Speech Communication 11 (1992) 31-44
North-Holland

Macro and micro features for automated pronunciation improvement in the SPELL system*

Jean-Paul Lefèvre**
OROS S.A., 13, Chemin des Prés, ZIRST-BP 26, 38241 Meylan cedex, France

Steven M. Hiller, Edmund Rooney, John Laver
The Centre for Speech Technology Research, University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, United Kingdom

Maria-Gabriella Di Benedetto
University of Rome, "La Sapienza", INFO-COM Dpt, Via Eudossiana 18, 00184 Roma, Italy

Received 4 September 1991
Revised 18 October 1991

Abstract. In this paper, 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 the relatively simple acoustic features of vowel quality and segmental duration. The micro feature analysis will focus on the segmental class of vowels. A distinctive feature approach is used to characterize non-native vowel pronunciation. Acoustic properties are sought which will be speaker-independent.

Zusammenfassung. In diesem Beitrag stellen wir eine Reihe von makroprosodischen und mikrosegmentalen Merkmalen vor um die Aussprache zu beschreiben. Diese Merkmale werden benutzt im Zusammenhang mit einem rechnergestützten Arbeitsplatz zur Verbesserung der Aussprache von Personen welche Fremdsprachen lernen. Drei Sprachen sind vorgesehen: Französisch, Englisch und Italienisch. Das System, welches entwickelt wird im Rahmen des SPELL Projekts, ist ein Mittel zum Studium für Benutzer mit mittlerem Können. Audio und visuelle Hilfsmittel werden angewendet um dem Studenten zu helfen seine Aussprache zu verbessern. Dies wird vollzogen in Rahmen eines allgemeinen Lehrparadigmas. Die prosodische Analyse, auf makroskopischem Niveau, bezieht sich auf die Melodie, die Dauer, den Akzent und den Rythmus. Der phonologische Rahmen für die Behandlung der Intonation beruht auf einem strukturierten System von Einheiten welche leicht mit diskreten linguistischen Funktionen in Verbindung gebracht werden. Der Akzent und der Rythmus werden behandelt auf Grund einfacher akustischer Merkmale wie die Vokalqualität und die relative Dauer der Segmente. Die Analyse der mikroprosodischen Merkmale konzentriert sich auf die Vokale. Unterscheidungsmerkmale werden dazu benutzt um die Aussprache der Vokale von Personen zu beschreiben welche nicht ihre Muttersprache sprechen. Es wird versucht akustische Merkmale zu bestimmen welche nicht vom Sprecher abhängen.

Résumé. Dans cet article, nous décrivons un ensemble de traits, tant macro que microscopiques, permettant de caractériser la prononciation. Ces traits forment la base d'un poste de travail développé en vue d'améliorer la prononciation d'une langue étrangère. Initialisé avec trois langues: le français, l'anglais et l'italien, le système mis au point dans le cadre du projet SPELL se veut un outil d'enseignement destiné à des utilisateurs de niveau moyen. Imbriquées dans un environnement intégré nommé DELTA, des informations de type graphique et sonore seront employées pour aider les étudiants à améliorer leur prononciation. L'analyse prosodique, de niveau macroscopique, prendra en compte les aspects intonation, durée, accent et rythme. L'approche phonologique utilisée lors du traitement de l'intonation fournit un système bien structuré d'unités contrastées faciles à corréler avec des fonctions linguistiques discrètes. Dans le système envisagé, l'accent et le rythme seront enseignés de manière moins approfondie en s'appuyant d'une part sur les traits acoustiques relativement simples liés à la qualité des voyelles et d'autre part sur la durée relative des segments. L'analyse des traits microprosodiques va être focalisée tout particulièrement sur les voyelles. Une approche à base de traits discriminants est utilisée pour caractériser la prononciation de voyelles par un locuteur ne parlant pas sa langue maternelle. Il est prévu de mettre en oeuvre des propriétés acoustiques si possible indépendantes du locuteur.

Keywords. Speech analysis, prosodic features, segmental features, pronunciation aid.

* This project is supported by the European Community's ESPRIT program, under contract No. 5192.
** J.-P. Lefèvre is now with AGORA CONSEIL, 185 Hameau du Château, 38360 Sassenage, France.

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 research and development of the SPELL workstation addresses many of the concerns of the ESPRIT program. Upon successful completion, the project will have produced a user friendly workstation which can be set alongside other learning tools in university and school language laboratories. 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. The analysis of these three languages naturally requires a high degree of international cooperation within the consortium. Expertise in phonetics, signal processing and systems development is distributed across the project members. The partners[1] in this collaborative project and their roles are as follows:

1. OROS S.A., 13 Chemin des Prés, ZIRST - BP 26, 38241 Meylan cedex, France (signal processing);
2. The Centre for Speech Technology Research, University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, Scotland (phonetic analysis);
3. ALCATEL FACE, Research Center, Via Generale Clark 21, 84100 Salerno, Italy (signal processing);
4. Tecnopolis CSATA Novus Ortus, 70010 Valenzano, Bari, Italy (user interface);
5. University of Rome, "La Sapienza", INFOCOM, Via Eudossiana 18, 00184 Rome, Italy (phonetic analysis).

This paper begins with a general description of the framework of the SPELL system within which the proposed macro and micro feature analysis (as defined below) will operate. This general discussion sets out some basic assumptions about the system and the phonetic issues which it must address. A general teaching paradigm for foreign language pronunciation called DELTA is also proposed. The discussion then focuses on the analysis of macro and micro features to be used within the SPELL system. The macro features comprise 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 feature analysis will focus on the segmental class of vowels using a distinctive feature approach to characterize non-native vowel pronunciation. Acoustic properties are sought which will provide speaker independence for vowels.

2. A general description of the SPELL system

This section discusses some general issues in the development of the SPELL workstation, beginning with an outline of the basic assumptions on which the research will be based.

2.1. Some basic assumptions about the SPELL system

ASSUMPTION 1. The SPELL workstation will be used as an autonomous teaching system.

A system for teaching foreign language pronunciation can be designed in two ways: (1) as an assistant workstation to a language teacher who organizes the particular pronunciation tasks to be

[1] Apart from the authors of this paper, the SPELL team involves P. Broccoli, F. Carraro, G. De Santis, F. Gabrieli, M. Jack, C. Maggio, F. McInnes, M. Refice, L. Santangelo, E. Savino and H. Wang.

practised by the student while using the system and (2) as an autonomous teaching system which is used without the need of direction by a teacher. The second approach is assumed here, thereby forcing a complete consideration of all the possible issues which might affect the performance of the system. Most importantly, courseware in some narrowly-defined topic areas will need to be created in order to achieve an autonomous teaching system.

ASSUMPTION 2. Users will be intermediate ability foreign language speakers.

This assumption avoids the need for the SPELL system to teach the basics of language use as well as pronunciation. It will need to be determined what is meant by 'intermediate ability' and pronunciation tests will be devised accordingly for the three SPELL languages.

ASSUMPTION 3. Audio and visual aids will be available to the user.

These are two of the major technical interests of the SPELL project. The audio aids will enable the user 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. Visual aids must produce intelligent displays which help a student to visualize relevant linguistic/phonetic concepts without requiring expert knowledge.

ASSUMPTION 4. Speaker intelligibility will be used to gauge pronunciation improvement.

Mimicry is often used by foreign language teachers in an attempt to achieve fluency in a given foreign language. In fact, there are very few foreign language learners who genuinely need to produce fully native versions of pronunciation. For the majority of students, improvement in intelligibility is a more practical objective in the successful use of a foreign language. It should be noted that mimicry is required to achieve success at pronouncing unfamiliar sounds but that functional ability to communicate with foreign listeners is of more interest than "speaking like a native" (see, for example, (Rivers, 1975; Harmer, 1983; Madsen, 1983)).

2.2. *The development of SPELL courseware*

In this section, some practical aspects in the development of courseware for the SPELL workstation are presented.

Courseware useful for improvements in pronunciation can be divided into two general types: practice courseware and teaching courseware. In the case of practice courseware, only quantitative feedback is provided to the student without any further directions from the system. The main drawback to this approach is that improvements in pronunciation performance are not very predictable. 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 development of teaching courseware will concentrate research efforts towards a system which touches on all levels of pronunciation teaching.

The typical aspects common to many foreign language teaching practices can be described in a paradigm called DELTA. This system is described as follows:

Demonstrate. Audio demonstrations by the system of various utterances are used to highlight the pronunciation features of interest. For example, differences in vowel quality would be demonstrated by playing out words which contain minimal pairs such as *beet* versus *bit* for the /i/ versus /I/ vowel distinction in English. The same pair of words might appear in sentences to demonstrate the phonological significance of the vowel distinction (e.g. "I would like a bit/beet").

Evaluate Listening. Small listening tests are completed by the student to evaluate his or her ability to perceive the pronunciation features of interest. For example, the student would take an ABX test: the two utterances (A) "bit" and (B) "beet" are played to the student, who must decide if the test utterance X is the same as A or B. If the student fails to perceive the distinction for a given pronunciation feature then he or she may be asked

to go back to the *Demonstrate* stage for more examples.

Teach. The actual teaching of the pronunciation features of interest takes place at this stage, with quantitative feedback for the student and directions for modifying inadequate performances.

Assess. This stage is the formal evaluation of the student's ability to pronounce the features of interest. For example, the student is given N attempts to produce utterances containing the features of interest. If the student achieves a certain percentage of corrent pronunciations then he or she should proceed onto the next features to be taught; otherwise the students should go back to the *Demonstrate* phase for this particular task.

In addition to the DELTA paradigm, a fuller assessment of proficiency can be given to the students after several SPELL lessons have been completed in order to evaluate the student's general performance while using the workstation. For example, a close test could be given in which the features of interest are embedded within a section of contrived text. The text is designed in such a way that listening judges can evaluate the features of interest without the student being aware of which features are being tested. (Bernstein et al. (1990) studied the feasibility of automatically grading the performance of Japanese students speaking English and found a good correlation between evaluations completed by human judges and the automatic system.) Most importantly, this assessment program provides the SPELL project with a means of evaluating the performance of the demonstrator system.

2.3. *Problem areas for research and development of SPELL*

A number of problem areas for the development of the SPELL workstation are discussed in this section.

The integration problem

A useful pronunciation-teaching system will not be created if it relies solely on the teaching of phonemes in isolation. In the early stages of teaching, isolated phonemes may be useful for simple demonstrations and training, but more natural data must also be used in which phonemes have been integrated in real speech items or phrases.

The equivalence problem

In speech technology research, there is often an assumption of a direct equivalence between information contained in the acoustic speech waveform and the resultant phonetic perception of that signal. However, this belief in equivalence is not completely valid. For example, the ability to mimic a given pitch contour exactly does not guarantee any generalization of pitch use within a language. The reasons for this are twofold. Firstly, native listeners are very sensitive to the linguistic relevance of small alignment/adjustment details of a pitch contour, and the linguistic relevance of the contour therefore springs partly from its integration with the segmental performance. Secondly, a given pitch contour acquires its functional value partly from its relative placement in the pitch range of the speaker concerned, not from its absolute value. It is therefore more desirable to concentrate on getting the student to imitate more abstract aspects of the contour such as the location of the pitch peak and the shape of the contour. In general, the system needs to be able to judge when the student has produced an acceptable version of a given feature of interest in relative terms rather than by absolute matching techniques.

The segmentation problem

Accurate feature analysis will depend on the location of phonetic segments within an utterance produced by a student. The analysis of vowel quality (micro) features certainly requires the prior location of the vowel segments. Proper prosodic analysis also cannot be completed without first locating those phonetic segments which have had durational and intonational features overlaid on their structures. Therefore, an automatic SPELL segmentation program is being developed for application to the speech signal prior to feature extraction and analysis.

Pronunciation errors

The SPELL system must be able to deal with a number of different types of pronunciation error produced by non-native speakers of a given language. Firstly, there are the structural errors,

brought from the mother tongue, which can take the form of additions to or omissions from the expected segmental sequence. For example, an Italian speaking English may tend to add vowels in order to preserve the syllabic structure of Italian, as in the word "bead" being pronounced /b i d e/. Secondly, systemic errors can occur in which the phoneme of interest does not exist within the speaker's mother tongue and the closest native phoneme is used as a regular substitution (e.g. a French person speaking English might tend to pronounce the word "bit" as "beat"). Thirdly, there are realization errors in which a version of the non-native phoneme of interest does exist within the native speaker's system but it is still not quite pronounced correctly (e.g. a French person speaking English may pronounce the vowel in the word "bead" using the correct quality but with the wrong duration). Finally, the speaker may produce gross mistakes such as misreading, stuttering, false starts, etc.

How can the input waveform be segmented correctly when such errors may exist? Firstly, the test utterances will be constructed in such a way as to limit the types of error which might occur. Secondly, the SPELL segmenter has been designed using a segmental transition network which includes the more predictable types of error which occur between two given languages.

Feedback

The appropriate feedback will have to be provided by the SPELL workstation for the user. In the case of vowel quality, it is felt that quantitative feedback on its own (i.e. without diagnostic information) would be appropriate, since it may be difficult to convey complex articulatory relationships to the linguistically unsophisticated foreign language learner. For example, basic vowel quality distinctions can be demonstrated by displaying targets within a vowel diagram which the student must hit with a "voice cursor" controlled by a vowel formant detection program. In effect, the student learns vowel quality distinctions via trial-and-error biofeedback rather than by active diagnostic feedback on the part of the SPELL system. Prosodic features appear to be more straightforward to describe to the student and therefore diagnostic feedback is appropriate (e.g. the workstation informs the student that he or she used a falling pitch contour at the end of the utterance but that a rising pitch would have been more appropriate). It should be emphasized that the feedback to a SPELL user will not be expressed in terms of explicit sophisticated linguistic or phonetic concepts: for example, the student need not be aware of the abstract phonological systems which form the bases of the training system.

3. The analysis of macro features in the SPELL system

The preceding sections set out a framework for the analysis and remediation of non-native pronunciations of a given language. In this section, the prosodic aspects of pronunciation are examined in more detail.

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 which may characterize an utterance as a whole. Features normally seen as prosodic 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 (e.g. (Lehiste, 1970, p. 83)). All utterances, including words spoken in isolation, have an intonation "contour". *Stress* is the term used to refer to a number of ways in which certain syllables are made more prominent than surrounding syllables. Stress functions at two levels: within the word (lexical stress or accent) and within the utterance as a whole (rhythmic stress), where it is closely integrated with intonation and rhythm. The *rhythm* of an utterance is given by the patterning in time of the segments, syllables and stresses; its actual definition and description remain the subject of controversy, and its measurement in terms of acoustic features is notoriously difficult (Adams, 1979).

3.2. Phonological approaches to intonation

The analysis of speech at the segmental level into phonemes is taken for granted, and it is often not

realized that there is a considerable body of linguistic theory underlying this analysis, which makes it possible to assume, for example, that the "p" sounds produced in words such as "put", "top" and "spin" can be regarded phonologically as "the same thing" despite some major acoustic differences. The analysis of prosody is still not as developed as that of segmental phenomena, but it is recognized that there is a similar need for a theoretical basis to make sense of the multiplicity of acoustic realizations of prosodic features which are found when real speech is examined (Cutler and Ladd, 1983). This is particularly important for language teaching. For example, 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. This is the essence of phonological analysis: a phonological treatment aims to establish a system of structures which can be used as a vehicle for meaning and a set of contrasting elements which can be inserted into those structures.

The development of a complete phonological description for each of the three target languages is beyond the scope of the SPELL project. Instead, published analyses of English, French and Italian will be used to achieve the aim of teaching the basic prosodic patterns of each language to students.

In intonation analysis, English has the most developed treatment with two main traditions: the so-called "British" school, which treats the pitch contour as the unit of analysis (e.g. (O'Connor and Arnold, 1961; Crystal, 1969, 1975; Halliday, 1973)); and the "American" school, which deals with distinctive pitch levels (e.g. (Trager and Smith, 1951; Liberman, 1975; Pierrehumbert, 1979, 1980)). Rather less work has been done on French intonation (examples, in a variety of approaches, are (Faure, 1973; Grundström, 1973; Martin, 1975, 1982; Kenning, 1979, 1983; Leach, 1988)). Analyses of Italian are fairly rare (Chapallaz, 1964, 1979; Muljacic, 1972).

Comparing intonational systems amongst these three languages in terms of their phonology is quite difficult given the varying depth of treatment and the differing approaches to the problem. 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 related to or anchored to the segmental structure of the utterance (this is the segmentation problem discussed above).

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.

3.3. The analysis of intonation

This section presents the analysis of intonation for French, Italian and English for the SPELL workstation.

The phonetic transcription of pitch phenomena is conventionally achieved by limiting the transcription to comment on relative pitch movements within a speaker's linguistic range. Two parallel horizontal lines are usually drawn to act as a stave representing the upper and lower limits of the linguistic pitch span; the use of horizontal lines means that the effect of pitch declination has been ignored (i.e., the tendency for fundamental frequency to slowly decrease from the beginning to the end of an utterance has been eliminated). This pitch transcription will be used in the remainder of the paper.

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.

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.

French intonation

According to Vaissiere (personal communication), French intonation is based on unitary pitch contours which are often called "tunes". These tunes are relatively simple in structure and the choices within them are very limited. Tune 1 is used for declarative statements, qu-questions and inverted polar questions, while Tune 2 is for non-inverted polar questions. For simple utterances, the overall contour for Tune 1 consists of a rise–fall pattern: the contour begins at a mid pitch level, rises to a high level located in the first lexical word (the only pitch anchor point in this tune) and then falls to a low level by the end of the utterance. For more complex utterances where Tune 1 contours are concatenated, there are two pitch anchor points since a small continuation rise occurs at the end of each contour but the last. Tune 2 starts at a medium pitch level, with a pitch trajectory towards a medium-high level anchor point, and then rises rapidly on the last syllable to a high level. Figure 1 displays schematic representations for the two French tunes to be taught as part of SPELL prosodic features.

Italian intonation

According to Chapallaz (1964, 1979), a tune analysis is also appropriate for Italian intonation. If these whole contours are examined in further detail then they can be described in terms of pitch anchor points and trajectories. Tune 1 is the usual intonation for statements, qu-questions, commands and exclamations. Tune 2 is the typical intonation for short, introductory non-final statements and polar questions; for these utterances, intonation is the only marker of interrogativity, since Italian, unlike French and English, does not have inverted questions forms or special lexical markers. These two tunes are generally similar in structure and differentiated by the final movement of the contours. Both tunes are broadly shaped by two pitch anchor points. The first anchor point is located in the first stressed syllable of the utterance at a high pitch level, while the second occurs on the last stressed syllable at a low level. The pitch trajectory falls evenly between these two anchor points. Any unstressed syllables before the first anchor point form a rising trajectory from a medium pitch level. The two tunes are differentiated by the final trajectory of the contours. For Tune 1,

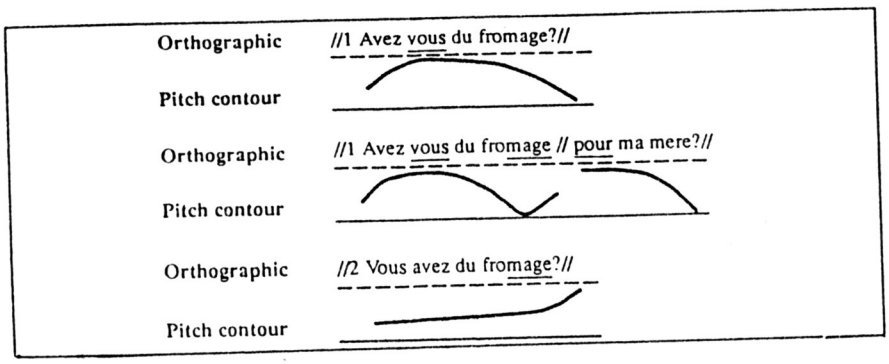

Fig. 1. Schematic representations of the two primary tunes for French intonation. The pitch anchor point for each utterance is underlined in the orthographic representations. The top part of the figure is for French Tune 1 for a simple utterance, while the middle part displays two Tune 1 contours linked together with a continuation rise. Tune 2 is displayed in the bottom part.

the contour remains at the low pitch reached by the falling trajectory at the second anchor point. The Tune 2 pitch trajectory rises sharply after the second anchor point. Figure 2 displays schematic representations for the two Italian tunes to be taught as part of SPELL prosodic features.

English intonation

Of the three target languages, English exhibits the greatest complexity for the structuring of intonation. That is, the whole contour can be divided into a chain with choices to be made at various locations. The phonology of English intonation has attracted a considerable body of recent research. A classic (and for SPELL's purposes, a very usable) treatment is by Halliday (1973). According to Halliday, there are three phonological systems at work in the intonation of English, namely tonality, tonicity and tone. *Tonality* is the system of options for dividing the stream of speech into units of intonational structure called tone-groups. *Tonicity* is the system of options for the location within the tone-group of the syllable receiving the most prominent pitch-movement (the tonic syllable). *Tone* is the system of choices of the type of pitch pattern over the tone-group up to and including the tonic syllable. The stretch of speech within the tone-group leading up to the tonic syllable is called the pre-tonic. The stretch of speech after the tonic syllable is called the post-tonic.

Within a single tone group, British English evidences a large variety of primary and secondary tones as well as numerous associated pre-tonic contours. For the SPELL project, Halliday's primary Tones 1 and 2 have been selected since they provide a substantial coverage of intonational uses within English. Tone 1 is used for declarative statements, wh-questions and imperatives. It consists of a falling pitch trajectory which originates at a high level anchor point and terminates at a low level anchor point. Tone 2 is used for polar questions and certain other attitudinal information and consists of a rising or falling/rising slope. For the SPELL project, Tone 2 will consist of a rising slope which originates at a low level anchor point and rises to a high level anchor point.

The choice of pitch contour shown by the pre-tonic stretch is contextually somewhat limited by the nature of the choice of tone. Tone 1 has the widest choice of different pre-tonic patterns that can precede the tonic syllable. In order to simplify the teaching task, one set pre-tonic contour will be taught to the student learning English. In the case of Tone 1, the pre-tonic will be a relatively flat pitch trajectory anchored at the mid level of the speaker's speaking pitch range. The pre-tonic for Tone 2 will be a relatively flat pitch trajectory anchored at the low level of the speaker's pitch range. The post-tonic choice of pitch pattern is completely prescribed, in that a tonic choice with a low terminal anchor point can only be followed by a low level post-tonic. Any tonic with a rising, non-low terminal tendency can only be followed within the same tone-group by a post-tonic pattern that continues the rising tendency. Figure 3 displays schematic representations for the two English tones to be taught as part of SPELL prosodic features.

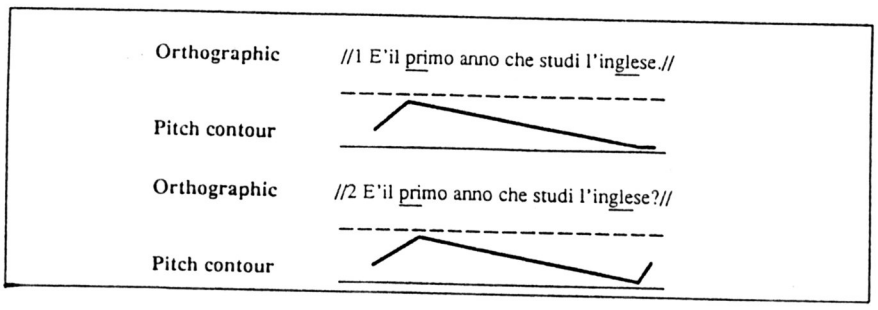

Fig. 2. Schematic representations of the two primary tunes for Italian intonation. The two pitch anchor points for each utterance are underlined in the orthographic representations. The upper part of the figure is for Italian Tune 1, while Tune 2 is displayed in the lower part.

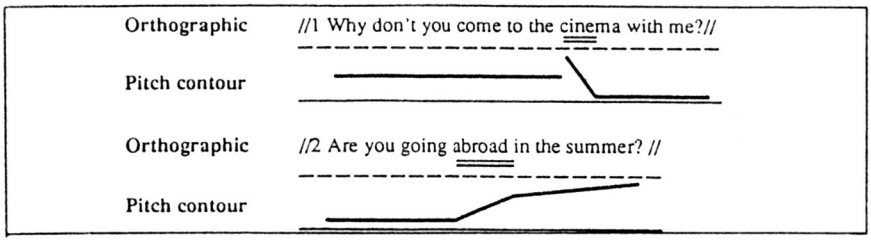

Fig. 3. Schematic representations of the two primary tunes and associated pre- and post-tonics for English intonation. The tonic syllable for each utterance is underlined in the orthographic representations, while double bars indicate that two anchors points are required here. The upper part of the figure is for English Tone 1 while Tone 2 is displayed in the lower part.

3.4. Stress and rhythm

Stress and rhythm are considered together here because, in English and Italian at least, they are intimately connected.

In most European languages, including English, French and Italian, stress or salience is marked acoustically by modulation of fundamental frequency, intensity, duration and segmental features. The way in which it is marked depends partly on the type of stress involved: lexical stress functions within a word to convey semantic or syntactic differences (e.g. English "concert" /'kɔnsət/ (noun) versus /kən'sə:t/ (verb)), while rhythmic stress conveys the rhythmic structure within an utterance as a whole.

Most work done on "stress" in the literature examines lexical stress, and the relative contribution made by the different acoustic parameters. Control of rhythmic stress is possibly of more significance for foreign language learners. It is particularly important for 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 are important anchor points for the pitch movements on which intonation depends. In both languages the differences between stressed and unstressed syllables are quite marked, although it is difficult to define the exact relationship between the underlying linguistic structure and measurable acoustic parameters. French 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).

The difference between languages which make rhythmic use of stressed syllables and those which do not has been formalized into a theory of speech rhythm which sees all languages as belonging to one of two types: stress-timed languages, in which the intervals between stressed syllables are controlled, and syllable-timed languages, in which control is concentrated on syllable durations (e.g. (Pike, 1945; Abercrombie, 1967)). The theory of stress timing has been used as a teaching device for English, learners being encouraged to maintain an equal interval of time between stressed syllables no matter how many unstressed syllables come between. However, the empirical basis of the theory has been called into question by a large amount of experimental work which has failed to find any objective equality either of inter-stress intervals in "stress-timed" languages (e.g. (Roach, 1982; Dauer, 1983)) or of syllable durations in "syllable-timed" languages (e.g. (Wenk and Wioland, 1982)). In addition, there appear to be a large number of languages which do not fit neatly into either category (e.g. (Miller, 1984)).

A more useful approach to rhythm appears to be emerging from these studies. It has been suggested that the perception of "stress-timing" and "syllable-timing" actually owes more to other factors in the languages under consideration, such as syllable structure, the nature of stress and the use of vowel reduction (Smith, 1976; Roach, 1982; Dauer, 1983)). According to Dauer (1983), languages which have been classified as "stress-timed" have a greater variation in syllable length and a greater variety of permitted syllable structures. They also permit a greater degree of vowel reduction: in English, for example, unstressed vowels are

typically reduced to /ə/ or /I/, while in French, the only permitted reduction – of "e-muet" – typically leads to the loss of the whole syllable. Stress-timed languages also tend to have full lexical stress. Dauer proposes that languages should therefore be considered as being more or less "stress-based" according to their tendencies on parameters such as these.

This leads to a convenient classification for the purposes of the SPELL project, and one which is more amenable to automatic processing than that provided by adherence to a strict stress-timing/syllable-timing approach (see Fig. 4). Thus, English is at one extreme 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 (that is, the vowels have been centralized). Italian, while also strongly 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. Indeed, the perception of stress by native French speakers is notoriously unreliable (Leon and Martin, 1980), and many linguists deny that French has word accent at all. In addition, the absence of vowel reduction produces a rhythm entirely different from that of Italian and English.

Thus 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, rather than by completing a full analysis of the stress locations and segmental timing (see Fig. 5). Italian and English make a greater contrast by lengthening stressed vowels and shortening unstressed vowels, while French makes no such distinction. English centralizes unstressed vowels, but Italian and French only evidence non-centralized vowels. Thus 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

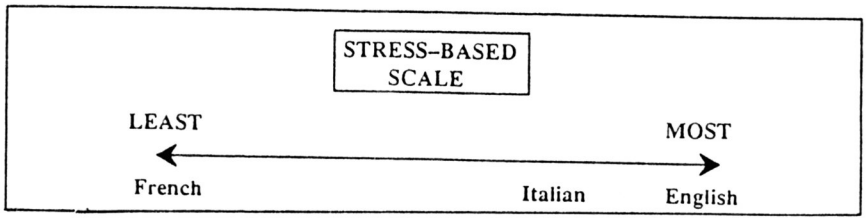

Fig. 4. Degrees of dependence on rhythmic stress following Dauer (1983). The degree of stress is dependent on several factors including lexical accent, syllable structure, segmental duration and degree of vowel reduction.

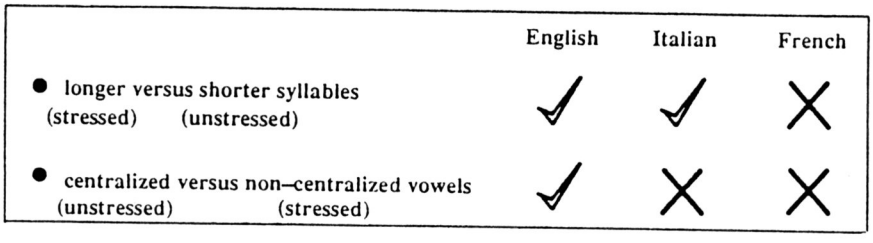

Fig. 5. Acoustic correlates of rhythmic stress for the three SPELL languages. A check mark means that a correlate is present, while a "×" represents an absence of it.

rhythm, and they are used for stress marking for all three languages.

4. The analysis of micro features in the SPELL system

In this section, the micro features of the SPELL system are presented. Micro feature is another term for phonetic segmental aspects of the speech signal. The SPELL research will be focusing on the segmental class of vowels for the final demonstrator system.

The aim of this aspect of the SPELL project is the identification of distinctive phonetic features and their acoustic correlates for the vowels of English, Italian and French. One of the main objectives is the determination of pronunciation features from a speaker's native language which affect the phonetic quality of non-native vowels uttered by that speaker.

An articulatory description of vowels in terms of distinctive phonetic features will be used in this area. The problem is then to identify the acoustic correlates of these features, and to verify that these correlates are perceptually relevant and independent of the speaker and the language.

There will be two main areas of research for micro features. The first area will address the characterization of non-native vowel production in terms of the distinctive features of the target language. For example, when an Italian speaker attempts to produce the French nasal vowel /õ/, a sequence of the non-nasal vowel /o/ followed by the nasal /n/ will often result. The vowel will be nasalized owing to the presence of the following nasal; however, in terms of distinctive features for French, the vowel is labelled as [-nasal]. The second area of research is to find those acoustic properties of vowels which prove to be speaker-independent. The system will need to represent, in terms of a speaker's own native vowel system, any non-native vowel for which there is no direct correspondent.

4.1. Identification of errors in foreign pronunciations of vowels

This section discusses the most common errors found for the non-native pronunciation of English, French and Italian vowels.

English vowels

The tense/lax high vowel contrasts /i ~ I/ and /u ~ U/, present in English, cannot be made by Italian and French speakers. The failure to distinguish these vowels often leads to confusion since there are several minimal pairs based on this distinction (e.g. the pair "beet" and "bit"). Though the extension of the tense-lax distinction to other vowel pairs is somewhat controversial, other pairs worthy of consideration are /ɛ ~ ae/, /ɑ ~ ʌ/ and /o ~ ʌ/.

The tense/lax pairs mentioned above also include some vowels which are not present in the Italian and French vowel systems (namely, /I, ae, ʌ, ɑ, U/). Training on the tense/lax pairs would also be appropriate, therefore, for the teaching of the pronunciation of these new vowels.

Note that Italian and French have only one low vowel /a/ which is centrally located (that is, no back/front distinction is made for low vowels). However, the English vowel system has two low vowels /ɑ/ and /ae/, the former having [+back] and the latter having [+front].

French vowels

There are a number of problems for non-native speakers pronouncing French vowels, namely the need to produce nasal and front rounded vowels as well as avoiding the tendency to diphthongize pure vowels.

Italian and English do not have vowels corresponding to the nasal vowels /ɛ̃, ã, õ/ of the French vowel system. Speakers of these languages tend to produce a nasalized vowel followed by the extra nasal /n/. A possible training paradigm for this problem would consist in pronouncing pairs of non-nasal/nasal vowels such as /ɛ ~ ɛ̃/, /a ~ ã/ and /o ~ õ/, which only differ in the position of the velum during articulation.

English and Italian lack the use of contrastive lip rounding in front vowels, and speakers therefore have difficulty achieving the correct lip position for the French front rounded vowels /y, Ø, œ/. Training for lip position will use pairs of the French rounded and unrounded vowels /y ~ i/, /Ø ~ e/ and /œ ~ ɛ/ which differ only in the degree of lip rounding or spreading.

In the case of native English speakers, the problem of diphthongization must also be considered.

The nearest English equivalents to the French pure vowels /e, o/ are the diphthongs /eI, ou/. While not classified as diphthongs in English, the vowels /i, u/ also tend to be diphthongized in English and this habit is carried over to English speakers' pronunciation of the corresponding pure vowels in French.

Italian vowels

There is little problem for French speakers learning Italian since all Italian vowels have correspondences in the French vowel system. In the case of native English speakers, the main problem is the diphthongization of pure vowels, as mentioned in the previous section.

4.2. Vowel coarticulation and vowel normalization

Two problems arise when representing vowels by means of acoustic parameters: vowel coarticulation and vowel normalization.

Vowel coarticulation

The phonetic context in which a vowel is spoken has a major influence on its articulation. This coarticulation effect gives rise to a range of different formant frequency values for that vowel within the production of one speaker, and may cause the acoustic parameters of two different vowel-phonemes to overlap. This causes considerable problems for vowel representation, since a given formant pattern then cannot be identified uniquely. Analyses have been carried out to find parameters which might be invariant with respect to the phonetic context (Di Benedetto, 1989a, b), but no useful alternative to vowel formant frequencies has been found.

Vowel context must therefore be held constant when vowels are being compared. This is to be achieved by the use of minimal pairs, where the consonantal context in which the vowels are embedded is identical.

Vowel normalization

There are two problems to be considered when comparing the same vowels produced by different speakers. Firstly, the same vowel phoneme may have a different acoustic realization for two speakers, owing to the differences in their vocal tract shape and dimensions. Thus, a given formant frequency pattern, measured in absolute terms, may be identified with one vowel in the speech of one person but with a different vowel produced by another. Some form of between-speaker vowel normalization is therefore required before the vowel spaces of two speakers can be compared. Secondly, cross-language normalization is also required owing to the multi-lingual nature of the SPELL project.

Between-speaker vowel normalization can be achieved by considering the normalized bark-scaled values using the first three formants plus the fundamental frequency, $F1-F0$, $F2-F1$, $F3-F2$, as suggested by Syrdal and Gopal (1986). Acoustic analyses have shown that vowels are adequately represented by these parameters in American English (Syrdal and Gopal, 1986) and Italian (Di Benedetto and Flammia, 1990). The $F1-F0$ dimension is associated with the distinctive feature high–low, and the $F3-F2$ and $F2-F1$ dimensions with the feature front–back. As an example, Fig. 6 shows the representations of all the Italian vowels in the $F3-F2$ versus $F1-F0$ plane for 13 male and 11 female speakers (Di Benedetto and Flammia, 1990). The parameters appear to normalize vowel differences across speakers successfully, though some overlap between vowel areas remains.

An alternative method of normalization being considered involves obtaining a representation of the speaker's peripheral vowels /i, a, u/ during a limited training phase (possibly covert). This could be used to locate the speaker's entire vowel space using the dimensions of $F1/F3$ versus $F1/F2$ (see, for example, (Minifie, 1973)).

Cross-language normalization presents a more complex problem, with two major issues. The first issue is whether a speaker uses similar formant frequency values for a foreign vowel which corresponds to a vowel of his or her native vowel system. In a SPELL micro feature pilot study, results from an acoustic analysis on isolated vowels of Italian and French uttered by a bilingual speaker (who did not know the purpose of the experiment) suggest that the Italian vowels have similar $F1$ and $F2$ values to those characterizing the corresponding vowels in French. The second issue is whether it is possible to predict the location within a speaker's

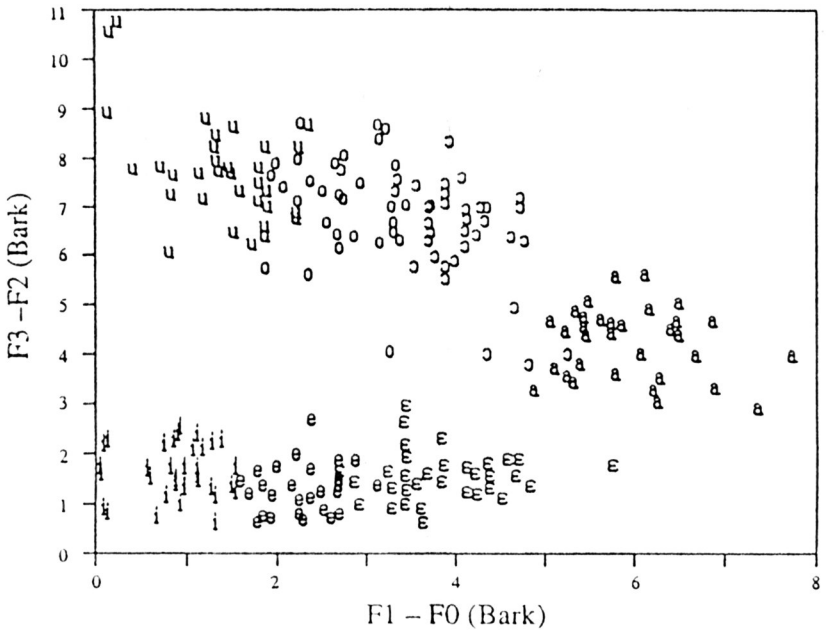

Fig. 6. A representation in the $F3-F2$ versus $F1-F0$ plane of all the Italian vowels. Data come from 13 male and 11 female Italian speakers. All formant values are expressed in units of Bark.

vowel space of a vowel which does not exist in his or her native language. This question is now under investigation.

5. Summary

This paper has presented the initial work completed by the SPELL project. Firstly, a basic description of the SPELL workstation was given: it is intended to be a teaching device aimed at intermediate ability foreign language learners with audio and visual aids being used to help students improve their general intelligibility within a basic teaching paradigm called DELTA. The two major phonetic areas of interest being covered by SPELL include macro (prosodic) and micro (segmental) features. A practical approach has been taken for teaching prosodic features of intonation, rhythm and stress. Whole intonation contours are broken down into easily teachable sub-units based on phonology where applicable. Rhythm and stress will be taught indirectly using the stress-based features of segmental duration and vowel quality. The micro features research is focused on the class of vowels. A number of problem areas have been identified including the English tense/lax vowel distinction and French nasal and rounded vowels. In addition, diphthongization of Italian and French pure vowels will need to be avoided by native English speakers.

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. An important feature of this segmenter is its ability to cope with the variety of realization, systemic and structural errors produced by a non-native speaker of a language. 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.

References

D. Abercrombie (1967), *Elements of General Phonetics* (Edinburgh Univ. Press, Edinburgh).

C. Adams (1979), *English Speech Rhythm and the Foreign Learner* (Mouton, The Hague).

J. Bernstein, M. Cohen, H. Murveit, D. Ritschev and M. Weintraub (1990), "Automatic evaluation and training in English pronunciation", *Proc. Internat. Conf. Spoken Language Process.*, Kobe, pp. 1185–1188.

M. Chapallaz (1964), "Notes on the intonation of questions in Italian", in *In Honour of Daniel Jones*, ed. by D. Abercrombie (Longman, London), pp. 306–312.

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

D. Crystal (1969), *Prosodic Systems and Intonation in English* (Cambridge Univ. Press, Cambridge).

D. Crystal (1975), *The English Tone of Voice* (Arnold, London).

A. Cutler and D.R. Ladd, eds. (1983), *Prosody: Models and Measurements* (Springer, Berlin).

R. Dauer (1983), "Stress-timing and syllable-timing reanalyzed", *J. Phonetics*, Vol. 11, pp. 51–62.

M.G. Di Benedetto (1989a), "Vowel representation: Some observations on temporal and spectral properties of the first formant frequency", *J. Acoust. Soc. Amer.*, Vol. 86, pp. 55–66.

M.G. Di Benedetto (1989b), "Frequency and time variations of the first formant: Properties relevant to the perception of vowel height", *J. Acoust. Soc. Amer.*, Vol. 86, pp. 66–77.

M.G. Di Benedetto and G. Flammia (1990), "Vowel distinction along auditory dimensions: A comparison between a statistical and a neural classifier", *Proc. Internat. Conf. Speech Technologies*, Rome, Italy, pp. 248–255.

G. Faure (1973), "La description phonologique des systèmes prosodiques", in *Interrogation et Intonation en Français Standard et en Français Canadien*, ed. by A.W. Grundström and P.R. Léon (Didier, Montréal), pp. 1–16.

A.W. Grundström (1973), "L'intonation des questions en français standard", in *Interrogation et Intonation en Français Standard et en Français Canadien*, ed. by A.W. Grundström and P.R. Léon (Didier, Montréal), pp. 19–51.

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

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

M.M. Kenning (1979), "Intonation systems in French", *J. Internat. Phonetics Assoc.*, Vol. 9, pp. 15–30.

M.M. Kenning (1983), "The tones of English and French", *J. Internat. Phonetics Assoc.*, Vol. 13, pp. 32–48.

P. Leach (1988), "French intonation: Tone or tune?", *J. Internat. Phonetics Assoc.*, Vol. 18, pp. 125–139.

I. Lehiste (1970), *Suprasegmentals* (MIT Press, London).

P.R. Léon and P. Martin (1980), "Des accents", in *The Melody of Language*, ed. by L.R. Waugh and C.H. Van Schooneveld (Univ. Park Press, Baltimore).

M. Liberman (1975), The intonational system of English, Massachusetts Institute of Technology PhD Dissertation (distributed by Indiana Linguistics Club).

H.S. Madsen (1983) *Techniques in Testing* (Oxford Univ. Press, Oxford).

P. Martin (1975), "Analyse phonologique de la phrase française", *Linguistics*, Vol. 146, pp. 35–68.

P. Martin (1982), "Phonetic realisations of prosodic contours in French", *Speech Communication*, Vol. 1, Nos. 3–4, pp. 283–294.

M. Miller (1984), "On the perception of rhythm", *J. Phonetics*, Vol. 12, pp. 75–83.

F.D. Minifie (1973), "Speech acoustics", in *Normal Aspects of Speech, Hearing and Language* ed. by F.D. Minifie, T.J. Hixon and F. Williams (Prentice Hall, Englewood Cliffs, NJ), pp. 235–284.

Z. Muljacic (1972), *Fonologia della Lingua Italiana* (Bologna).

J.D. O'Connor and G.F. Arnold (1961), *Intonation of Colloquial English* (Longman, London).

J. Pierrehumbert (1979), "Intonation synthesis based on metrical grids", in *Speech Communication Papers*, ed. by J. Wolf and D. Klatt (Acoust. Soc. Amer.), pp. 523–526.

J.B. Pierrehumbert (1980), The phonology and phonetics of English intonation, Massachusetts Institute of Technology PhD Dissertation, Cambridge, MA.

K.L. Pike (1945), *The Intonation of American English* (Univ. of Michigan Press, Ann Arbor).

W.M. Rivers (1975), *A Practical Guide to the Teaching of French* (Oxford Univ. Press, New York).

P. Roach (1982), "On the distinction between 'stress-timed' and 'syllable-timed' languages", in *Linguistic Controversies*, ed. by D. Crystal (Arnold, London), pp. 73–79.

A. Smith (1976), "The timing of French, with reflections on syllable timing", *Edinburgh University Department of Linguistics Work in Progress*, Vol. 9, pp. 97–108.

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

G.L. Trager and H.L. Smith (1951), *An Outline of English Word Structure* (Battenburg, Norman, OK).

B. Tranel (1987), *The Sounds of French: An Introduction* (Cambridge Univ. Press, Cambridge).

B. Wenk and F. Wioland (1982), "Is French really syllable-timed?", *J. Phonetics*, Vol. 10, pp. 193–216.