciation of foreign language students. The SPELL workstation uses techniques of speech analysis to assess and improve learners' pronunciation in modules for teaching consonant production, vowel quality, rhythm and intonation, in three European languages (English, French and Italian). Each teaching module is discussed in terms of its phonetic basis, the implementation of its analysis modules and a description of the associated graphic user interface.

Keywords: *Pronunciation, Speech Analysis, Speech Technology, Consonants, Vowels, Rhythm, Intonation*

10.21437/Eurospeech.1993-314

1. INTRODUCTION

The development of automated systems for teaching pronunciation is a recent innovation in the field of speech technology. The main concern of speech technology research in the 1980's was the development of core technologies for speech synthesis, speech/speaker recognition and coding. SPELL (Interactive System for Spoken European Language Training) is a four–year ESPRIT project which has adapted many of these core technologies to the task of improving the pronunciation of foreign language students [1]. This project is developing systems for computer–aided pronunciation teaching for English, French and Italian. University and industrial groups from the three member countries have joined together to contribute expertise in digital speech processing, phonetics/phonology and computer–aided instruction.

The SPELL project is being conducted in two 2–year phases. The result of the recently completed first phase is a demonstrator system running on an IBM–PC compatible platform using the Microsoft Windows® graphical environment. The speech signal processing for the analysis of vowels, rhythm and intonation is executed on the OROS AU–21 DSP board. The second phase is responsible for the completion of the basic phonetic research (in particular, the addition of consonant teaching) and the preparation of the SPELL system for future commercialisation.

This paper presents an overview of the research work of the SPELL project, to be read in conjunction with other papers which detail specific research themes from the project (see [2, 8, 9, 12]).

2. CONSONANTS

The analysis of consonants provides a useful illustration of the integration of linguistic knowledge and speech technology being achieved on the SPELL project. The first stage in the development of consonant teaching modules has been a survey of the likely errors in consonant production by non–native speakers of English, French and Italian. This has involved a comparison of the consonant systems of each language, to characterise the principal differences between each language pair and to predict the range of possible errors in consonant production by non–native speakers. Errors have been identified at a *systemic* level (e.g. the substitution of /ʃ/ for /tʃ/ by native French speakers learning Italian), at a *structural* level (e.g. the omission of pre–consonantal and pre–pausal /r/ by native English speakers learning French), and at a *realizational* level (e.g. the absence of aspiration in English voiceless stops as produced by native French and Italian learners).

While there is a large number of possible errors in each source–target language combination, particularly at the level of acoustic realisations, not all such errors are equally important in terms of teaching priorities. The principal consideration is the effect of the error on the student's intelligibility. Some errors, such as the use of dental /t/ for English alveolar /t/ by native French and Italian learners, have very little effect on the listener's understanding of the speaker, while others such as systemic substitutions may be highly destructive of intelligibility. The errors identified in each language combination have therefore been ranked according to their expected effect on intelligibility, and a "short–list" of candidates drawn up for the development of teaching modules. This has considerably reduced the number of sounds which need to be considered as a matter of priority for the language student. Further prioritisation is also being undertaken, on the basis of the frequency with which these sounds occur in the target language.

Most of the errors identified take the form of substitutions of one sound – such as an allophone of one of the native language phonemes, or another phoneme of the target language – for the

intended target language sound. In the SPELL consonant analysis system, these categorical errors are detected by using a Hidden Markov Model (HMM) segmenter [5], which labels the incoming speech using the sequence of Acoustic Phonetic Units (APUs) specified in an associated segment transition network (see Figure 1). This network contains the desired se-

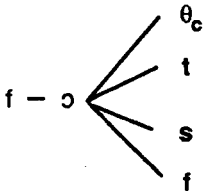

Figure 1. Segment transition network for the English word *fourth* showing the use of alternative APUs which may be produced by a native Italian speaker. The subscript $_c$ indicates the correct APU.

quence of segments in the test phrase, along with all the likely substitutions which the student might make. APU models corresponding to each of these substitutions are available to the segmenter. The choice made by the segmenter in processing a given phrase is thus used to determine the error produced by the student.

Consonant teaching modules are now being developed for the high priority errors identified in each language pair. These teaching modules will use minimal pair examples to highlight the contrast between the target sound and the typical non–native speaker substitutions. In many cases, these substitutions themselves may be used as starting points for teaching the articulation of target sounds: French and Italian dental stops /t/ and /d/, for example, provide a useful starting point for teaching the dental fricatives of English. The use of some basic articulatory information (the position of the lips, tongue and jaw, for example) will also be considered as part of the teaching strategy, using multimedia graphics where appropriate.

3. VOWELS

A SPELL module has already been developed for teaching monophthong vowels in the three target languages. A similar phonetic approach to that of the consonants has being adopted, involving the identification of the most important errors in vowel production for each language pair. However, the analysis required to assess the quality of a student's vowels is rather different.

A vowel similarity metric determines whether a given student's vowel token falls within a vowel space derived from multiple tokens spoken by a group of native speakers. Vowel tokens are represented in terms of two normalised formant parameters ([10]; see next paragraph). Each target vowel space represents an elliptical area in this two–dimensional space, covering up to 2 standard deviations on either side of the mean on each formant parameter.

The acoustic analysis which provides this representation first uses the HMM segmenter to isolate the vowel from surrounding speech sounds and silence. The first three formant frequencies are estimated using a modified McCandless algorithm [4], along with fundamental frequency (F0). The most stable region of the vowel is then chosen by a "steady–state" finder algorithm [11], and the four acoustic parameters are then averaged across this stable vowel region. The normalized formant parameters are then obtained by transforming measurements to a Bark scale and calculating the formant differences F1–F0 (corresponding approximately to the articulatory dimension of tongue height) and F2–F1 (corresponding approximately to tongue frontness–backness) [10].

These acoustic parameters are then used to provide feedback in a graphical display for the student. Figure 2 shows a typical example of the user interface for a native English student studying isolated French vowels. The general interface con-

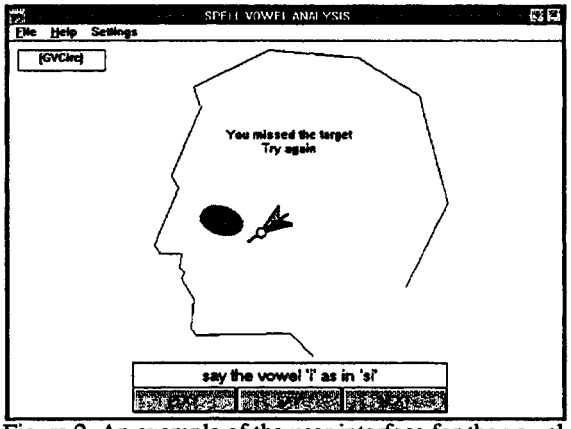

Figure 2. An example of the user interface for the vowel teaching module for an English student learning French.

forms to the common user access conventions of the Microsoft Windows® graphical environment. The main vowel display is located above a small prompt window and three user buttons. PLAY allows the student to listen to the teacher's model, SAY records and analyzes the student's own attempt, and NEXT lets the student choose another vowel for practice. In the main display, a two–dimensional elliptical vowel target for the French vowel /i/ is superimposed on the outline of a head, to provide an approximate reference for the articulatory position of the vowel target (high–front). The position of the normalized vowel formants derived from the student's attempt is marked by the tip of a dart. A feedback message indicates that in this case the student has missed the target (the token is slightly low and too far back). The dart's position relative to the target may be used by the students in a "biofeedback loop" to correct their pronunciation. Courseware for vowels, along the lines of that envisaged for consonants, is currently being implemented to make use of this interface within full teaching modules.

4. RHYTHM

A major contributor to rhythm in English, French and Italian is the relationship between strong and weak syllables. In all three languages, syllable strength is marked acoustically by variations in one or more of the parameters of F0, intensity, duration and vowel quality (i.e. formants). However, the frequent occurrence of strong syllables in English and Italian, contrasting markedly with intervening weak syllables, produces a characteristic rhythmic alternation which does not occur in French.

In addition, French minimizes the distinction between strong and weak syllables when compared with Italian and English.

The SPELL module for rhythm aims to improve the rhythmic quality achieved by learners by concentrating on a sub–set of the available parameters, namely duration and vowel quality. Thus learners of English are taught to produce weak syllables with centralized vowel quality and reduced duration; learners of French, conversely, must avoid any centralization of vowel quality or reduction in duration; while learners of Italian should aim for an intermediate position, contrasting duration but keeping vowel qualities uncentralized (the parameters of F0 and intensity have yet to be included in this analysis).

In the analysis of rhythm, the parameters of duration and vowel quality are derived indirectly, using the HMM segmenter. APU models are created for a variety of realizations of a given vowel, and are included as options in the segment transition network which governs the operation of the segmenter for a prescribed phrase. The choices made by the segmenter in processing the student's utterance are used to decide the rhythmic status of each syllable (as *strong* or *weak*). An example for English is given in Figure 3. The lessons in English rhythm concen-

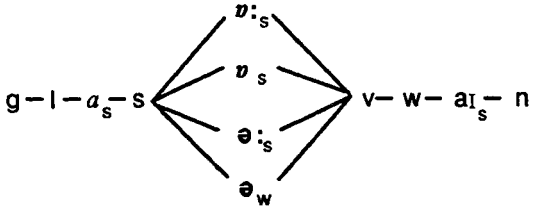

Figure 3. Segment transition network for the phrase *glass of wine* showing the use of alternative APUs to determine the rhythmic status of a syllable. Subscript $_s$ indicates a strong syllable nucleus, subscript $_w$ a weak one.

trate initially on teaching a small set of weak forms of function words (*for, the, to, some* etc.), to give the students control over strong and weak syllables. The segment transition network for a phrase such as *glass of wine* (Figure 3) allows several realizations of the word *of*. The choice of the top citation form by the segmenter 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 the bottom schwa would indicate that the student had made the syllable *weak*, and the student's attempt would be judged to be correct. Two intermediate levels are also allowed for, one using the citation–form vowel quality but with an appropriately short duration, the other using the correct central vowel quality with a prolonged duration.

Figure 4 shows a typical example of the user interface for the rhythm teaching module, in this case for a student studying English rhythm. The student controls the rhythm teaching module using the PLAY , SAY and NEW buttons (see the Vowel interface above). At the top, the *teacher window* displays the target rhythmic pattern for the current phrase; in this example, the student's task is to achieve a weak syllable for the function word *of* in the phrase *glass of wine*. The target display includes the orthographic representation of the phrase, labels indicating the target rhythmic status (*S* = strong; *w* = weak), and graphic blocks for reinforcement (tall = strong; short = weak). The *student window* displays in a similar form the results of

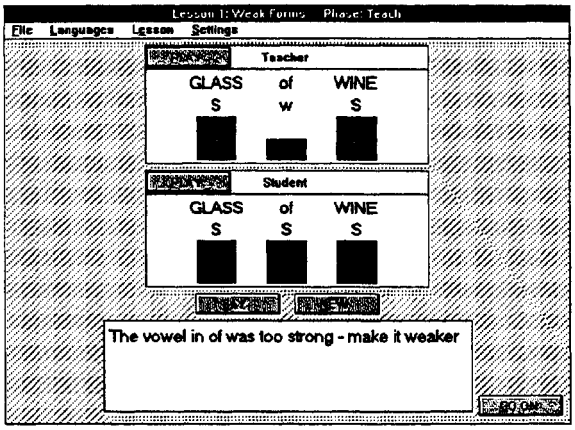

Figure 4. An example of the user interface for the rhythm teaching module for a student learning English.

analyzing the student's utterance. In this example, the label and graphic block indicate that the student has incorrectly used a strong syllable in the word *of*. The *diagnosis window* at the bottom displays a message indicating that the vowel in the word *of* was too strong and that the student should weaken it.

Complete teaching modules for rhythm in English, French and Italian are now under development.

5. INTONATION

The description and analysis of intonation adopted for the SPELL system has a practical phonetic basis which allows the major F0 movements associated with the contours of all three languages to be described using a common terminology and analyzed with a single similarity metric [3]. Central to this analysis is the relationship between the fundamental frequency contour of an utterance and its associated segmental sequence.

A number of stylized contours in each language have been chosen for the initial implementation (see [3] for details). Each contour is schematised using a set of *pitch anchors* – specifying the segmental locations of the major turning points and discontinuities within a stylized contour – and a corresponding set of *pitch tunnels* – describing the tolerances allowed for the path taken by an intonation contour between two pitch anchors. Figure 5 shows the schematization of the intonation contour used to teach simple French declarative statements, which consists of a *rise–fall* pattern. The three pitch anchors (i.e. the rec-

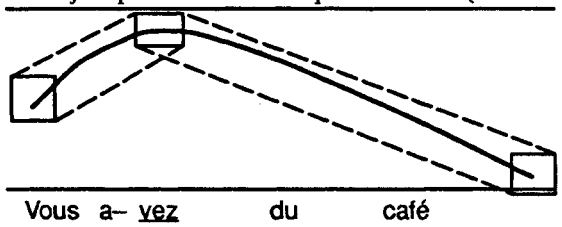

Figure 5. The use of pitch anchors (rectangles) and tunnels (dashed lines) to schematise the model pitch contour used for teaching French statement intonation. The solid line represents the rise–fall pitch pattern.

tangles) indicate that the pitch goes from mid to high to low. The location of the turning point (the middle rectangle) is de-

termined by rule, occurring on the final syllable of the first lexical word. The height and width of each anchor indicate the variability allowed in terms of pitch height and segmental location respectively. Linear interpolation between these zones of tolerance at each anchor produces a pitch tunnel (the dashed lines) through which any acceptable F0 contour must pass. Feedback can be given on each component of the contour between any two anchors, making this approach suitable for both whole–contour and componential treatments of intonation. Each contour selected for teaching in the three languages can be modelled phonetically in this way.

Two analyses are performed on each input utterance: the derivation of a smoothed, normalized fundamental frequency contour and the extraction of the segmental sequence. The F0 contour is derived from the low–pass–filtered speech waveform by a modified super–resolution pitch determination algorithm [2, 6]. The contour is heavily smoothed by a non–linear smoother [7], 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 the Hidden Markov Model segmenter described above; the segment transition network which specifies the possible sequence of segment labels allows for a variety of alternative pronunciations, including errors predictable from the student's mother tongue, to ensure an accurate segmentation.

Figure 6 shows a typical example of the user interface for the SPELL intonation module, for a student studying English statement intonation. The student controls the intonation

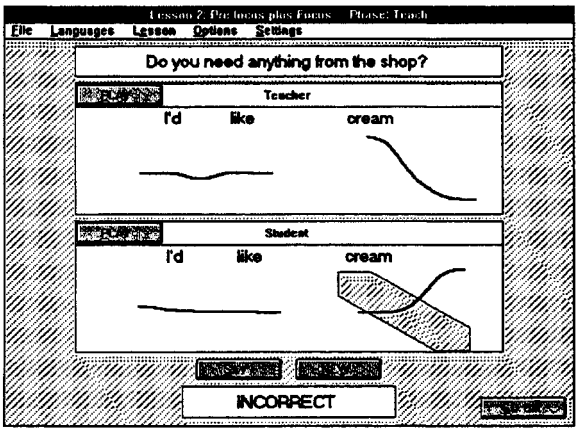

Figure 6. An example of the user interface for the SPELL intonation teaching module.

teaching module using the PLAY, SAY and NEW buttons (see the Vowel interface above). Statement intonation is taught using simulated dialogue, with the *question window* (top) providing a suitable context for the student's "reply". The *teacher window* (middle) displays the target intonation for the current phrase, with the teacher's smoothed F0 contour and a time–aligned orthographic representation to indicate the location of the intonational movements. The *student window* (bottom) displays the results of analyzing the student's utterance in a similar form, but with the facility to highlight errors in the student's production. In this example, the student has used the wrong intonation at the tonic ("cream"), and the error is highlighted by the appearance of the correct pitch tunnel at this point. In addition, a feedback window (bottom) displays a CORRECT or INCORRECT message, and a system voice announces "Well done" or "Try again" accordingly.

Complete courseware modules for English intonation have now been implemented, and work on modules for French and Italian is currently under way.

6. SUMMARY

This paper has presented the phonetic and speech technology aspects of the SPELL project on computer–aided pronunciation teaching. Modules have been implemented for teaching vowels, rhythm and intonation to learners of English, French and Italian. Each module has been built on a solid phonetic foundation and structured using sophisticated signal processing techniques on a PC. Work is now in progress to complete a module for teaching those consonants of each language which have the greatest effect on intelligibility when mispronounced. It is the aim of the present research to improve all the modules in order to bring the SPELL system to the market.

REFERENCES

[1] This project is supported by the European Community's ESPRIT program, Contract Nos. 5192 and 7153.

[2] *Bagshaw, P.C.; Hiller, S.M.; Jack, M.A.:* Enhanced Pitch Tracking and the Processing of F0 Contours for Computer Aided Intonation Teaching. In these proceedings.

[3] *Hiller, S.; Rooney, E.; Laver, J.; di Benedetto, M.–G.; Lefèvre, J.–P.:* Macro and Micro Features for Automated Pronunciation Improvement in the SPELL System. Proc. ESPRIT '91, pp. 378–392, 1991.

[4] *McCandless, S.S.:* An Algorithm for Automatic Formant Extraction using Linear Prediction Spectra. IEEE Trans. Signal Processing, Vol. ASSP–22, pp. 135–141, 1974.

[5] *McInnes, F.R.; Carraro, F.; Hiller, S.M.; Rooney, E.J.:* Evaluation and Optimisation of a Segmenter for a PC–based Pronunciation Teaching System. Proc. Institute of Acoustics, Vol. 14, pp. 109–116, 1992.

[6] *Medan, Y.; Yair, E.; Chazan, D.:* Super Resolution Pitch Determination of Speech Signals. IEEE Trans. Signal Processing, Vol. ASSP–39, pp. 40–48, 1991.

[7] *Rabiner, L.R.; Sambur, M.R.; Schmidt, C.E.:* Applications of Nonlinear Smoothing Algorithm to Speech Processing. IEEE Trans. Signal Processing, Vol. ASSP–23, pp. 552–557, 1975.

[8] *Rooney, E.; Eckert, M.; Vaughan, R.; Hiller, S.; Laver, J.:* Training consonants in a computer–aided system for pronunciation teaching. In these proceedings.

[9] *Rooney, E.; Vaughan, R.; Hiller, S.; Carraro, F.; Laver, J.:* Training Vowel Pronunciation using a Computer Aided Teaching System. In these proceedings.

[10] *Syrdal, A.K.; Gopal, H.S.:* A Perceptual Model of Vowel Recognition based on the Auditory Representation of American English Vowels. J. Acoust. Soc. Am., Vol. 68, pp. 1465–1475, 1986.

[11] *Van Bergem, D.R.:* The Influence of Sentence Accent, Word Stress and Word Class on the Quality of Vowels. Proc. Eurospeech 91, pp. 1455–1458, 1991.

[12] *Wang, H.D.; Degryse, D.; Carraro, F.:* A Prosody Modification Approach for Auditory User feedback in the SPELL Pronunciation Teaching System. In these proceedings.