Int J Speech Technol (2009) 12: 161–184
DOI 10.1007/s10772-010-9065-1

Prosodic tools for language learning

Rodolfo Delmonte

Received: 7 November 2009 / Accepted: 20 January 2010 / Published online: 3 March 2010

Abstract In this paper we will be concerned with the role played by prosody in language learning and by the speech technology already available as commercial product or as prototype, capable to cope with the task of helping language learner in improving their knowledge of a second language from the prosodic point of view. The paper has been divided into two separate sections: Section One, dealing with Rhythm and all related topics; Section Two dealing with Intonation. In the Introduction we will argue that the use of ASR (Automatic Speech Recognition) as Teaching Aid should be under-utilized and should be targeted to narrowly focussed spoken exercises, disallowing open-ended dialogues, in order to ensure consistency of evaluation. Eventually, we will support the conjoined use of ASR technology and prosodic tools to produce GOP useable for linguistically consistent and adequate feedback to the student. This will be illustrated by presenting State of the Art for both sections, with systems well documented in the scientific literature of the respective field.

In order to discuss the scientific foundations of prosodic analysis we will present data related to English and Italian and make comparisons to clarify the issues at hand. In this context, we will also present the Prosodic Module of a courseware for computer-assisted foreign language learning called SLIM—an acronym for Multimedia Interactive Linguistic Software, developed at the University of Venice (Delmonte et al. in Convegno GFS-AIA, pp. 47–58, 1996a; Ed-Media 96, AACE, pp. 326–333, 1996b). *The Prosodic Module* has been created in order to deal with the problem of improving a student's performance both in the perception and production of prosodic aspects of spoken language activities. It is composed of two different sets of Learning Activities, the first one dealing with phonetic and prosodic problems at word level and at syllable level; the second one dealing with prosodic aspects at phonological phrase and utterance suprasegmental level. The main goal of Prosodic Activities is to ensure consistent and pedagogically sound feedback to the student intending to improve his/her pronunciation in a foreign language.

Keywords Computer assisted language learning · Speech technology for language learning · Prosody · ASR · Phonetics · Phonology · Acoustic phonetics

1 Introduction

The teaching of the pronunciation of any foreign language must encompass both segmental and suprasegmental aspects of speech. In computational terms, the two levels of language learning activities can be decomposed at least into phonemic aspects, which include the correct pronunciation of single phonemes and the co-articulation of phonemes into higher phonological units; as well as prosodic aspects which include

- the correct position of stress at word level;
- the alternation of stress and unstressed syllables in terms of compensation and vowel reduction;
- the correct position of sentence accent;
- the generation of the adequate rhythm from the interleaving of stress, accent, and phonological rules;
- the generation of adequate intonational pattern for each utterance related to communicative functions.

R. Delmonte (✉)
Linguistic Computational Laboratory (LCL), Università
Ca' Foscari—Ca' Bembo, Dorsoduro, 1075-30123 Venezia, Italy
e-mail: delmont@unive.it
url: project.cgm.unive.it

As appears from above, for a student to communicate intelligibly and as close as possible to native-speaker's pronunciation, prosody is very important (Bagshaw 1994). We also assume that an incorrect prosody may hamper communication from taking place and this may be regarded a strong motivation for having the teaching of Prosody as an integral part of any language course.

Learners of a second language create an interlanguage using both physical and psychological elements of their native language (Delmonte 1988b). As a result, they will transfer both the phonological rules (the grammar) and the acoustic cues for stress and rhythm identification from their native language to the second language. It is a fact that second language learners not only have problems in production but also in perception of prosodic features (Delmonte 1988a).

From our point of view, what is much more important to stress is the achievement of a successful communication as the main objective of a second language learner rather than the overcoming of what has been termed "foreign accent", which can be deemed as a secondary goal. In any case, the two goals are certainly not coincident even though they may be overlapping in some cases. We will discuss these matter in the following sections.

All prosodic questions related to "rhythm" will be discussed in the first section of this chapter. In Roach (2000) the author argues in favour of prosodic aids in particular because a strong placement of word stress may impair understanding from the listener's point of view of the word being pronounced. He also argues in favour of acquiring correct timing of phonological units to overcome the impression of "foreign accent" which may ensue from an incorrect distribution of stressed vs. unstressed stretches of linguistic units such as syllables or metric feet. Timing is not to be confused with speaking rate which need not be increased forcefully to give the impression of a good fluency: trying to increase speaking rate may result in lower intelligibility.

The question of "foreign accent" is also discussed at length in Jilka (2009). This work is particularly relevant as far as intonational features of a learner of a second language which we will address in the second section of this chapter. Correcting the Intonational Foreign Accent (hence IFA) is an important component of a Prosodic Module for self-learning activities, as categorical aspects of the intonation of the two languages in contact, L1 and L2 are far apart and thus neatly distinguishable. Choice of the two languages in contact is determined mainly by the fact that the distance in prosodic terms between English and Italian is maximal, according to Ramus and Mehler (1999), Ramus et al. (1999).

In all systems based on HMMs (Kawai and Hirose 1997; Ronen et al. 1997), student's speech is segmented and then matched against native acoustic models. The comparison is done using HMM loglikelihoods, phone durations, HMM phone posterior probabilities, and a set of scores is thus obtained. They should represent the degree of match between non-native speech and native models. In the papers quoted above, there are typically two databases, one for native and another for nonnative speech which are needed to model the behaviour of HMMs. As regards HMMs, in Kim et al. (1997) the authors discuss the procedure followed to generate them: as expected, they are trained on the native speakers database where dynamic time warping has applied in order to eliminate the dependency of scoring for each phone model on actual segment duration. Duration is then recovered for each phone from each frame measurements and normalized in order to compensate for rate of speech. Phonetic time alignment is then automatically generated for the student's speech.

HMModels are inherently inadequate to cope with prosodic learning activities since statistical methods can only produce distorted results in a teaching environment. First of all, they need to produce a set of context-independent models for all phone classes and this fact goes against the linguistically sound principle that says that learning a new phonological system can only be done in a context-dependent fashion. Each new sound must be learnt in its context, at word level, and words should be pronounced with the adequate prosody, where duration plays an important role. One way to cope with this problem would be that of keeping the amount of prosody to be produced under control: in other words to organize tasks which are prosodically "poor" in order to safeguard students from the teaching of bad or wrong linguistic habits.

Then there is the well-know problem of the quantity of training data to be used to account for both inter-speaker and intra-speaker variability. In addition, since a double database should be used, one for native and one for non-native speakers, the question is what variety of native and non-native is being chosen, seen that standard pronunciation is an abstract notion. As far as prosody is concerned, we also know that there is a lot of variability both at intraspeaker and interspeaker level: this does not hinder efficient and smooth communication from taking place, but it may cause problems in case of a student learning a new language.

Other problems are related to well-known unsuitability of HMM to encode duration seen that this parameter cannot be treated as an independent variable (but see the discussion in the sections below). Other non-independent aspects regard transitions onto and from a given phonetic segment together with carryover effects due to the presence of previous syllabic nasal or similar sonorant units. In addition, the maximum likelihood estimate and smoothing methods introduce errors in each HMM which may be overlooked in the implementation of ASR systems for dictation purposes; but not in the assessment of Goodness of Pronunciation for a given student with a given phoneme. Generally speaking, HMMs will only produce decontextualized standard models to follow for the student, which are intrinsically unsuited to be used for assessment purposes in a teaching application.

In pronunciation scoring, technology is used to determine how well the expected word/utterance was said. It is simple to return a score; the trick is to return a score that "means" something (Price 1998: 105). Many ASR systems have a score as a by-product. However, this score is tuned for use by native speakers, and does not tend to work well for language learners. Therefore, unacceptable or unintelligible utterances may receive good scores (false positives), and intelligible utterances may receive poor scores (false negatives). Students want detailed scores on their pronunciation, not overall scores or sentence scores. The challenge is that reliable scores at the individual sound level are difficult for experts as well as for our algorithms. A score alone is not likely to be as useful as diagnostics on the source of error and possible remedies.

1.1 SLIM and ASR

SLIM is an interactive multimedia system for self-learning of foreign languages and is currently addressed to Italian speakers. It has been developed partly under HyperCard™, and partly under MacroMedia Director™. However at present, the Prosodic Module interacts in real time only with HyperCard™ (Delmonte 2000).

SLIM makes use of Speech Recognition in a number of tasks which exploit it adequately from the linguistic point of view. We do not agree with the use of speech recognition as adequate assessment tool for the overall linguistic competence of a student. In particular, we do not find it suited for use in language practice with open-ended dialogues given the lack of confidence in the ability to discriminate and recognize Out-Of-System utterances (Meador et al. 1998). We use ASR only in a very controlled linguistic context in which the student has one of the following tasks:

- repeat a given word or utterance presented on the screen and which the student may listen to previously—the result may either be a state of recognition or a state of non-recognition. The Supervisor will take care of each situation and then allow the student to repeat the word/utterance a number of times;
- repeat in a sequence "minimal pairs" presented on the screen and which the student may listen to previously—the student has a fixed time interval to fulfil the task, and a certain number of total possible repetitions (typically twenty)—at the end, feedback will be number of correct repetitions;
- speak aloud one utterance from a choice among one to three utterances appearing on the screen as a reply to a question posed by a native speaker's voice or by a character in a video-clip. This exercise is called Questions and Answers and is usually referred to a False Beginner-Intermediate level of proficiency of the language. The student must be able to understand the question and to choose the appropriate answer on the basis of grammatical/semantic/pragmatic information available. The outcome may be either a right or a wrong answer, and ASR will in both cases issue the appropriate feedback to the student;
- do role-play, i.e. intervene in a dialogue of a video-clip by producing the correct utterance when a red light blinks on the screen, in accordance with a given communicative function the student is currently practising. This is a more complex task which is only allowed to be accessed by advanced students: the system has a number of alternative utterances connected with each communicative function the student has to learn. The interaction with the system may be both in real time or in slow-down motion: in the second case the student will have a longer time to synchronize his/her spoken utterance with the video-clip.

One might question the artificiality of the learning context by reminding the well-known fact that a language can only be learnt in a communicative situation (Price 1998). However we feel that the primary goal of speech technology is to help the student develop good linguistic habits in L2, rather than engaging the student in the use of "knowledge of the world/context" creatively in a second language.

Thus we assume that speech technology should focus on teaching systems which incorporate tools for prosodic analysis focussing on the most significant acoustic correlates of speech in order to help the student imitate as close as possible the master performance, contextualized in some communicative situation.

Some researcher have tried to cope with the problem of identifying errors in phones and prosody within the same ASR technology (Eskénazi 1999). The speech recognizer in a "forced alignment mode" can calculate the scores for the words and the phones in the utterance. In forced alignment, the system matches the text of the incoming signal to the signal, using information about the signal/linguistic content that has already been stored in memory. Then after comparing the speaker's recognition scores to the mean scores for native speakers for the same sentence pronounced in the same speaking style, errors can be identified and located (Bernstein and Franco 1995). On the other hand, for prosody errors, duration can be obtained from the output of most recognizers. In rare cases, fundamental frequency may be obtained as well. In other words, when the recognizer returns the scores for phones, it can also return scores for their duration. On the other hand, intensity of the speech signal is measured before it is sent to the recognizer, just after it has been preprocessed. It is important that measures be expressed in relative terms—such as duration of one syllable compared to the next—since intensity, speaking rate, and pitch vary greatly from one individual to another.

The FLUENCY system—which will be illustrated further on in the chapter—uses the SPHINX II recognizer to

detect the student's deviations in duration compared to that of native speakers. The system begins by prompting the student to repeat a sentence. The speech signal and the expected text are then fed to the recognizer in forced alignment mode. The recognizer outputs the durations of the vowels in the utterance and compares them to the durations for native speakers. If they are found to be far from the native values, the system notifies the user that the segment was either too long or too short.

Bagshaw et al. (1993a) compared the student's contours to those of native speakers in order to assess the quality of pitch detection. Rooney et al. (1992) applied this to the SPELL foreign language teaching system and attached the output to visual displays and auditory feedback. One of the basic ideas in their work was that the suprasegmental aspects of speech can be taught only if they are linked to syllabic information. Pitch information includes pitch increases and decreases and pitch anchor points (i.e., centers of stressed vowels). Rhythm information shows segmental duration and acoustic features of vowel quality, predicting strong vs. weak vowels. They also provided alternate pronunciations, including predictable cross-linguistic errors.

As we will argue extensively in this section, we assume that segmental information is in itself insufficient to characterize non-native speech prosody and to evaluate it. In this respect, "forced alignment mode" for an ASR working at a segmental/word level still lacks hierarchical syllabic information as well as general information on allowable deviations from mother-tongue intonation models which alone can allow the system to detect prosodic errors with the degree of granularity required by the application.

The paper is then organized as follows:

- Section 1—The Introduction
- Section 2—Prosodic tools for self-learning activities in the domain of rhythm
  - General problems related to rhythm
  - Building and exploiting a prosodic syllable database: our approach
  - Self-learning activities in the prosodic module: word stress and timing
- Section 3—Prosodic tools for self-learning activities in the domain of intonation
  - General problems related to intonation
  - Intonation practice and visualization: our approach
  - Self-learning activities in the prosodic module: utterance level exercises
- Section 4—Conclusions

Table 1 Mean consonant durations

	EN	SP	IT	Orthog.	Tokens
θ	99			thin	101
ð	48	41		then	970
h	62			hat	593
ʃ	***104***		***84***	ship	850
ʃ:	168		147		16
ʒ	***49***	47	***35***	leisure	456
z	***86***	77	***67***	zeal	1615
s	119	79	98	seal	5149
s:	***164***	91	***149***	lesso	172
v	***64***		***55***	vat	1251
v:	***92***		***98***		11
f	***104***	81	***104***	fat	1330
f:	***136***		***126***	goffo	33
x		86		Bach	695
#	269	303	212	pause	3494
?	57	65	36	oh	3325
p:			129	cappa	30
p	101	70	79	pit	2005
b:			87	gobba	42
b	***72***	43	***66***	bit	925
d:	98		79		114
d	***61***	43	***57***	dig	3166
t:	***115***		***108***	cotto	346
t	94	64	72	tip	5683
g	76	35	55	god	554
k:	154		130	pacco	60
k	***103***	71	***85***	cat	2912

2 Prosodic tools for self-learning activities in the domain of rhythm

2.1 General problems related to rhythm

In prosodic terms, Italian and English are placed at the two opposite ends of a continuum where languages of the world are placed (Ramus and Mehler 1999; Ramus et al. 1999).

This is dependent on the two overall phonological systems, which in turn are bound by the vocabulary of the two languages. The Phonological system will typically determine the sound inventory available to speakers of a given language; the vocabulary will decide the words to be spoken. The Phonological system and the vocabulary in conjunction will then determine the phonotactics and all suprasegmental structures and features. We will at first, look at data computed by Grover et al. (1998) for the two languages at hand English (EN) and Italian (IT) and relate them to Spanish (SP) in order to highlight similarities and differences among language families of main acoustic correlates of prosodically relevant features. The Tables 1, 2 report mean duration

Table 2 Mean sonorant & vowel durations

	EN	SP	IT	Orthog.	Tokens
j	74	52	46	yes	1448
w	62	39	37	why	911
m	***69***	71	***64***	man	2586
m:	***94***	104	***96***	mamma	50
n	***72***	65	***58***	not	7095
ŋ	80			ring	428
n:	135	107	101	canna	138
ñ		66		ragno	53
λ:			83	aglio	57
l	***58***	53	***54***	lamp	4222
l:	***85***	92	***75***	collo	195
r	***45***	42	***43***	rovo	5632
r:	***83***		***77***	corro	48
R		69		carro	947
eɪ	155			fail	912
oɪ	173			foil	79
əʊ	141		126	goal	1216
aɪ	178			file	404
a	170			allow	255
ʊ					
i	122	64	70	seal	4976
I	63			bit	1943
ə	51			allow	4889
ʊ	58			full	516
ɔ	135			fall	2119
ɛ	***106***		***111***	leisure	2510
ʊ	77			put	483
ɒ	124	78	94	got	5424
e	***82***	61	***73***	este	5777
æ	102			cat	451
o	109	69	78	door	4012
u	113	57	71	fool	1502

data for all sound classes of three languages relevant for our comparative analysis.

We turned figures into bold and italics or we underlined Spanish data, in order to highlight mean data for those sounds where the difference in absolute values is below or equal to 20 msec, a threshold which has been regarded perceptually "noticeable" (Pisoni 1977).

We shall comment the EN/IT comparison first. As can be seen, as far as their common sound inventory is concerned, over a total of 55 sounds, only 21 (amounting to 38%) show similar enough data. Of the 34 remaining sounds: half, 17 (amounting to 30.9%), are not shared between the two languages: 14 only belong to English and not to Italian, 3 only belong to Italian and not to English; the other half, 17, are part of a similar inventory but show a different enough prosodic behaviour.

On the contrary if we look at data from Spanish and compare them to Italian, we see that as far as the inventory is concerned, we are now left with 45 sounds, almost half of which, 22 (48.9%) show similar prosodic behaviour of the remaining 23 sounds, 17 (amounting to 37.7%), are not shared between the two languages: 3 only belong to Spanish and not to Italian, and 14 only belong to Italian and not to Spanish; 6 (amounting to 13.4%) are part of a similar inventory but show a different enough prosodic behaviour.

If we look at sound classes we see that similarities are concentrated in Sonorant class consonants where most sounds are shared and half have a similar prosodic behaviour. As far as dissimilarities are concerned, the Vowel sounds have the highest number of nonshared sounds between English and Italian/Spanish: just the opposite would apply to the language pair Italian and Spanish. Thus the number of contrastively different sounds to be learned totally anew for an Italian L2 student of English when compared to the same L2 learner of Spanish would fare 14 to 3; if we look at differences which requires the student to perform a more finely tuned learning process, prosodically based, this would disfavour again English as L2, with a 31% of sounds belonging to the same inventory but requiring a different tuning in English, vs. a 13.4% of sounds in Spanish. However this tuning in our opinion is cannot be accomplished at a segmental level but only at a prosodic suprasegmental level: and the prosodic unit to be addressed would be the syllable. This is the unit which allows the speaker to realize durational differences which will then carry over to rhythmic variations.

As far as syllables are concerned, we should also note that their most important structural component, the nucleus, is a variable entity in the two language families: syllable nuclei can be composed of just vowels or of vowels and sonorants. Vowel and sonorant sounds being similar would account for the greatest impression of two languages sounding the same or very close: from a simplistic segmental point of view, English and Italian/Spanish would seem to possess similar prosodic behaviour as far as sonorants are concerned. On the contrary, we should note the fact that English would syllabify a sonorant as syllable nucleus—as would German—but this would be totally unknown to a Romance Italian/Spanish speaker. Contrastive studies have clearly pointed out the relevance of phonetic and prosodic exercises both for comprehension and perception. In general prosodic terms, whereas the prosodic structure of Italian is usually regarded as belonging to the syllable-timed type of languages, that of English is assumed to belong to stress-timed type of languages (Bertinetto 1980; Lehiste 1977). This implies a remarkable gap especially at the prosodic level between the two language types. Hence

the need to create computer aided pronunciation tools that can provide appropriate feedback to the student and stimulate pronunciation practice.

Reduced vowels typically affect duration of the whole syllable, so duration measurements are usually sufficient to detect this fact in the acoustic segmentation. In stress-timed languages the duration of interstress intervals tends to become isochronous, thus causing unstressed portions of speech to undergo a number of phonological modifications detectable at syllable level like phone assimilation, deletion, palatalization, flapping, glottal stops, and in particular vowel reduction. These phenomena do not occur in syllable-timed languages—but see below—which tend to preserve the original phonetic features of interstress intervals (Bertinetto 1980). However a number of researcher have pointed out that isochrony is much more a matter of perception than of production (see in particular, Lehiste 1977). Differences between the two prosodic models of production are discussed at length in a following section.

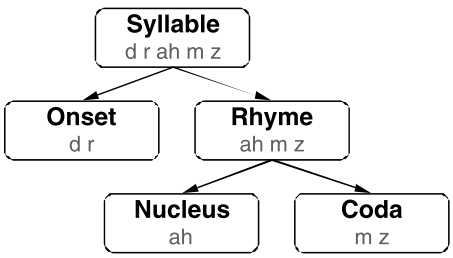
2.1.1 Building and exploiting a prosodic syllable database: our approach

Prosodic data suffer from a well-known problem: sparsity (Delmonte 1999). In order to reach a better understanding of this problem however, we would like to comment on data in the literature (Delmonte 1999; van Son and van Santen 1997; Umeda 1977; van Santen et al. 1997) basically related to English, apart from (van Santen et al. 1997), and compare them with data available on Italian. We support the position also endorsed by Klatt and theoretically supported by Campbell and Isard in a number of papers (Campbell and Isard 1991; Campbell 1993), who consider the syllable the most appropriate linguistic unit to refer to in order to model segmental level phonetic and prosodic variability.

The reason why the coverage of data collected for training corpus is disappointing is not simply a problem of quantities, which can be solved by more training data. The basic problem seems to be due to two ineludible prosodic factors:

- the need to encode structural information in the syllable, which otherwise would belong to higher prosodic units such as the Metric Foot, The Clitic Group, The Phonological Group (which will be discussed in more detail below);
- the prosodic peculiarity of the English language at syllable level.

I am here referring to the great variety of syllabic nuclei available in English due to the high number of vowels and diphthongs and also to the use of syllabic consonants like nasals or liquids as syllable nuclei. the presence of a too large feature space, or too great number of variables to be considered. When compared with a language like Chinese, we see two languages at the opposite sides: on the one side a language like Chinese where syllables have a very limited distribution within the word and a corresponding limitation in the type of co-occurring vowel; on the other side very high freedom in the distribution of syllables within the word as our data will show. As to stressed vs unstressed syllables the variability is very limited in Chinese due to the number of stressable vowels, and also due to the fact that most words in Chinese are monosyllabic. In addition, syllable structure is highly simplified by the fact that no consonant clusters are allowed. In fact (van Santen et al. 1997: 321) reports the number of factors and parameters used to compute the multilingual prosodic model for Chinese, French and German we see that Chinese has less than one third the number of classes and less than half the number of parameters than the other two languages. English, which is not listed, is presented in van Son and van Santen (1997) with the highest number of factors, 40. Sparsity in prosodic data is then ultimately linked to the prosodic structure of the language, which in turn is partly a result of the interaction between the phonological and the lexical system of the language.



If we look at syllable distribution in Italian (Delmonte 1991) we see that syllables are for the great majority present in more than 2 positions in the word; syllables present in only one position constitute 15.6% (see Delmonte 1999: 321). Thus syllables are very predictable: 66 syllable types cover 87% of the total corpus analyzed. In order to build a database that contains syllable-level information along with word-level and phoneme-level information we used the WSJCAM0—the Cambridge version of the continuous speech recognition corpus produced by the Wall Street Journal, distributed by the Linguistic Data Consortium (LDC). We worked on a subset of 4165 sentences, with 70,694 words that constitute half of the total number of words in the corpus amounting to 133,080. We ended up with 113,282 syllables and 287,734 phones. The final typology is made up of 44 phones, 4393 syllable types and 11,712 word types. As far as syllables are concerned, we considered only 3409 types. Read speech databases usually contain hand-annotated time-aligned phoneme-level and word-level transcriptions of each utterance. Our attempt was to use the information available in order to build a syllable-level transcription of each utterance. We found that using only

Table 3 Diagram of the syllable-level general algorithm

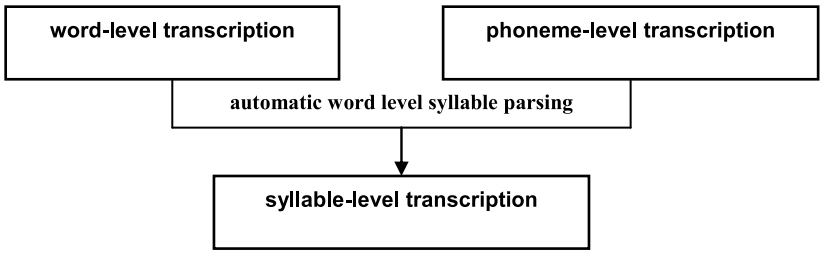


Table 4 A LALR grammar based on syllable structure

SYLLABLE	→	[ONSET] NUCLEUS [CODA]
ONSET	→	FRICATIVE STOP {LIQUID | GLIDE}
		FRICATIVE {STOP | GLIDE | LIQUID | NASAL | FRICATIVE}
		STOP {LIQUID | GLIDE }
		LIQUID GLIDE
		NASAL GLIDE
		{FRICATIVE | STOP | GLIDE | LIQUID | NASAL}
NUCLEUS	→	{SHORT VOWEL | VOWEL | DIPHTHONG} SCHWA
		{REDUCED VOWEL | SHORT VOWEL | VOWEL | DIPHTHONG}
CODA	→	STOP FRICATIVE {STOP | FRICATIVE} FRICATIVE
		LIQUID {{FRICATIVE FRICATIVE} | {STOP STOP}} FRICATIVE
		NASAL STOP {{FRICATIVE STOP} | {STOP FRICATIVE}}
		FRICATIVE {{STOP {STOP | FRICATIVE}} | {FRICATIVE FRICATIVE}}
		STOP {{STOP FRICATIVE} | {FRICATIVE {FRICATIVE | STOP}}}
		LIQUID {{STOP FRICATIVE} | {NASAL {STOP | FRICATIVE} }}
CODA	→	LIQUID {FRICATIVE {STOP | FRICATIVE}}
		NASAL {STOP | FRICATIVE} {STOP | FRICATIVE}
		{FRICATIVE | STOP} {STOP | LIQUID | NASAL | FRICATIVE}
		LIQUID {STOP | NASAL | FRICATIVE}
		NASAL {STOP | FRICATIVE}
		{FRICATIVE | STOP | LIQUID | NASAL}

phoneme-level information was difficult because the continuous syllable parsing is not as simple at utterance-level as it is at word-level. So both phoneme-level and word-level time-aligned transcriptions have been used.

2.1.2 Automatic syllable parsing

The algorithm for word-level syllable parsing that we used is based on the structure of English syllables and on a certain number of phonological rules (see below). The syllable is made of a nucleus, which is a vowel or a vowel-like consonant—usually a sonorant—that can be optionally prefixed and suffixed by a number of consonants, termed the onset and coda respectively.

A LALR(1) grammar has been written based on this syllable structure and on the phoneme-level structure of the onset, nucleus and coda.

We found this grammar useful for dividing the syllable into onset, nucleus and coda, but modifying the grammar to parse sequences of syllables by adding the rule SYLLABLELIST → [SYLLABLELIST] SYLLABLE resulted in an ambiguous grammar. Trying to use the LALR(1) finite state automata there are two types of conflicts to be resolved: *shift-reduce* and *reduce-reduce* conflicts corresponding to vowel-to-vowel and consonant-to-consonant transitions. Some phonological rules (Bannert 1987; Kahn 1976) have to be applied and more look-ahead has to be done in order to resolve the conflicts during the parsing process.

Any consonant cluster that can start or end an English word must also be a permissible onset or coda of an English syllable. The *maximum onset rule* states that the syllable boundaries must be placed so as to maximize the number of consonants at the beginning of each syllable. The rules in our grammar may be considered too general. Stronger constraints on possible English syllable-initial and syllable-final consonant clusters can be used.

Based on these considerations a modified finite state automata has been built in order to parse words as sequences of syllables. In the case of consonant-to-consonant transi-

tion conflicts we found that in order to take the right decision it is enough to look-ahead until a vowel is encountered and to check the table of all permissible syllable onsets and codas (Kahn 1976). We also used the fact that some permissible onsets and codas cannot occur in the medial position of multi-syllabic words. For example the four consonant codas can occur infrequently in word-final position such as *tempts* (t ɛ m p t s), sixths (s ɪ k s θ s), etc.

The algorithm has been tested first using the Carnegie Mellon University Pronouncing Dictionary that contains more than 100.000 entries. The syllabification generated by the algorithm was correct in 93% of the cases. The errors made by the algorithm were found to be caused mainly by foreign and by compound words. A large number of errors occur mainly at vowel-schwa transitions in the compound words such as *flyaway* (f l ay - e - w ey). Trying to relax the constraints in this case, so that a syllable boundary has to be inserted always between a vowel and a schwa, would result in an even larger number of wrong generated syllables. Other errors occur when the syllabification does not obey the *maximum onset rule* such as the word *meniscus* that has to be syllabified as (m e - n ɪ s - k e s) and not (m e - n ɪ - s k e s). In this case the so-called *weight-to-stress principle* (that states that a consonant before an unstressed syllable affiliates to the left syllable if this one is stressed in defiance to the *maximum onset rule*) has to be applied.

To limit the number of errors we organized a list of the most frequently used foreign and compound words already divided into syllables and asked the parser to search this list every time a new segmentation is tried at word level. Using the syllable-parsing algorithm the time-aligned syllable-level transcription of each utterance has been obtained.

2.1.3 The syllable database

The next step after the syllable-level transcription has been obtained was to put together all the information in an easy to use form. We found that the best choice for our purposes was to create a relational database. This allowed us to easily develop a set of software tools in order to provide context-dependent and context-independent statistical information about words, syllables and phonemes (such as frequency, average duration, standard deviation, etc.).

In this preliminary form the syllable database does not contain any syllable-level stress information. In order to assign stress information to the syllables we tried to develop an automatic procedure. The syllables have been divided into three categories: *reduced*, *light* and *heavy* depending on the structure of the syllable and the composition of its nucleus. The class of *reduced* syllables contains those syllables whose nucleus is a reduced vowel or a semivowel. The syllables that contain a short vowel and no coda belong to the class of *light* syllables. The other syllables that contain a long vowel or a diphthong, or a short vowel followed by at least one consonant are *heavy* syllables.

So, we started by classifying the words into functional and non-functional and labelled all the syllables of the functional words as unstressed. Then all the *reduced* syllables have been also labelled as unstressed. For all the words with only one syllable not already classified, the non-classified syllable has been labelled as stressed.

For the remaining words with non-classified syllables we tried to use a pronunciation dictionary that contains stress information. The problem was that the dictionary transcription doesn't always match the actual pronunciation of the word. If the pronunciation of the word was found in the dictionary then stress has been assigned to syllables using the information stored into the dictionary. To those words whose pronunciation was different than the one found in the dictionary a weight pattern has been attached. For example, if the first syllable of the word is *reduced,* the second syllable is *heavy* followed by a *light* syllable, in this way the weight pattern of the word is *reduced-heavy-light*. Then for all those words that have the same number of syllables and the same weight pattern as the corresponding dictionary transcription stress has been assigned to syllables according to the stress pattern of the dictionary transcription. For the rest of the words it was labelled as stressed the penultimate syllable if it is *heavy* or else the antepenultimate syllable if it is *light* or *heavy*, or else the rightmost *heavy* or *light* syllable. All the other syllables were labelled as unstressed.

To evaluate the performance of the algorithm we checked all the words with more than three syllables from our database (2896 out of 70694). We found that the algorithm performed well in more than 97% of the cases. The errors found had to be corrected by hand.

As we already said, the main reason for creating the syllable database was to use it in the Prosodic Module of SLIM (an acronym for Multimedia Interactive Linguistic Software), developed at the University of Venice. This module—described in detail in Delmonte et al. (1997), Delmonte (1998) but also below—is composed of learning activities dealing with phonetic and prosodic problems at word segmental level and at utterance suprasegmental level and has the goal to improve the student's performance both in the perception and production of prosodic aspects of spoken language activities.

We found that, in some cases, using the information at syllable level instead that at phoneme level dramatically improves the performance of the automatic alignment of the speech signal algorithm that we use in the prosodic module.

Also, while using the syllable database, we were able to build more reliable prosodic and duration models for the syllables and this allowed us to give a better feedback to the student, to tell him where or what the mistake is in his pronunciation.

Table 5 Vowels and glides mean durational values

	Mean	Minimum	Maximum	St. dev.	Occur. No.
Long vowels	99,02	48.00	406.86	43.24	28.323
Short vowels	60.58	32.00	276.00	25.22	39.182
Reduced vowel	37.56	16.00	336.00	24.57	24.923
Diphthongs	123.44	60.75	386.00	47.17	17.452
Glides	73.98	32.00	400.00	37.44	29.613

Table 6 Syllable types with higher frequency values

Syllable types	Mean	Minimum	Maximum	St. dev.	Occur. No.
ðə	91.30	48.00	400.00	29.31	4067
ə	38.29	16.00	240.00	20.48	2986
tə	110.94	48.00	416.00	42.59	2730
ɪn	110.25	64.00	416.00	38.29	1469
əv	101.50	48.00	400.00	36.65	1394
ən	97.64	48.00	288.00	29.55	1289
pə	113.34	64.00	288.00	25.76	1166
tɪ	127.52	64.00	352.00	44.39	793
ʃn	250.37	96.00	560.00	61.73	747
ðæt	171.04	112.00	432.00	50.62	726
ɪz	139.93	80.00	512.00	50.12	668
tʊ	205.35	80.00	576.00	75.35	651
fɔ	151.19	96.00	416.00	46.86	619
tɪd	209.53	96.00	672.00	70.69	556
kən	178.21	96.00	384.00	35.23	536

The problem that we encountered was that it was not possible to use syllable level information in all cases, due to the fact that we found syllables missing from the syllable database and for some syllables there were too few occurrences in the database, (the well-known sparseness problem, see works by van Son and van Santen (1997) and van Santen (1997)). In these cases context-dependent phoneme level information was used.

As to phones, we came up with the same conclusions reached by Klatt's experiment. According to the following theoretical subdivision of vowels we individuated four levels of duration for vowels:

/ə/ reduced; arpabet correspondence | ax |

/ɪ ʊ ɒ e/ short vowels; arpabet correspondence | ih eh ah uh |

/æ ʌ i: u: ɔ: a: ɜ:/ long vowels; arpabet correspondence | ae aa ao oh er iy uw |

diphthongs; arpabet correspondence | ia ea ey ow aw ay oy ua |

glides; arpabet correspondence | w y r l |

We considered syllables according to two factors: stressed vs unstressed, but also to position in word. We classified all syllables according to position: Word Initial, Medial Final. From the 3409 types the classification into stressed vs. unstressed gave us the following distribution:

A. Stressed syllable types 2581 with 33,351 occurrencies;
B. Unstressed syllable types 1108 with 54,128 occurrencies.

Thus, we can conclude that stressed syllable occurrencies are scattered amongst a high number of types, more than the double than the number of unstressed syllable types. We collected all syllable types with frequency of occurrence higher than 100 and we turned out with 110 types for unstressed vs. 55 stressed. In particular, then, no stressed syllable type overrides 500 occurrences: on the contrary, 15 unstressed syllables types have a frequency higher than 500, some of them are well over 1000 as shown by Table 6.

The intersection between the two sets produces however a very small number of types: only 280 syllable types may have both a stressed and an unstressed counterpart. If we eliminate from this intersection subset pairs containing hapax legomena, i.e. syllables with only 1 or 2 occurrencies, the final result drops down to 140 syllable types. As expected, distribution of stressed vs unstressed syllable types in English is strictly dependent on lexical structure of words: in particular, function words are destressed and usually constitute the great majority of cases. According then to word typology, which is strictly determined by the linguistic domain, there will be the supremacy of certain types of sylla-

Table 7 Absolute Values for 4 Corpora of read and telephone spoken Italian

Corpora/ prosodic elements	Corp. 1	Corp. 2	Corp. 3	Corp. 4	Total
Utterance	148	14	18	110	290
Words	2368	151	324	181	3024
Syllables	4504	377	354	505	5645
Phones	9956	806	786	1227	12757
Mean syll. length in number of phones	2.4	2.1	2.3	3.0	
Mean phone/words	4.7	5.3	4.3	6.8	
Mean syll/words	2.2	2.5	1.1	2.2	
Mean phone dur.	77 ms	66 ms	91 ms	68 ms	
Mean syllable dur.	132 ms	132 ms	197 ms	168 ms	
Total stressed syll.	1641	84	102	151	
Total unstress. syll.	2863	293	252	254	
Mean stress. syll.	244 ms	212 ms	252 ms	188 ms	
Mean unstress. syll.	134 ms	120 ms	179 ms	134 ms	

bles, with the obvious proviso of the unbalance in the proportion of unstressed vs. stressed syllables.

Coming now to syllable types distribution according to their position within the word, we ended with the following data:

A. Word Initial: 2700 syllables, divided up into 1968 higher frequency + 732 hapax legomena
C. Word Medial: 1607 syllables, divided up into 1023 higher frequency + 584 hapax legomena
D. Word Final: 1911 syllables, divided up into 1310 higher frequency + 601 hapax legomena.

Contrary to what happened with stress factor, where we saw an almost complementary distribution of syllables, in case of positional factor we find a lot of overlapping, which however is fairly randomly distributed. As to differences in duration due to position in the word our data are in accordance with (Delmonte 1999).

In order to build a comparative timing model for Italian speakers, we collected data in four different corpora which we present in a concise manner (Table 7).

One corpus, Corpus 2 was made up of utterances read aloud by an expert phoneticians with no repetitions, but which was meant to investigate the role of syntactic structure on timing (Delmonte et al. 1986; Delmonte 1987a). In this case we used utterances that contained a number of key words in different syntactic and phonological environment. Corpus 1 was repeated 44 times at different time intervals; Corpus 3 was repeated 3 times; Corpus 4 is made of spontaneous telephone conversations. Table 7 contains all values computed and also Mean Durational values in msec. for Phone and Syllable. Differences in speaking rate amount to 37% mean phone duration which applies also to syllable durations; moreover, approximately the same difference can be computed for stressed vs unstressed phone/syllable. These data are very significant and must be adequately computed within our model. As reported in Table 6 the four corpora are fairly balanced as to syllable structure, which is regarded a very important factor in Italian: not only consonant clusters in syllable onset, but also and foremost syllable codas are to regarded as a relevant parameter for timing. Generally speaking, open syllables are longer than closed syllables. Other important factors to be considered is whether a word is a function grammatical word or a content word. Destressing usually applies in those cases of function words which are treated as proclitic elements of a Phonological Word(PW). PWs in Italian are right headed (Delmonte and Dolci 1991). Prepausal lengthening always applies, but syllable lengthening also takes place at Phonological Phrase (PP) and Intonational Group (IG) boundaries. Other important factors are syllable compressions effects in long PP where compensation applies but only in co-occurrence with long utterances. For instance in the complex NP "il numero telefonico dell'MIT"/the phone number of MIT" in a comparatively long utterance (No. 5, Corpus 3), stressed and unstressed syllables are shorter than in other contexts. Elisions phenomena can only take place at word boundaries and are well documented in our corpora (see also Delmonte 1981).

2.2 Evaluation tools for timing and rhythm

As stated in the Introduction, assessment and evaluation are the main goal to be achieved by the use of speech technology, in order to give appropriate and consistent feedback to the student. Theoretically speaking, assessment requires the system to be able to decide at which point in a graded scale the student's proficiency is situated. Since students

usually develop some kind of interlingua between two opposite poles, non-native beginners and full native pronunciation, the use of two acoustic language models should be targeted to low levels of proficiency, where performance is heavily encumbered, conditioned by the attempts of the student to exploit L1 phonological system in learning L2. This strategy of minimal effort will bring as a result a number of typical errors witnessing to a partial overlapping between the two concurrent phonetic inventories: phonetic substitutes, for phonetic classes not attested in L2 will cause the student to produce words which only approximate the target sound sequence perhaps by manner but not by place of articulation as is the usual case with dental fricatives in English [ð, θ].

Present-day speech recognizers are sensitive exclusively to phonetic information concerning the words spoken—their contents in terms of single phones. Phonetically based systems are language-specific, not only because the set of phonemes is peculiar to the language but also because the specification of phonetic context means that only certain sequences of phonemes can be modeled. This presents a problem when trying to model defective pronunciations generated by non-native speakers. For example, it might be impossible to model the pronunciation [zæt]—typical of languages lacking dental fricatives—for the word *that* with a set of triphones designed only for normal English pronunciations.

Current large-vocabulary recognition systems use *sub-word* reference model units at the phoneme level. The acoustic form of many phonemes depends critically on their phonetic context, particularly the immediately preceding and following phonemes. Consequently, almost all practical sub-word systems use *triphone* units; that is, a phoneme whose neighbouring phoneme to the left and to the right is specified. Clearly, only in case some errors are detected and evaluated, the system may try to guess which level of interlingua the student belongs to.

Thus the hardest task ASR systems are faced with is segmentation. In Hiller et al. (1993) segmentation is obtained using a HMM technique where the labeling of the incoming speech is constrained by a segmental transition network which is similar to our lexical phonetic description in terms of phones with associated phonetic and phonological information. In their model however, a variety of alternative pronunciations are encoded, including errors predictable from the student's mother tongue. These prediction are obtained from a variety of different sources (see Hiller et al. 1993: 466). In our case, assessment of the student's performance is made by a comparative evaluation of the expected contrastive differences in the two prosodic models in contact, L1 and L2. Main factors to be controlled in the phonological model of a syllable-timed language may then be traced in the following linguistic elements:

– Prepausal lengthening at PP and IG boundaries;
– Syllable coda intraword effect;
– Stressed syllable type effect;
– PP syllabic length compensatory effect;
– PW function/content word compensatory effect;
– Stressed vs Unstressed syllable effect;
– Contextual consonant cluster effect;
– Interword sandhi rules effect.

According to van Son and van Santen (1997), Umeda (1977), van Santen et al. (1997) the effect of context in determining consonant duration can be schematized in the following conditions:

a. consonant identity;
b. relative position of the consonant in the word {initial, medial, final, prepausal}
c. preceding conditions of the consonant followed by a vowel {vowel, nasals, others}
d. following conditions of the consonant preceded by a vowel {voiceless cons, voiced cons, vowel, sonorant, nasal}
e. prosodic conditions {unstressed syllable, beginning of stressed syllable}
f. function word vs. content word.

As for vowels the most significant difference is between stressed and unstressed vowels and their position as commented above. In Klatt's experiment, the average duration of a stressed vowel in the connected discourse was 132 ms; while the average duration of unstressed vowels, including schwa, was 70 ms.

To summarize we have the following factors:

– Inherent duration for vowel and consonant type
– Phrase-final and word final syllables
– Postvocalic consonants effect on vowels
– Other syllable-position-in-word and number-of-syllables effects.

In English the main influencing factors are inherent durations. Stressed vowels in word-final syllables of phrase-final words were significantly longer in duration than vowels in any other position, by a percentage of 30%. In English the shortest vowels were in non-final syllable of non-phrase-final words. On the average a vowel in this position was 12% shorter than the median for that vowel type.

Compound words are to be treated as phonological words for the purpose of predicting segment duration.

Klatt's data are non compatible with any strong formulation of an equal-stress timing model for speech production. Equal stress timing is particularly at odds with the large differences observed in inherent durations. The final scheme for a Durational Model of English is as follows (Klatt 1987: 139, but also van Santen et al. 1997):

– each phone type has a different inherent duration;

– vowels not in phrase-final syllables are shorter;
– consonants in syllable onset are made longer by syllabic stress;
– vowels not in word-final syllables are slightly shorter;
– consonants in clusters are shorter;
– the influence of the final consonant in a syllable is small except in phrase-final position;
– unstressed vowels are shorter in duration than stressed vowels;
– segments are uncompressible beyond a certain duration boundary.

2.2.1 *Preprocessing phase and timing modeling*

As far as prosodic elements are concerned, prosodic evaluation is at first approximated from a dynamic comparison with the Master version of the current linguistic item to practice. In order to cope with L1 and L2 on a fine-grained scale of performance judgement, we devised two types of models:

MODEL I. Top-down Syllable-based Model It is a model in which durational structure for a phonological or an intonational phrase is specified first, and then the segmental duration in the words of the grammatical units are chosen as to preserve this basic pattern. The pattern is very well suited for syllable-timed languages, in which the number of syllables and the speaking rate could alone determine the overall duration to be distributed among the various phonetic segments according to phonological and linguistic rules. Mean values for unstressed and stressed syllables could be assigned and then refurbished according to number of phones, their position at clause and phrase level, their linguistic and informational role. Lengthening and shortening apply to mean durational values of segmental durations. In a partial version of this Model, inherent consonant durations are applied at general phonetic classes in terms of compressibility below/above a certain threshold and not at single segments. Since variability is very high at segment level, we apply an "elasticity" model (Campbell and Isard 1991; Campbell 1993) which uses both position and prosodic type to define minima and maxima, and then compute variations by means and standard deviations.

MODEL II. Bottom-up Segment-based Model In this model the starting point is the assignment of inherent duration to each phonetic segment which is followed by use of phonological rules to account for segmental interactions and influences of higher-level linguistic units. For English, Klatt chooses this model which reflects a bias toward attempting to account for durational changes due to local segmental environment first, and then looking for any remaining higher level influences. In this model, the relative terms lengthening and shortening of the duration of a segment has sense if related to inherent duration for a particular segment type. The concept of a limiting minimum duration or equivalently the incompressibility can be better expressed by beginning with the maximum segmental duration (Klatt 1987: 132). In fact, we resort again to the "elasticity" hypothesis at syllable level, since we found that working at segmental level does not produce adequate predictions.

The two models are further implemented as discussed below.

2.2.2 *Segmentation and stress marking*

Consider now the problem of the correct position of stress at word level and the corresponding phenomena that affect the remaining unstressed syllables of words in English: prominence at word level is achieved by increased duration and intensity and/or is accompanied by variations in pitch and vowel quality (like for instance vowel reduction or even deletion, in presence of syllabifiable consonant like "n, d"). To detect this information, the system produces a detailed measurement of stressed and unstressed syllables at all acoustic-phonetic levels both in the master and the student signal. However, such measurements are known to be very hard to obtain in a consistent way (Bagshaw 1994; Roach 2000): so, rather than dealing with syllables, we deal with syllable-like acoustic segments. By a comparison of the two measures and of the remaining portion of signal a corrective diagnosis is consequently issued.

The segmentation and alignment processes can be paraphrased as follows: we have a preprocessing phase in which each word, phonological phrase and utterance is assigned a phonetic description. In turn, the system has a number of restrictions associated to each phone which apply both at subphonemic level, at syllabic level and at word level. This information is used to generate suitable predictions to be superimposed on the segmentation process in order to guide its choices. Both acoustic events and prosodic features are taken into account simultaneously in order to produce the best guess and to ensure the best segmentation.

1. Each digitalized word, phonological phrase or sentence is automatically segmented and aligned with its phonetic transcript provided by the human tutor, with the following sequence of modules:
   - Compute acoustic events for silence detection, silence detection, fricatives detection, noise elimination;
   - Extract Cepstral coefficient from the input speech waveform sampled at 16 MHz, every 5 ms for 30 ms frames;
   - Follow a finite-state automaton for phone-like segmentation of speech in terms of phonological features;
   - Match predicted phone with actual acoustic data;
   - Build syllable-like nuclei and apply further restrictions.

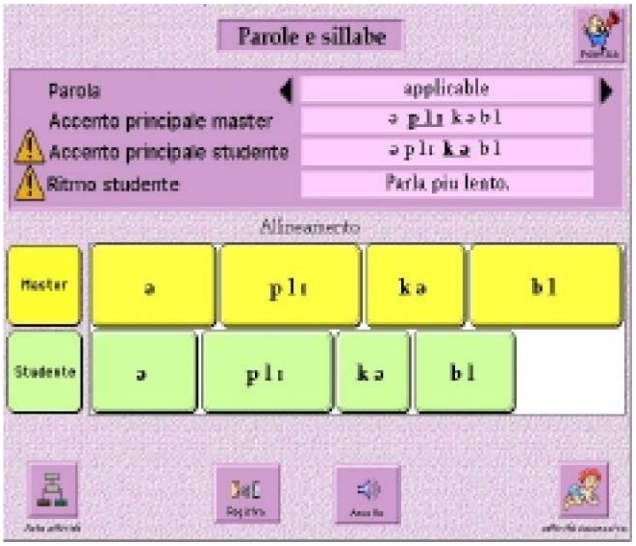


Fig. 1 Syllable level prosodic activities

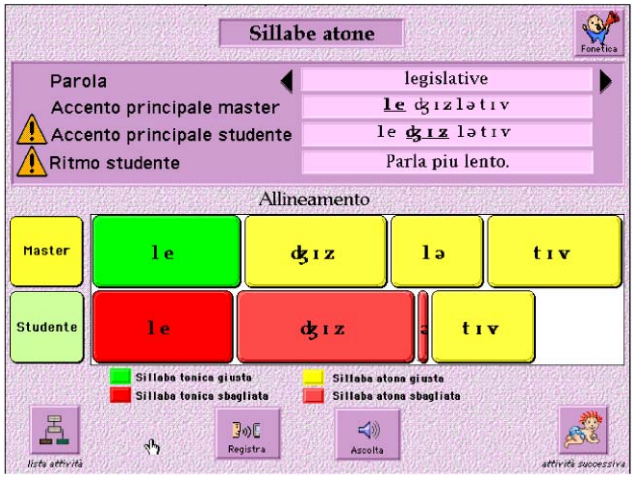


Fig. 2 Word stress prosodic activities

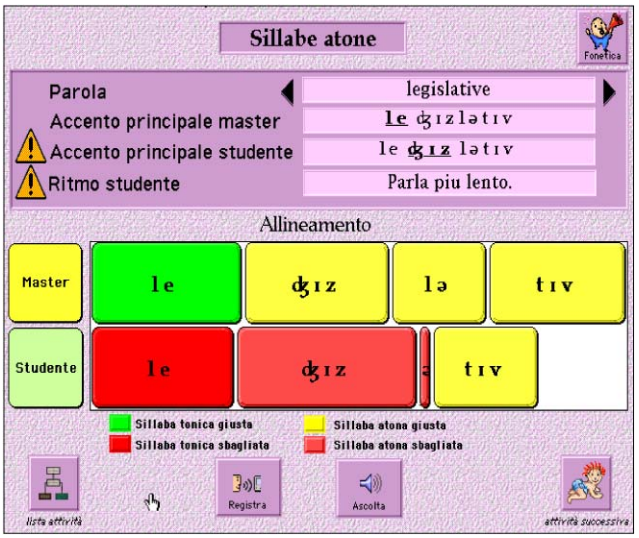


Fig. 3 Unstressed syllables prosodic activities

As mentioned above, the student is presented with a master version of an utterance or a word in the language he is currently practising and he is asked to repeat the linguistic item trying to produce a performance as close as possible to the original native speaker version. This is asked in order to promote fluency in that language and to encourage as close as possible mimicry of the master voice.

The item presented orally can be accompanied by situated visual aids that allow the student to objectivize the relevant prosodic patterns he is asked to mimic. The window presented to the student includes three subsections each one devoted to one of the three prosodic features addressed by the system: stressed syllable/syllabic segment—in case of words—or the accented word in case of utterances, intonational curve, overall duration measurement.

Word-level exercises (see Figs. 1–3) are basically concentrated on the position of stress and on the duration of syllables, both stressed and unstressed. In particular, Italian speakers tend to apply their word-stress rules to English words, often resulting in a completely wrong performance. They also tend to pronounce unstressed syllables without modifying the presumed phonemic nature of their vocalic nucleus preserving the sound occurring in stressed position: so the use of the reduced schwa-like sound [ə], which is not part of the inventory of phonemes and allophones of the source language, must be learned.

The main Activity Window for "Parole e Sillabe"/Words and Syllables is divided into three main sections: in the higher portion of the screen the student is presented with the orthographic and phonetic transcription (in Arpabet) of the word which is spoken aloud by a native speaker's voice. This section of the screen can be activated or disactivated according to which level of Interlingua the student belongs to. We use six different levels (Delmonte et al. 1996a, 1996b).

In particular, the stressed syllable is highlighted between a pair of dots. The main central portion of the screen contains the buttons corresponding to each single syllable which the student may click on. The system then waits for the student performance which is dynamically analysed and compared to the master's. The result is shown in the central section by aligning the student's performance with the master's. According to duration computed for each syllable the result will be a perfect alignment or a misalignment in defect or in excess. Syllables exceeding the master's duration will be shown longer, whereas syllables shorter in duration will show up shorter. The difference in duration will thus be evaluated in proportion as being a certain percentage of the master's duration. This value will be applied to parameters governing the drawing of the related button by HyperCard™. At the same time, in the section below the central one, two warnings will be activated in yellow and red, informing the

student that the performance was wrong: prosodic information concerns the placement of word stress on a given syllable, as well as the overall duration.

In case of error, the student practicing at word level will hear at first an unpleasant sound which is then followed by the visual indication of the error by means of a red blinking syllable button, the one in which he/she wrongly assigned word stress. This is followed by the rehearsal of the right syllable which always appears in green. A companion exercise takes care of the unstressed portion/s of the word: in this case, the student will focus on unstressed syllables and errors will be highlighted consequently in that/those portion/s of the word. Finally the bottom portion of the window contains buttons for listening and recording on the left, arrows for choosing a new item on the right; at the extreme right side a button to continue with a new Prosodic Activity, and at the extreme left side a button to quit Prosodic Activities.

2.2.3 Phonological rules for phonological phrases

Another important factor in the creation of a timing model of L2 is speaking rate, which may vary from 4 to 7 syllables/sec. Changes in speaking rate exert a complex influence on the durational patterns of a sentence. When speakers slow down, a good fraction of the extra duration goes into pauses. On the other hand, increases in speaking rate are accompanied by phonological and phonetic simplifications as well as differential shortening of vowels and consonants. This usually constitutes another important aspect of English self-learning courseware for syllable-timed L2 speakers.

Effects related to speaking rate include compression and elision which take place mainly in unstressed syllables and lead to syllabicity of consonant clusters and of sonorants. Examples at word level include cases such as the following:

a. electrical -> [ɛlɛktɹɪkɬ]
b. usually -> [juʒlɪ]
c. always -> [ɔwəz]
d. problem -> [pɹɒbm]
e. police -> [plis]
f. potassium -> [ptæsiəm]
g. kindly -> [kaynlɪ]
h. rambling -> [ræmlɪŋ]
i. wondering -> [wʌn̪dɪŋ].

Other interword elision and assimilation phenomena will be presented below. These cases require the intervention of phonological rules to produce the adequate representation for a given lexical item. As a result of the opposition between weak and strong syllables at word level (Eskénazi 1999), native speakers of English apply an extended number of phonological rules at the level of Phonological Phrase, i.e. within the same syntactic and phonological constituent. These rules may result in syllable deletion, resyllabification and other assimilation and elision phenomena, which are unattested in syllable-timed languages where the identity of the syllable is always preserved word-internally. In rapid/quick colloquial/familiar style of pronunciation in RP of free conversation and dialogue the effects of elision and compression of vowels and consonants can reach 83% elision at word boundary and 17% internal elision (Delmonte 2000). Here below we list some of the most interesting cases attested in our corpus:

- **Elision**
  - whether you -> [wɛðjə]
  - Baker Street -> [beɪkstɹit]
  - about another -> [əbaəʔnnʌðə]
  - find here -> [faɪnhɪə]
  - bit short -> [bɪʃɔt]
  - with the -> [wɪðə]
  - had happened -> [ədæpm]
  - there's an -> [ðəz n]
  - there are lots -> [ðəɹ lɒts]
  - to question -> [tkwɛʃn]
  - goes on -> [gəz n]
  - they do -> [ðe du]
  - they can go -> [ðe ʔŋ gəə]
  - I was -> [aəz]
  - don't know -> [dənə]
  - to your -> [tjɔ]
  - but we've -> [bəwɪv]

  As far as assimilation is concerned, the main phenomena attested are alveolarization, palatalization, velarization and nasalization some of which are presented here below together with cases attested in our corpus of British English.
- **Homorganic Stop Deletion**. The process of homorganic stop deletion is activated whenever a stop is preceded by a nasal or a liquid with the same place of articulation and is followed by another consonant
- In front of voiced/unvoiced fricative
  - you want some chocolate -> [juwɒnsəm]
  - and this is my colleague -> [ənðɪs]
- **Homorganic Stop Deletion with Glottalization**
  - what can I do for you -> [wɔʔkən]
- Homorganic Liquid and Voiced Stop Deletion in Consonant Cluster
  - you should be more careful -> [juʃəbbɪ]
- **Palatalization Rules affect all alveolar obstruents: /t, d, s, z/**
- **Palatalization of Alveolar Fricative**
  - can I use your phone -> [kenajuʒjɔ]
- **Palatalization of Alveolar Nasal**
  - may I join you? -> [meɪaiʤɔinju]
- **Palatalization of Alveolar Stop**
  - nice *to meet you* -> [tmɪʧjə]

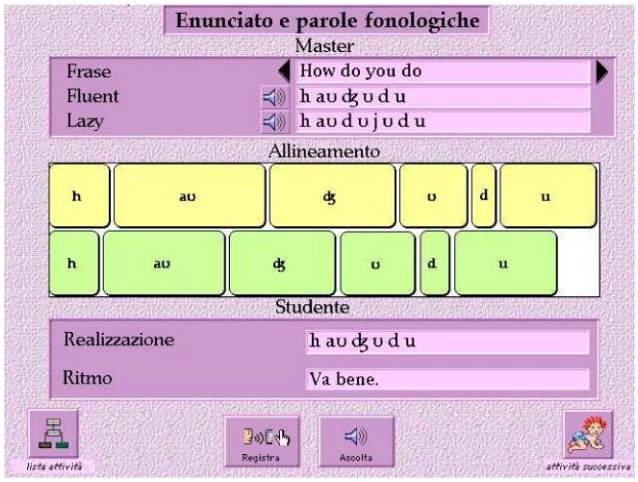


Fig. 4 Phonological phrase level prosodic activities

- **Degemination**
  - *just take a* seat -> [ʤasteɪkə]
- **Velarization**
  not quite -> [nɒʔkwaɪt]

In order to have Italian students produce fluent speech with phonological rules applied properly we decided to set up a Prosodic Activity which offered the two versions of a single phrase taken from the general course being practised. The student could thus hear both the "lazy" version, with carefully pronounced words, and no rule application taking place; then, the second version, with a fluent and quicker speech is spoken twice. This latter version starts flashing and stops only when the student records his/her version of the phrase.

A comparison then follows which automatically checks whether the student has produced a phrase which is close enough to the "fluent" version. In case the parameters computed are beyond an allowable threshold, the comparison proceeds with the "lazy" version in order to establish how far the student is from the naive pronunciation. The assessment will be used by the Automatic Tutor to decide, together with similar assessments coming from Grammar, Comprehension and Production Activities, the level of Interlingua the student belongs to.

3 Prosodic tools for self-learning activities in the domain of intonation

3.1 General problems related to intonation in language teaching

In his PhD dissertation and in a number of recent papers M. Jilka (Jilka and Möhler 1998; Jilka 2000) analyzes the problem of intonational foreign accent (IFA) in the speech of American speakers of German. The definition of what constitutes a case of intonational foreign accent seems fairly straightforward: the intonation in the speech of a non-native speaker must deviate to an extent that is clearly inappropriate for what is considered native. The decision of what intonation is inappropriate or even impossible strongly depends on the surrounding context, much more so than it is the case for deviations in segmental articulation. It is therefore a prerequisite for the analysis of intonational foreign accent that the context be so clear and narrow as to allow a decision with respect to the appropriateness of a particular intonational realization.

This can be done in terms of a categorical description of intonation events based on ToBI labelling. Results show that IFA does indeed include categorical mistakes involving category type and placement, transfer of categories in analogous discourse situations, and deviating phonetic realization of corresponding tonal categories.

While such an identification of IFA based on ToBI labeling can be easily achieved in an experimental situation, where transcriptions are all done manually, in a self-learning environment the same results would all be based on the ability of the underlying algorithm to achieve a confident enough comparison between a Master and Student signal.

To comply with the idea that only categorical deviations are relevant in the determination of IFA and that it is sensible to propose appropriate corrective feedback only in such cases we need to start from semantically and pragmatically relevant intonational countours as will be discussed in a section below.

As Jilka (2000: Chap. 3) suggests, the main difference in evaluating segmental (allophones) vs suprasegmental (allotones) variations in an L2 student's speech, is that a broader variational range seems to be allowed in the realization of intonational features. We are then faced with the following important assumptions about the significance of variation in the identification of intonational deviations:

- intonation can be highly variable without being perceived as foreign accented (A1)
- context-dependent variation in intonational categories is greater than in segmental categories (A2).

The first assumption (A1) presupposes that the fact that intonation allows a high degree of variation in the choice and distribution of tonal categories is a major aspect aggravating the foreign accent identification process. Noticeable variations may retain the same or a slightly different interpretation, but are not perceived as inappropriate, i.e. foreign-accented. Measurable variations from an assumed prototypical realization may not be perceived at all (thus being basically irrelevant), perceived as different, but not interpreted as such, or actually interpreted as different, but not as foreign.

Table 8 Tone inventories of American English and Italian

	American English	Italian
Pitch accents	H*, L*, L + H*, L* + H, H + !H*	H*, L*, L + H*, L* + H, H + L*
Initial Phrasal tones	%H	%H
phrase accents	H-, L-	H-, L-
boundary tones	L-L%, L-H%, H-L%, H-H%	L-L%, L-H%, H-L%, H-H%

Consequently, a second assumption (A2) about variation in intonation must contend that intonational categories may have more context-dependent different phonetic realizations ("allotones") than segmental categories. This further increases the difficulty in identifying intonational foreign accent, even though, as already mentioned, a number of those additional phonetic realizations do not contribute to foreign accent.

We will compare the two tone inventories as they have been reported in the literature and then we will make general and specific comments on the possibility for an automatic comparing tool to use them effectively.

The American English inventory (Jilka 2000: Chap. 4), contains five types of pitch accent, two of them monotonal (H*, L*), the other three bitonal (L* + H, L + H*, H + !H*), thus implying an inherent F0 movement (rise or fall) between two targets. Phrasing in American English is determined by two higher-level units, intermediate phrases (ip's) and intonation phrases (IP's). Phrasal tones either high or low in the speaker's pitch range mark the end of these phrases. For intermediate phrases they are called phrase accents (H-, L-), for intonation phrases the term boundary tone (H%, L%) is used. As the terminology suggests, ip's and IP's are ordered hierarchically. An IP consists of one or more ip's and one or more IP's make up an utterance. For this reason, the end of an IP is by definition also the end of an ip, and a boundary tone is always accompanied by a phrase accent, allowing four possible combinations: L-L%, L-H%, H-L% and H-H%.

Even though the two inventories are almost identical, the range of variation in intonation contours is used in a much richer way in American English rather than in Italian (Avesani 1995).

The deviations are summarized in an inventory of nine major differences in the productions of the Dutch speakers Willems (1983). The listed deviations, which correspond to distinct instances of intonational foreign accent, include what Willems terms

- the direction of the pitch movement (sometimes Dutch speakers may use a rise where British English speakers use a fall)
- the magnitude of the pitch excursion (smaller for the Dutch speakers)
- the incorrect assignment of pitch accents
- differences in the F0 contour associated with specific tonal/phrasal contexts and discourse situations such as continuations (Dutch speakers often produce falls)
- the F0 level at the beginning of an utterance (low in Dutch speakers, but mid in British English speakers) or
- the magnitude of final rises in Yes/No-questions (much greater in Dutch speakers).

Taking into consideration theory-dependent differences in terminology, a number of Willems' results are confirmed in this study' s comparison of German and American English.

3.1.1 Teaching intonation as discourse and cultural communicative means

Chun (1998) emphasizes the need to look at research been conducted to expand the scope of intonation study beyond the sentence level and to identify contrasting acoustic intonational features between languages. For example, Hurley (1992) showed how differences in intonation can cause sociocultural misunderstanding. He found that while drops in loudness and pitch are turn-relinquishing signals in English, Arabic speakers of English often use non-native like loudness instead. This could be misinterpreted by English speakers as an effort to hold the floor (Hurley 1992: 272–273). Similarly, in a study of politeness with Japanese and English speakers, Loveday (1981) found more sharply defined differences in both absolute pitch and within-utterance pitch variation between males and females in Japanese than between English males and females in English politeness formulas. In addition, the Japanese subjects transferred their lower native language pitch ranges when uttering the English formulas. Low intonation contours are judged by native speakers of English to indicate boredom and detachment, and if male Japanese speakers transfer their low contours from Japanese to English when trying to be polite, this could result in misunderstandings by native English speakers.

As evidence for culture-specificity with regard to the encoding and perception of affective states in intonation contours, Luthy (1983) reported that although a set of "nonlexical intonation signals" (Luthy 1983: 19) (associated with expressions like uh-oh or mm-hm in English) were interpreted consistently by a control group of English native speakers, non-native speakers of varied L1 backgrounds tended

to misinterpret them more often. He concluded that many foreign students appear to have difficulty understanding the intended meanings of some intonation signals in English because these nuances are not being explicitly taught.

Kelm (1987), acknowledging that "correct intonation is a vital part of being understood" (Kelm 1987: 627), focused on the different ways of expressing contrastive emphasis in Spanish and English. He investigated acoustically whether the range of pitch of non-native Spanish speakers differed from that of native Spanish speakers. Previous research by Bowen (1975) had found that improper intonation in moments of high emotion might cause a non-native speaker of Spanish to sound angry or disgusted. Kelm found that the native Spanish-speaking group clearly varied in pitch less than the two American groups; that is, native English speakers used pitch and intensity to contrast words in their native language and transferred this intonation when speaking Spanish. Although the results showed a difference between native and non-native Spanish intonation in contrasts, they did not show the degree to which those differences affect or interfere with communication.

In intonation teaching, one focus has traditionally been contrasting the typical patterns of different sentence types. Pitch-tracking software can certainly be used to teach these basic intonation contours, but for the future, in accordance with the current emphasis on communicative and sociocultural competence, more attention should be paid to discourse-level communication and to cross-cultural differences in pitch patterns. According to Chun (1998), software programs must have the capability to:

- Distinguish the meaningful intonational features with regard to four aspects of pitch change: (a) direction of pitch change (rise, fall, or level), (b) range of pitch change (difference between high and low levels), (c) speed of pitch change (how abruptly or gradually the change happens), and (d) place of pitch change (which syllable(s) in an utterance)
- Go beyond the sentence level and address the multiple levels of communicative competence: grammatical, attitudinal, discourse, and sociolinguistic.

3.1.2 From syllable structure to intonational phrase

As has been clearly shown in recent experiments (Bannert 1987; Batliner et al. 1998; Breen 1995) prosodic information can be fruitfully used in speech recognition tasks, contributing higher level linguistic knowledge and improving the overall performance of the system. Stress and rhythm related phenomena seem to be explainable mainly by means of syllable level prosodic labeling. From a linguistic point of view this is however not sufficient: we are convinced of the need of enhancing our database labeling, in order to encode higher level suprasegmental information which could be used at utterance level to detect syllable structure modification phenomena. In particular, we are interested in encoding "core factors" which include the following ones:

– phone identity factors: current segment, previous segment(s), next segment(s);
– stress related factors: degree of discourse prominence, lexical stress;
– locational factors: segment in the syllable, syllable in the word, word in the phrase, phrase in the utterance.

Ten factors listed by degree of contribution were reported by Campbell and Isard (1991), Campbell (1993: 1083) as relevant to model durations at syllable level ending up in 39 different types:

a. number of segments in the syllable—seven levels;
b. break index—four levels;
c. nature of the rhyme—open/closed;
d. function/content distinction;
e. nature of the peak—four classes;
f. stress index—four levels;
g. type of foot—headed or not;
h. number of syllables in the foot—six levels;
i. position in the word—four classes;
k. position of the phrase in the utterance—four classes.

The list of breaks and boundaries proposed in Verbmobil project is reported here below (see Batliner et al. 1998, Table 8, pp. 207–208), obeys to the following Hierarchy: Main, Subordinate, Coordinate—M, S, C (see Table 14, p. 209) and contains the following 24 types, plus a jolly one. Prosodic syntactic strength: strong(3), intermediate(2), weak(1), very weak(0).

Our classification (Bacalu and Delmonte 1999) starts from Syllable Type and then deals with Syllable Structure and Stressed Syllable Structure; then at a higher level it defines the Clitic Group and proposes a number of labels starting from word level up to Intonational Group. At word level it is important to take into account both number of syllables, the position of stressed syllable and the sequence of unstressed syllables in order to be able to forecast compensation and other restructuring phenomena in English and Italian.

CG CLITIC GROUP

CG A word may be a proclitic or be a head of a CG.
CG1 A word may be a head of a CG and be final
CG2 A word may be a head of a CG and not be final
CG3 A word may be a proclitic, depending on its grammatical or lexical tag
CG4 A word may be a head not CG. final, and act as a local focalizer.

IG INTONATIONAL GROUP

Table 9 List of breaks and boundaries used by verbmobil

SM3—Main clause and main clause	IC0—every other boundary
SM2—Main clause and subordinate clause	PM3—free Phrase, stand alone
SS2—Subordinate clause and main clause	PC2—sequence in free Phrase
SM1—Main clause and subordinate clause, prosodically integrated (complement clause)	PM1—free Phrase, prosodically integrated
SS1—Subordinate clause and main clause, prosodically integrated (complement clause)	SC3—Coordination of main clause and subordinate clause
LS2—Left dislocation	RS2—Right dislocation
SC2—Subordinate clause and subordinate clause	RC2—sequence of Right dislocations
FM3—pre-/postsentential particle, with pause etc.	RC1—Right dislocation at open verbal brace
DS3—pre-postsentential discourse particle, ambisentential	EM3—embedded sentence/phrase
AM3—between sentences, Ambiguous	DS1—pre-postsentential discourse particle, no pause
AC1—between constituents, Ambiguous	AM2—between free phrases, Ambiguous
IC1—asyndetic listing of Constituents	IC2—between Constituents

IG A CG may be sentence final or not

IG1 A CG is sentence final if it is included in the higher S symbol or at functional level within a single clause with its own main predicate; sentences included in sentential complement do not count as Phonological Sentences;

IG2 A subordinate clause is a closed Adjunct and is included in a separate IG: it may precede or follow the Main Clause;

IG3 The same applies to Interjections like "per favore/ please", which may be regarded as free phrases, with a separate IG;

IG4 A different status must be assigned to open Adjuncts, like Relative Clauses;

IG5 A CG may be sentence final, receive sentence stress and be prepausal this is the canonical position for sentence stress, whenever there is no extraposed or right dislocated constituent;

IG6 A CG may be sentence final and receive sentence stress: this is the position for sentence stress followed by a PhPh right dislocated or extraposed;

IG7 A CG may be sentence final and have no sentence stress: this is the role played by extraposed or right dislocated constituents;

IG8 A CG may be sentence final, have no sentence stress and be prepausal: this is the role played by all extraposed or right dislocated constituent in prepausalposition;

IG9 A CG may be sentence final, have no sentence stress and not be prepausal: this is the case of more than one constituent right dislocated or extraposed;

IG10 A CG not sentence final with sentence stress: this is the case of all constituents in object position either inverted subjects or verb phrases with sentence stress, followed by an extraposed constituent;

IG11 A CG not sentence final: any other PhPh, like all subject Nps which are not sentence final and receive no sentence stress, nor are prepausal.

UG UTTERANCE GROUP

UG1 A Focussed Phrase is coincident with Semantic Focus.

UG2 A Focussed Phrase is not coincident with Semantic Focus.

UG3 A Topic Dislocated Phrase is Extraposed and is UG final.

UG4 A Topic Dislocated Phrase is Extraposed and is not UG final.

UG5 A Topic Dislocated Phrase is Extraposed, is not UG final, and has parenthetical intonation.

3.2 Intonation practice and visualization: our approach

As to Intonational Group detection and feedback, from a number of studies in Dialog Acts it seems clear that intonation is very important in the development of DA classifiers and automatic detector for conversational speech. From the work published in Shriberg et al. (1998) however, we may assume that in the 42 different DA classified only 2 acoustic features were actually considered relevant for the discrimination task: duration and FØ curve. This same type of information is used by our system for intonation teaching. We also assume that word accent is accompanied by ***FØ*** movement so that in order to properly locate pitch accent we compute ***FØ*** trajectories first. Then we produce a piecewise stylization which appears in the appropriate window section and is closely followed by the ***FØ*** trajectory related to the student's performance so that the student can work both at an auditory and at a visual level.

The stylization of an ***FØ*** contour aims at removing the microprosodic component of the contour. Prosodic representation is determined after ***FØ*** has been resolved, since

FØ acts as the most important acoustic correlate of accent and of the intonational contour of an utterance. Basically, to represent the intonational contour, two steps are executed: reducing errors resulting from automatic pitch detection and then stylisation of ***FØ*** contour. The stylisation of ***FØ*** contour results in a sequence of segments, very closed to local movements in speaker's intonation. We tackled these problems in a number of papers (see Delmonte 1983a, 1983b, 1984, 1985a, 1985b, 1987b, 1988a, 1988b) where we discuss the relation existing between English and Italian intonational systems both from a theoretical point of view and on the basis of experimental work.

3.2.1 FØ tracking algorithm

At utterance level, word accent is accompanied by ***FØ*** movement so that in order to properly locate pitch accent we compute ***FØ*** trajectories first. Then we produce a piecewise stylization which appears in the appropriate window section and is closely followed by the ***FØ*** trajectory related to the student's performance so that the student can work both at an auditory and at a visual level.

The method we propose takes as pitch period the time interval between two significant peaks (Delmonte et al. 1997). Informally, the algorithm consists of three basic actions: deciding whether a peak is actually significant, trying to start a new pitch trace and trying to continue a trace already detected with a new significant peak. There are two data structures used in pitch detection: a buffer containing peaks and an array keeping information about the pitch "paths" in progress. The buffer is a queue of cells, each of them carrying information on time and amplitude of the current peak in the original signal, but also (and most important) the different traces the peak can participate in. The TRACES array simply keeps the pitch traces under development, each trace being uniquely identified by a label. As already suggested, the algorithm tries to develop simultaneously traces which could intersect with one another, and finally chooses the FØ trajectory among them. The first action of the algorithm decides which peaks are significant and is performed conforming to two criteria:

– amplitude of peaks;
– permissible range of FØ.

Then the filtered peaks are entered into the buffer, as candidates participating in different traces. Periodicities are searched for within this buffer. The length of the buffer is calculated according to the formula:

$$\text{buffer\_length} > 2 * \text{Tmax}/\text{Tmin}$$

where Tmin and Tmax denote the permissible range of the fundamental frequency. Two adjacent peaks cannot occur closer than the minimum permissible period and also, the buffer must be long enough to permit tracing, therefore at least three peaks of the longest possible period. As emphasized above, each item in the buffer keeps a record of the participation of the corresponding peak in different traces; so, each peak entering in the buffer can trace at most Tmax/Tmin traces and this limits also the number of channels that must be allocated for each of the elements of the buffer.

The algorithm (Delmonte et al. 1997; Bagshaw et al. 1993b; Medan et al. 1991) proceeds as follows: suppose that at a certain moment the buffer snapshot displays already a number of traces in progress, each trace "knowing" about the last period introduced (that is, the time interval between the two last peaks in trace), lets call it T1; suppose also that the supervisor performing the first action already decided which was the next significant peak and entered it in a buffer. What the supervisor further does is to verify whether (and which of) the traces under development can be continued with the current tail of the buffer. First, it verifies if the time distance between the tail-peak in buffer and the tail-peak of a trace, lets call it T2, is greater then Tmax (T2 > Tmax); if it is, the trace is closed. Otherwise, the supervisor checks whether T1 > T2. If it isn't, the supervisor expects another peak to continue the trace; otherwise, the tail peak in the buffer is marked as participating in the trace which is continuing, where we also annotate its last peak and last period (that is, T2). The other action done by the supervisor is to start new traces, that is, it checks whether three peaks, the tail in buffer and two previously introduced in buffer, can start a new trace; the criteria used are the same: the periods can not get out of the range between Tmin and Tmax, and the two periods corresponding to the three peaks must have values which are close enough.

Pitch tracking is prone, as explained, at finding multiple traces simultaneously. Therefore a post-processing phase is needed to disambiguate these multiple traces. One can simply ask: why this aptitude at finding more traces simultaneously, to finally choose only one of them? One answer is that the final pitch trajectory should reflect the periodicity of some patterns in signal's behaviour; for instance, inside a vowel the signal is slightly changing in time, but there is usually a pattern which is repeating. More traces could be a way to reflect the patterns and also to reflect their periodicity. They can bring a lot of information when detecting phoneme boundaries in signal's segmentation. There is a drawback however when dealing with multiple traces; since the range of frequencies a priori established for FØ is rather large, the algorithm can result in traces which are simply mirroring of halves or doubles of a real pitch trace. Fortunately, these situations can be detected in most cases, using a deeper analysis of peaks' amplitudes and this is a complementary phase to the first action of the algorithm. Basically, post-processing consists of two phases: the first one

eliminates doubling and halving errors, and proceeds simultaneously to traces development (in a sense is acting as an eliminating criterion), and the other independently chooses the final FØ trace. As emphasized, in most cases the first phase is enough. There are also extreme cases when simply analyzing energy amplitudes doesn't suffice and conflicting traces remain. The last phase intends to solve them.

The method implemented relies on fuzzy logics (Delmonte et al. 1996a). When choosing fuzzy methods the aim was that of postponing a decision about a local phenomenon until uncertainty has been reduced or even eliminated by successive processing phases. In this way, the algorithm has only to take care of a higher number of possible candidates which are ranked according to subjective qualitative fuzzy propositions. That is, there are fuzzy propositions and fuzzy overlapping sets created to give a score with a truth degree to each uncertain trace. Finally, the trace with the higher score and with the higher truth degree is chosen as FØ trace.

One of the fuzzy propositions used (and subsequently illustrated) has the following motivation: the closer a pitch segment is to one of its neighbours in time (let's say the right one), the closer the first period of this neighbour and the last period of the current segment inspected should be. That is, in short time intervals, the FØ contour is gradually changing, while in long ones it can change significantly. Consequently, when a pitch segment is added to a partial pitch trace, it is inspected to determine whether it naturally continues what was already constructed. If the trajectory is developing right-to-left, then the contribution of the current pitch segment inspected to a non-conflicting trace is asserted as BAD, ACCEPTABLE or GOOD. This is done according to a function of the last pitch period of the current segment and the first pitch period of the right neighbour, and the truth degree is computed by a function of time interval between the segment and the neighbour.

The actual implementation uses backtracking; only the leaves of the tree (identifying complete traces) are object to defuzzification. The performance is still good, for the uncertain traces are in small number due to the first phase. However, the algorithm could be further improved with a branch-and-bound method, which could be synchronized to the development of uncertain traces, heuristically based on defuzzification of partial trajectories.

3.2.2 Intonational curve representation

In the generation of an acoustic-phonetic representation of prosodic aspects of speech for computer aided pronunciation teaching, the stylization of an FØ contour aims to remove the microprosodic component of the contour. Prosodic representation is determined after the fundamental frequency has been resolved, since fundamental frequency acts as the most important acoustic correlate of accent and of the intonational contour of an utterance. Basically, to represent the

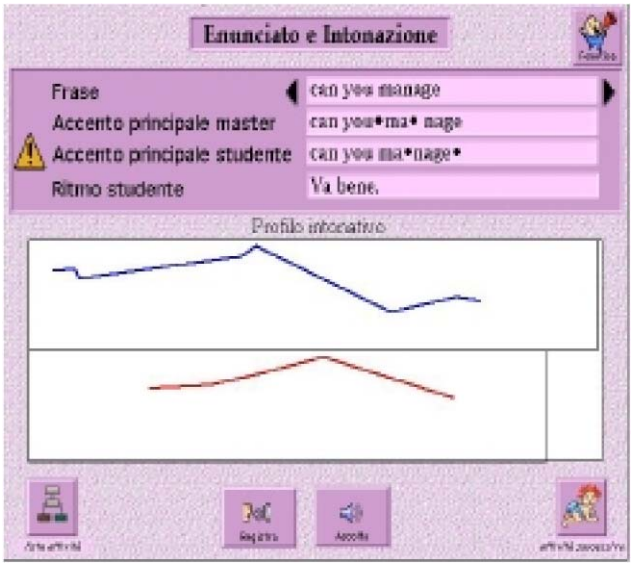


Fig. 5 Utterance Level Prosodic Activities: 1

intonational contour, two steps are executed: reducing errors resulting from automatic pitch detection and then stylization of FØ contour. The stylization of FØ contour results in a sequence of segments, very closed to local movements in speaker's intonation. As highlighted above, the pitch resulted is a "direct-period" mirroring. To compute FØ, one might implement the frequency function $\text{FØ}(t) = 1/T(t)$. However, by this method dissimmetries will eventually result: on rising portions of $T(t)$, $\text{FØ}(t)$ is normally compressed, while on falling portions of $T(t)$, $\text{FØ}(t)$ is stretched. As the displayed pitch is intended to put in evidence the rising portions of $\text{FØ}(t)$ where accent appears, we prefer to simply compute a symmetric function of the $T(t)$ slope instead of calculating the $\text{FØ}(t)$ as $1/T(t)$. In this way we achieve two goals at one time: the normal compression is thus eliminated, and we save computation time (Delmonte et al. 1997).

To classify pitch movements we use four tone types: rising, sharp rising, falling and sharp falling, where the "sharp" versions coincide in fact with main sentence accent and should be time aligned with it. The classification is based on the computation of the distance to the line between beginning and the end of a section, compared on the basis of an a priori established threshold. For instance, in Fig. 5, some parameters of the utterance "Can you manage?" are described pronounced by a female speaker, where the intonational contour smoothly falls in the first word, then follows a level pattern in "you", to reach a sharp rise and then a fall in the first syllable of "manage", with a smooth final rising (as expected, since the utterance is a yes/no question).

The following tagging conventions were used.

Regarding **pitch level**:

L—low (-> 140 Hz men, 140 -> 170 women);

H—high (180 -> 220 Hz men, 210 -> 280 women);
Z—extremely high (240 -> Hz men, 330 -> women);

Regarding **pitch movements**:
+—rising
+*—sharp rising (main accent)
-—falling
-*—sharp falling (main accent).

3.3 Self-learning activities in the prosodic module: Utterance Level Exercises

In Utterance Level Prosodic Activities the student is presented with one of the utterances chosen from the course he is following. Rather than concentrating on types of intonation contours in the two languages where performance-related differences might result in remarkable intraspeaker variations, we decided to adopt a different perspective. Our approach is basically communicative and focuses on a restricted number of communicative functions from the ones the student is practising in the course he is following (for a different approach see 41 on Japanese-English). Contrastive differences are thus related to pragmatic as well as performance factors.

In the course, the student will address some or all of the following communicative functions:

1. Describing actions: habitual, future, current, past; 2. Information: ask for, indicate something/someone, denoting existence/non existence; 3. Socializing: introduce oneself; on the phone; 4. Expressing Agreement and Disagreement; 5. Concession; 6. Rational enquiry and exposition; 7. Personal emotions: Positive, Negative; 8. Emotional relations: Greetings, Sympathy, Gratitude, Flattery, Hostility, Satisfaction; 9. Categories of Modal Meaning, Scales of certainty: i. Impersonalized: Affirmation, Certainty, Probability, Possibility, Negative Certainty; ii. Personalized: Conviction, Conjecture, Doubt, Disbelief; iii. Scale of commitment; iv. Intentionality; v. Obligation; 10. Mental Attitudes: Evaluation; Verdiction; Committal; Release; Approval; Disapproval; Persuasion; Inducement; Compulsion; Prediction; Tolerance.

All these communicative functions may be given a compact organization within the six following more general functions or macrofunctions:

1. ASK; GIVE, OFFER, CONSENT;
2. DESCRIBE; INFORM;
3. SOCIALIZE.
4. ASSERT, SAY, REPLY;
5. EXPRESS EMOTIONS, MODALITIES;
6. MENTAL ATTITUDES;

Each function has been given a grading according to a scale of six levels. The same applies to the grading of grammatical items, be they syntactic or semantic, by classifying each utterance accordingly. The level index is used by the Automatic Tutor which has to propose the adequate type of exercise to each individual student (Delmonte et al. 1996a, 1996b). As far as the Activity Window is concerned—"Enunciato e Intonazione"/Utterance and Intonation, the main difference from Word Level Prosodic Activities discussed above concerns the central main portion of the screen where, rather than a sequence of syllable buttons, the stylized utterance contours appear in two different colours: red for student, blue for master. After each student's rehearsal, the alignment will produce a redrawing of the two contours with different sizes in proportion with the master's one. In the example shown in Fig. 5, above, sentence accent goes on first syllable of the verb "manage" in the Master version, while the student version has accent on the second syllable of the same word "manage".

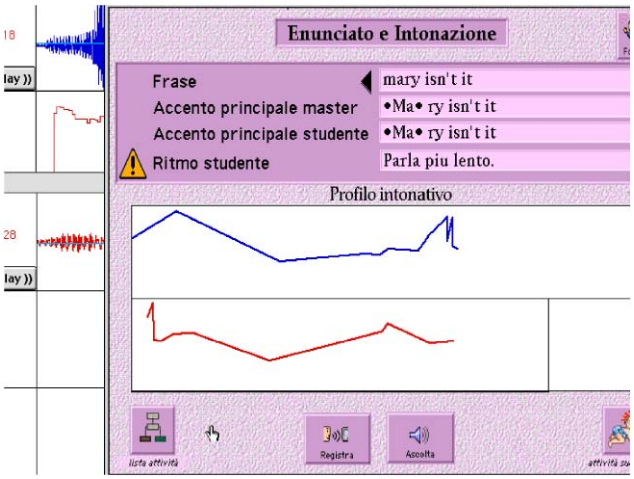


Fig. 6 Utterance Level Prosodic Activities: 2

In the second example, see Fig. 6, we show a Tag-Question, where the difference between the two performances are only in rhythm. Both the initial accent on "Mary" and the final rising pitch on "it" are judged satisfactory by the system which can be seen on the back of the student's activity window.

The third and final example, see Fig. 7, is a simple utterance "Thank you", which however exhibits a big FØ range from the high level of the first peak on the word "Thank" to the low level of the word "you", making it particularly hard for Italian speakers to reproduce correctly.

4 Conclusions

A method of creating a syllable database using an available read speech database that contains hand-annotated time-aligned phoneme-level and word-level transcriptions of each utterance has been presented. We also showed how using this database enabled us to improve the performance of the

Table 10 Evaluation of Prosodic and Phonetic Abilities

Linguistic elements/ levels	Word stress pos.	Pitch accents word/utter.	Tense—Lax feature acquisition	Syllable reduction	Grapheme-phoneme transduction	Syllabic consonants	Execution time
Lev. 3	95%	100%	100%	98%	98%	100%	100%
Lev. 2	75%	70%	80%	90%	80%	70%	80%
Lev. 1	30%	10%	25%	25%	40%	10%	50%

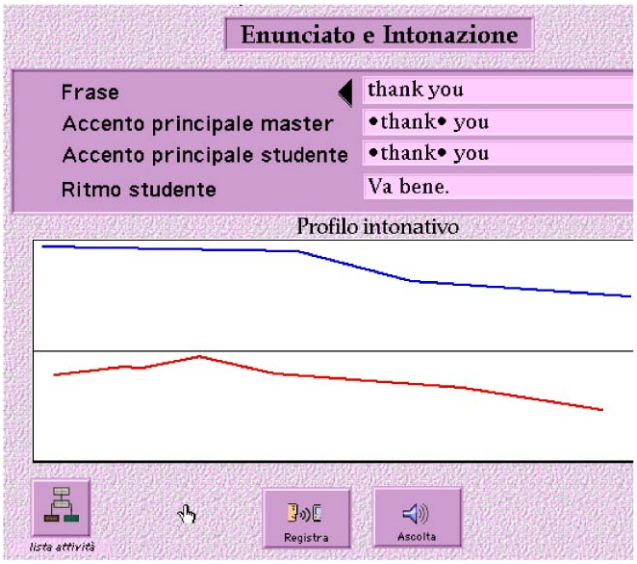


Fig. 7 Utterance Level Prosodic Activities: 3

Prosodic Module in our Multimedia Interactive Linguistic Software.

As a first step in the valuation of our Prosodic Module, we designed an experiment made of data collection phase in order to spot common mistakes followed by a period of two weeks exposure to Prosodic Activities and then a final test. In order to conduct a segmental and suprasegmental error analysiswe worked on the segmental inventory of the target language by creating a word list of 100 items that contains all of the vowel, diphthong, and consonant phonemes in as many syllable positions as possible. We also included in the list several polysillabic words designed to elicit typical word-level stress placement errors.

The errors produced in the readings have been collated and counted, yielding a subset of the words of the inventory list which contain the words in which pronunciation errors were actually made by the talkers. We also eliminated errors that occur only once to insure that the errors are representative of the group of speakers.

We not only found errors in vowel tenseness, and consonant voicing, but a lot of epenthetic consonants added when words ended with fricative consonants, lots of errors in velar nasal pronunciation, dental fricatives, /r/ as a vibrant, /t, d/ as dental stops to list only the most frequent ones. None of the students produced reduced vowels and syllabic consonants properly.

In phase 1 production task, visual prompts are displayed and subjects respond by speaking the word with no immediate auditory model. The system records each performance without evaluating it. On the contrary, when using the Prosodic Module, students receive a lot of feedback centered mainly on syllable level and on correcting the errors we expected. Preliminary final achievement tests on 20 Italian students show that an average of 10 hours practice sufficient for the student to go from Level 2 to Level 4.

The system is currently undergoing its first evaluation phase with students of English for Foreign Languages and Literatures, in self-access modality. More information is needed on the efficiency and feasibility of computer-based self-instruction in order to be able to assess its impact in a real University course. From the first feedback provided by students who chose to enroll in a half-tutored half-self-access course, work on Prosodic Activities has proven very useful and rewarding. In the following Table we report preliminary results of a screening on prosodic abilities acquisition applied on a sample of 50 students from different levels of linguistic abilities.

Acknowledgements I would like to thank Dan Cristea, Ciprian Bacalu and people who worked on SLIM project.

References

Avesani, C. (1995). ToBIT: un sistema di trascrizione per l'intonazione italiana. In *Atti delle 5 Giornate di Studio GFS*, Povo (TN) (pp. 85–98).

Bacalu, C., & Delmonte, R. (1999). Prosodic modeling for syllable structures from the VESD—venice English syllable database. In *Atti 9° Convegno GFS-AIA*, Venezia.

Bagshaw, P. (1994). *Automatic prosodic analysis for computer aided pronunciation teaching*. Unpublished PhD Dissertation, Univ. of Edinburgh, UK.

Bagshaw, P., Hiller, S., & Jack, M. (1993a). Computer aided intonation teaching. In *Proceedings of Eurospeech, 93* (pp. 1003–1006).

Bagshaw, P. C., Hiller, S. M., & Jack, M. A. (1993b). Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching. In *Proc. Eurospeech93*, Berlin (pp. 1003–1006).

Bannert, R. (1987). From prominent syllables to a skeleton of meaning: a model of a prosodically guided speech recognition. In *Proceedings of the XIth ICPhS* (Vol. 2, p. 22.4).

Batliner, A., Kompe, R., Kiessling, A., Mast, M., Niemann, H., & Noeth, E. (1998). M—Syntax + Prosody: a syntactic-prosodic labelling scheme for large spontaneous speech databases. *Speech Communication*, *25*(4), 193–222.

Bernstein, J., & Franco, H. (1995). Speech recognition by computer. In N. Lass (Ed.), *Principles of experimental phonetics* (pp. 408–434). New York: Mosby.

Bertinetto, P. M. (1980). The perception of stress by Italian speakers. *Journal of Phonetics*, *8*, 385–395.

Bowen, J. D. (1975). *Patterns of English pronunciation*. New York: Newbury House.

Breen, A. P. (1995). A simple method of predicting the duration of syllables. In *Eurospeech'95* (pp. 595–598).

Campbell, W. (1993). Predicting segmental durations for accommodation within a syllable-level timing framework. In *Eurospeech '93* (pp. 1081–1085).

Campbell, W., & Isard, S. (1991). Segment durations in a syllable frame. *Journal of Phonetics*, *19*, 37–47.

Chun, D. M. (1998). Signal analysis software for teaching discourse intonation. *LLTJ, Language Learning & Technology*, *2*(1), 61–77.

Delmonte, R. (1981). L'accento di parola nella prosodia dell'enunciato dell'Italiano standard. In *Studi di Grammatica Italiana, Accademia della Crusca*, Firenze (pp. 69–81).

Delmonte, R. (1983a). A phonological processor for Italian. In *Proceedings of the 1st conference of the European chapter of ACL*, Pisa (pp. 26–34).

Delmonte, R. (1983b). *Regole di Assegnazione del Fuoco o Centro Intonativo in Italiano Standard*. CLESP, Padova.

Delmonte, R. (1984). *On certain differences between English and Italian in phonological processing and syntactic processing*. Manuscript, Università di Trieste.

Delmonte, R. (1985a). Parsing Difficulties & Phonological Processing in Italian. In *Proceedings of the 2nd conference of the European chapter of ACL*, Geneva (pp. 136–145).

Delmonte, R. (1985b). Sintassi, semantica, fonologia e regole di assegnazione del fuoco. In *Atti del XVII congresso SLI*, Bulzoni, Urbino (pp. 437–455).

Delmonte, R. (1987a). The realization of semantic focus and language modeling. In *Proceedings of the XIth ICPhS* (Vol. 2, 24.1, pp. 101–104).

Delmonte, R. (1987b). The realization of semantic focus and language modeling. In *Proceeding of the international congress of phonetic sciences*, Tallinn, URSS (pp. 100–104).

Delmonte, R. (1988a). Analisi Automatica delle Strutture Prosodiche. In R. Delmonte, G. Ferrari, & I. Prodanoff (Eds.), *Studi di linguistica computazionale* (pp. 109–162). Padova: Unipress, Cap. IV.

Delmonte, R. (1988b). Focus and the semantic component. In *Rivista di Grammatica Generativa* (pp. 81–121).

Delmonte, R. (1991). Linguistic tools for speech understanding and recognition. In P. Laface & R. De Mori (Eds.), *NATO ASI Series: Vol. F 75. Speech recognition and understanding: recent advances* (pp. 481–485). Berlin: Springer.

Delmonte, R. (1998). Prosodic modeling for automatic language tutors. In *Proc. STiLL'98, ESCA*, Sweden (pp. 57–60).

Delmonte, R. (1999). Prosodic variability: from syllables to syntax through phonology. In *Atti IX Convegno GFS-AIA*, Venezia (pp. 133–146).

Delmonte, R. (2000). SLIM prosodic automatic tools for self-learning instruction. *Speech Communication*, *30*, 145–166.

Delmonte, R., & Dolci, R. (1991). Computing linguistic knowledge for text-to-speech systems with PROSO. In *Proc. EUROSPEECH'91*, Genova (pp. 1291–1294).

Delmonte, R., Mian, G. A., & Tisato, G. (1986). A grammatical component for a text-to-speech system. In *Proceedings ICASSP '86*, IEEE, Tokyo (Vol. 4, pp. 2407–2410).

Delmonte, R., Cristea, D., Petrea, M., Bacalu, C., & Stiffoni, F. (1996a). Modelli fonetici e prosodici per SLIM. In *Convegno GFS-AIA*, Roma (pp. 47–58).

Delmonte, R., Cacco, A., Romeo, L., Dan, M., Mangilli-Climpson, M., & Stiffoni, F. (1996b). SLIM—a model for automatic tutoring of language skills. In *Ed-Media 96*, AACE, Boston (pp. 326–333).

Delmonte, R., Petrea, M., & Bacalu, C. (1997). SLIM prosodic module for learning activities in a foreign language. In *Proc. ESCA, Eurospeech'97*, Rhodes (Vol. 2, pp. 669–672).

Eskénazi, M. (1999). Using automatic speech processing for foreign language pronunciation tutoring: some issues and a prototype. *Language Learning & Technology*, *2*(2), 62–76.

Grover, C., Fackrell, J., Vereecken, H., Martens, J.-P., & Van Coile, B. (1998). Designing prosodic databases for automatic modelling in 6 languages. In *Proceedings of ESCA/COCOSDA workshop on speech synthesis*, Australia (pp. 93–98).

Hiller, S., Rooney, E., Laver, J., & Jack, M. (1993). SPELL: an automated system for computer-aided pronunciation teaching. *Speech Communication*, *13*, 463–473.

Hurley, D. S. (1992). Issues in teaching pragmatics, prosody, and nonverbal communication. *Applied Linguistics*, *13*(3), 259–281.

Jilka, M. (2000). *The contribution of intonation to the perception of foreign accent*. Doctoral Dissertation, Arbeiten des Instituts für Maschinelle Sprachverarbeitung (AIMS) (Vol. 6(3)). University of Stuttgart.

Jilka, M. (2009). Ph.D. Dissertation. Available at http://www.ims.uni-stuttgart/phonetik/matthias/.

Jilka, M., & Möhler, G. (1998). Intonational foreign accent: speech technology and foreign language teaching. In *Proceedings of the ESCA workshop on speech technology in language learning*, Marholmen (pp. 115–118).

Kahn, D. (1976). *Syllable-based generalizations in English phonology*. MIT doctoral dissertation, distributed by IULC.

Kawai, G., & Hirose, K. (1997). A call system using speech recognition to train the pronunciation of Japanese long vowels, the mora nasal and mora obstruent. In *Proc. Eurospeech97* (Vol. 2, pp. 657–660).

Kelm, O. R. (1987). An acoustic study on the differences of contrastive emphasis between native and non-native Spanish speakers. *Hispania*, *70*, 627–633.

Kim, Y., Franco, H., & Neumeyer, L. (1997). Automatic pronunciation scoring of specific phone segments for language instruction. In *Proc. Eurospeech97* (Vol. 2, pp. 645–648).

Klatt, D. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, *82*, 737–797.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, *3*, 253–263.

Loveday, L. (1981). Pitch, politeness and sexual role: an exploratory investigation into the pitch correlates of English and Japanese politeness formulae. *Language and Speech*, *24*, 71–89.

Luthy, M. J. (1983). Nonnative speakers' perceptions of English "nonlexical" intonation signals. *Language Learning*, *33*(1), 19–36.

Meador, J., Ehsani, F., Egan, K., & Stokowski, S. (1998). An interactive dialog system for learning Japanese. In *Proc. STiLL '98*, op.cit. (pp. 65–69).

Medan, Y., Yair, E., & Chazan, D. (1991). *Super resolution pitch determination of speech signals*. New York: IEEE Press.

Pisoni, D. (1977). Identification and discrimination of the relative onset times of two component tones: implications for voicing perception in stops. *Journal of the Acoustic Society of America*, *61*, 1352–1361.

Price, P. (1998). How can speech technology replicate and complement good language teachers to help people learn language? In *Proc. STiLL '98*, op.cit. (pp. 103–106).

Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: a study based on speech resynthesis. *Journal of the Acoustic Society of America*, *105*(1), 512–521.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *26*(2), 145–171.

Roach, P. (2000). Studying rhythm and timing in English speech: scientific curiosity, or a classroom necessity?

Ronen, O., Neumeyer, L., & Franco, H. (1997). Automatic detection of mispronunciation for language instruction. In *Proc. Eurospeech97* (Vol. 2, pp. 649–652).

Rooney, E., Hiller, S., Laver, J., & Jack, M. (1992). Prosodic features for automated pronunciation improvement in the SPELL system. In *Proceedings of the International Conference on Spoken Language Processing*, Banff, Canada (pp. 413–416).

Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., & Van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, *41*(3–4), 439–487. Special Issue on Prosody and Conversation.

Umeda, N. (1977). Consonant duration in American English. *Journal of the Acoustical Society of America*, *61*, 846–858.

van Santen, J. (1997). Prosodic modeling in text-to-speech synthesis. In *Proc. Eurospeech97* (Vol. 1, pp. 19–28).

van Santen, J., Shih, C., Möbius, B., Tzoukermann, E., & Tanenblatt, M. (1997). Multi-lingual durational modeling. In *Proc. Eurospeech97* (Vol. 5, pp. 2651–2654).

van Son, R., & van Santen, J. (1997). Strong interaction between factors influencing consonant duration. In *Proc. Eurospeech97* (Vol. 1, pp. 319–322).

Willems, N. (1983). *English intonation from a Dutch point of view*. Doctoral Dissertation, University of Utrecht.