5

Exploring Speech Technologies for Language Learning

Rodolfo Delmonte
Università Ca' Foscari - Ca' Bembo,
Linguistic Computational Laboratory,
Italy

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 rhymth from the interleaving of stress, accent, and phonological rules;
- the generation of adequate intonational pattern for each utterance related to communicative functions;

As appears from above, for a student to communicate intelligibly and as close as possible to native-speaker's pronunciation, prosody is very important [2.]. 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.From our point of view it is much more important to stress the achievement of 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 about these matter in the following sections.

All prosodic questions related to "rhythm" will be discussed in the first section of this chapter. In [62.] 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 M., 1999). 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, F. and J. Mehler, 1999; Ramus F., et al., 1999).

1.1 Speech recognition and acoustic models

In all systems based on HMMs (Kawai G., K.Hirose, 1997; Ronen O. 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 Y., et al. 1997) the authors discuss the procedure followed to generate them: 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 inappropriate to cope with prosodic learning activities since theymay produce distorted results in a teaching environment. This may be due, first of all, to the fact that they 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 P., 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).

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 J., 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 P., 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, M., 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, J., & Franco, H., 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.

In Bagshaw et al. (1993) student's contours are compared 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 part of the Chapter, 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.

2. Section I: Prosodic tools for self-learning activities in the domain of rhythm

2.1 General problems related to Rhythm

In prosodic terms, Italian/Spanish and English are placed at the two opposite ends of a continuum where languages of the world are placed (Ramus, F. and J. Mehler, 1999; Ramus

F., et al. 1999). This is dependent on their overall phonological systems, which in turn are bound by the vocabulary of the 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.

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 P.M., 1980; Lehiste I., 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 P.M., 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 I., 1977). Differences between the two prosodic models of production are discussed at length in a following section.

2.1.1 Segmental vs. syllable-based modeling

Prosodic data suffer from a well-known problem of sparsity (Delmonte R., 1999). In order to reach a better understanding of this problem however, we would like to comment on data in the literature (van Son R., J. van Santen, 1997; Umeda N., 1977; van Santen J., 1997) basically related to English, apart from the latter, 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 W., S.Isard, 1991; Campbell W., 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-occuring 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 J., et al., 1997:321; Grover et al., 1998) 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 R., J. van Santen, 1997) with the highest number of factors, 40. Sparsity in prosodic data is then ultemately 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.

2.1.2 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 ibid., 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.

As Klaus Zechner, et al. (2009) comment, while speech scoring systems for linguistically simpler tasks such as reading or providing a short response have been in operation for some time (Bernstein J., 1998, 1999; H. Franco et al.), few attempts have been made to automatically score spontaneous, non-native speech where the term 'spontaneous' is referred to high entropy speech where a large-vocabulary continuous speech recognition (LVCSR) system needs to be used for recognizing speakers' utterances. ETS has, after several years of research (see K. Zechner, I. I. Bejar, and R. Hemat), designed and implemented an operational system, SpeechRater™, for scoring spontaneous non-native speech in the context of the TOEFL® iBT Practice Online (TPO) Speaking practice program. In the currently operational Version 1, however, the main area of feature coverage is fluency. The architecture of the SpeechRater system is a concatenation of these three components: a arge-vocabulary continuous speech recognition (LVCSR) system trained on non-native speech, a feature computation module, and a multiple regression scoring module.The interesting point is that the speech recognizer has been trained on "non-native" speech: in particular 30 hours of speech have been used and 100 hours for the language model training. The ASR then computes a total of 40 features which are appropriate for the task and their usage fits well with human raters' judgements.

C. Cucchiarini, S. Strik, and L. Boves (1997a) and C. Cucchiarini, S. Strik, and L. Boves (1997b) describe a system for Dutch pronunciation scoring along similar lines. Their feature set, however, is more extensive and contains, in addition to log likelihood Hidden Markov Model scores, various duration scores, and information on pauses, word stress, syllable structure, and intonation. In an evaluation, correlations between four human scores and five machine scores range from 0.67 to 0.92.In a more recent paper on an algorithm called the Goodness of Pronunciation, Sandra Kanters, Catia Cucchiarini, Helmer Strik compile an inventory of pronunciation errors frequently made by foreigners speaking Dutch. On the basis of this inventory they create artificial errors in a native development corpus, which in turn were used to optimize thresholds for the Goodness of Pronunciation (GOP) algorithm, which they use to give corrective feedback to users at the phoneme level.As the authors comment, in pronunciation learning corrective feedback is particularly required because very often learners are not awareof the pronunciation errors they make. Since exposure to the L2 and L2 output will not automatically guarantee this kind of awareness, corrective

feedback is required to make learners aware of their errors and stimulate them to attempt self-improvement (Havranek, G.).

2.2 State of the art in CALL tools: rhythm

Here below we list and briefly present those CALL systems that are located on the web which have tackled the problem of student's assessment in the field of word and subword syllable units using automatic visualization and correction methods. The comments and pictures are taken from the website of come from a publication of the author.

2.2.1 WebGrader

WebGraderTM (Neumeyer L., et al, 1998) is a pronunciation grading tool designed for practicing pronunciation in a second language. The system uses SRI's speech recognition and pronunciation scoring technologies. The application client was implemented by using the Java platform to facilitate deployment and updates of software and content over the World Wide Web. We present the overall system architecture, user-interface design, scoring algorithms, and a preliminary user study. WebGraderTM is organized in lessons. A lesson is a collection of related sentences organized by themes such as transportation or eating in a restaurant. Students can listen to natives saying the phrases, part of the phrases, or individual words. They can also record themselves and obtain pronunciation scores for the phrase and for individual words. Words that are hard to produce can be practiced by selecting the target word and obtaining scores for that particular word. The content can easily be updated, and additional lessons can be downloaded from a content server.

Fig. 1. WebGrader Visualization of graded pronunciation of French utterance

2.2.2 BetterAccent Tutor for English

BetterAccent Tutor (Komissarchik E., Julia Komissarchik, 2000a, 2000b) is designed for non-native speakers of English, who would like to speak clearly, effectively and be easily understood. Using advance unique patented speech analysis technology, BetterAccent Tutor

presents instant audio-visual feedback of users' pronunciation. In American English three components of speech that contribute the most to comprehensibility are *intonation, stress and rhythm*. BetterAccent Tutor analyzes intonation, stress and rhythm patterns of a user-recorded utterance and visualizes these patterns in an easy- to- understand manner. By pinpointing the exact mistakes, BetterAccent Tutor allows users to focus on the problems that are unique to their speech. It allows users to *record*and*playback* utterances. Analyzes and visualizes*intonation, intensity* and *rhythm* patterns of recorded utterances. Visualizes *the syllabic structure* of recorded utterances and *highlights the syllables* as they are played back. Allows users to *visually compare* the user's and native speaker's *intonation, intensity* and *rhythm* patterns. Contains an *extensive set of exercises* specially- designed for the BetterAccent Tutor. Includes *detailed explanations* of each exercise. Includes a *large collection of utterances by native speakers* to provide users with guidance and a yardstick for correct pronunciation. Works well as a course supplement or as an interactive pronunciation coach for students' independent study.

BetterAccent Tutor's purpose is to help students speak clearly and effectively and to be easily understood. We believe that there is no such thing as right or wrong pronunciation; not even two native speakers speak alike. But to be understood by native and non-native speakers, it is imperative for non-native speakers to match native speakers at certain key points. With visual feedback, the Tutor shows users' speech characteristics that are most important. As commented above, the three factors that have the biggest impact on intelligibility of speech are intonation, stress and rhythm. BetterAccent Tutor analyzes and visualizes intonation, stress and rhythm patterns of users' speech. By visualizing users' pronunciation, the Tutor allows users to focus on the problems that are unique to their speech. The Tutor is designed to give users the power to identify, understand and correct pronunciation errors. BetterAccent Tutor Comprehensive Curriculum includes: Word Stress; Simple Statements; Wh-Questions; General Questions; Repeated Questions; Alternative Questions; Tag Questions; Commands; Exclamations; Direct Address; Series of Items; Long Phrases; Tongue Twisters

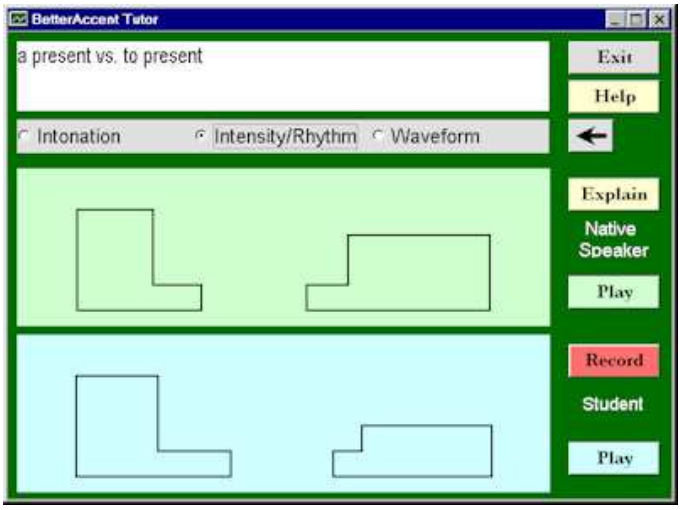

Fig. 2. BetterAccent Visualization of word stress example

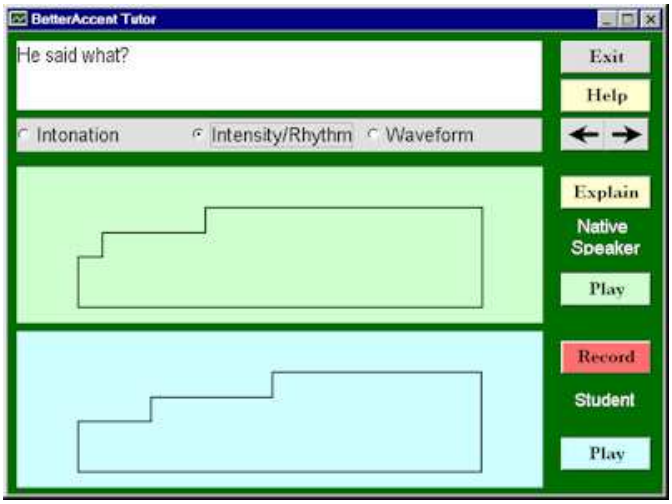

Fig. 3. BetterAccent Visualization of utterance example

2.2.3 Fluency

The FLUENCY (Eskénazi, M., 1999; Eskénazi M., et al. 2000) project has investigated the detection of changes in duration, amplitude, and pitch that can reliably detect where non-native speakers deviate from acceptable native values, independently of L1 and L2. Thus, if a learning system is applied to a new target language, its prosody detection algorithms do not have to be changed in any fundamental way. Since they are separate from one another, the three aspects of prosody can easily be sent to visual display mechanisms that show how to correctly produce pitch, duration, or amplitude changes as well as compare a native speaker's production to that of a non-native speaker.

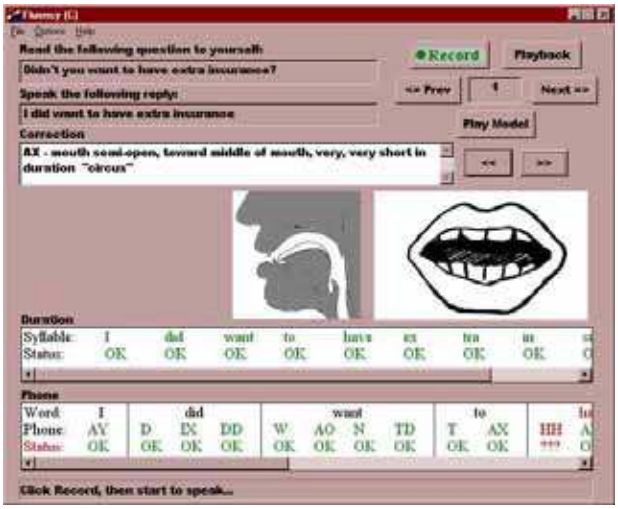

Fig. 4. FLUENCY Visualization of utterance example

2.2.4 Ordinate's PhonePass

Ordinate's patented PhonePass® (Bernstein J., 1998; Bernstein J., et al. 1998) testing system is based on of years of research in speech recognition, statistical modeling, linguistics, and testing theory. The technology uses a speech recognition system that is specifically designed to analyze speech components from native and non-native speakers of English. In addition to recognizing words, the system also locates and evaluates relevant segments, syllables, and phrases in speech. The PhonePass system then uses statistical modeling techniques to assess the spoken performance.Independent studies have shown that Ordinate's SET tests (Spoken English Tests), which are powered by the PhonePass testing system, are more objective and reliable in operation than today's best human-rated tests, including one-on-one oral proficiency interviews. Using criteria developed by expert linguists, the PhonePass testing system provides items and scores that have been validated with reference to human judgments of proficiency, fluency, and pronunciation.

The PhonePass testing system uses speech recognition technology that was built to handle the different rhythms and varied pronunciations used by native and non-native English speakers. The system generates scores based on the exact words used in the spoken responses, as well as the pace, fluency, and pronunciation of those words in phrases and sentences. In addition to recognizing the words uttered, the system also aligns the speech signal, i.e., it locates the part of signal containing relevant segments, syllables, and words.Base measures are then derived from the linguistic units (segments, syllables, words), based on statistical models of native speakers. The base measures are combined into four diagnostic sub-scores using advanced statistical modeling techniques. Two of the diagnostic sub-scores are based on the content of what is spoken, and two are based on the manner in which the responses are spoken. An Overall Score is calculated as a weighted combination of the diagnostic sub-scores.

For the SET-10 test, responses to four item tasks are currently used for automated scoring. These are: reading aloud, repeating sentences, building sentences, and giving short answers to questions. In scoring, there is exactly one correct word sequence expected for each response to the reading and repeat items. Expert judgment was used to define correct answers to the short-answer question and sentence-build items. Most of the short-answer and some of the sentence-build items have multiple answers that are accepted as correct. All short-answer questions were pre-tested on diverse samples of native and non-native speakers. All items retained in the item banks were answered correctly by at least 90% of the native sample.

2.2.5 SRI's EduSpeak

EduSpeak ® (Franco H., et al., 2000) is a speech recognition system that, through its Software Development Kit, enables developers of multimedia applications to incorporate continuous speaker-independent speech recognition into their applications.Developed in the Speech Technology and Research (STAR) Laboratory of the Information and Computing Sciences Division at SRI International, EduSpeak® is now available for licensing in the Language Education, Reading Development, and Corporate Training markets. Interactive English as Second Language (ESL) instructional CDs for elementary school children, using EduSpeak®'s unique pronunciation scoring technology

- Computer-aided collection and grading of spoken language in education and corporate settings
- Multimedia edutainment software with speech enhanced interactivity
- Language training courses for corporate travelers

Features & Benefits:

Speaker independence: No user training required. Continuous speech capability: No need for artificial pauses. State-of-the-art performance: High level of accuracy. Compact engine and models downloads: Fast application loading and internet. Multiple native speech models: Multiple language capability. Non-native speech models: Robustness to strong accents. Children's speech models: Increased accuracy for children. Pronunciation grading capability: Pronunciation feedback. Dynamically loadable vocabulary: Application flexibility. Arbitrary grammars: Increased flexibilty in task design. Dynamically loadable grammars task: Dynamic configuration of recognizer

Fig. 5. EduSpeak website advertisement

2.2.6 CMU Native Accent

NativeAccentTM(Eskénazi, M., 2007) is a pronunciation tutor using automatic speech recognition from the CMU. It has gone through a full-fledged assessment by real users in real situations, based on the customer's own criteria instead of more academic measures, and the variations in the customers' measures. Results in one study show that subjects who used NativeAccentTM did more than twice as well as the control group while both groups had human instruction.

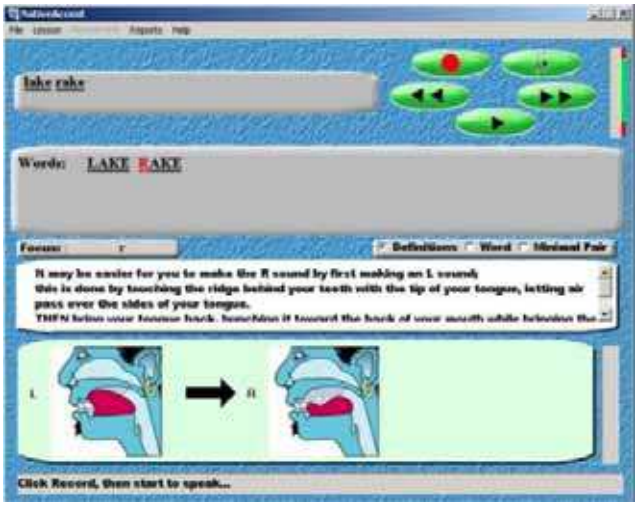

Fig. 6. Main screen of NativeAccent showing feedback

The system has been implemented for pronunciation error detection but also for a complete course of study, with leveled corrective feedback information, a curriculum, a student model, astrategy on how to proceed through the curriculum for differentlearners (fast and slow, for example) and a reporting mechanism forthe teacher (to follow individual and grouped student progress).

2.3 Self-aearning activities in the prosodic module: word stress and timing

We now present*SLIM* an interactive multimedia system for self-learning of foreign languages which 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™ [24].

2.3.1 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 and used in our system two types of models:

MODEL I: - Top-down Syllable-based Model for Syllable-Timed languages

It is a model in which durational structure for a phonological or an intonational phrase is specified first, and then the segmental duration of the grammatical units in the words 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 W., S.Isard, 1991; Campbell W., 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 for Stress-Timed languages

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 (1987:132) 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, D., 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.

2.3.2 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. First of all, 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 P., 1994; Roach P., 1982): 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.

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.

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. 7-8) 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.

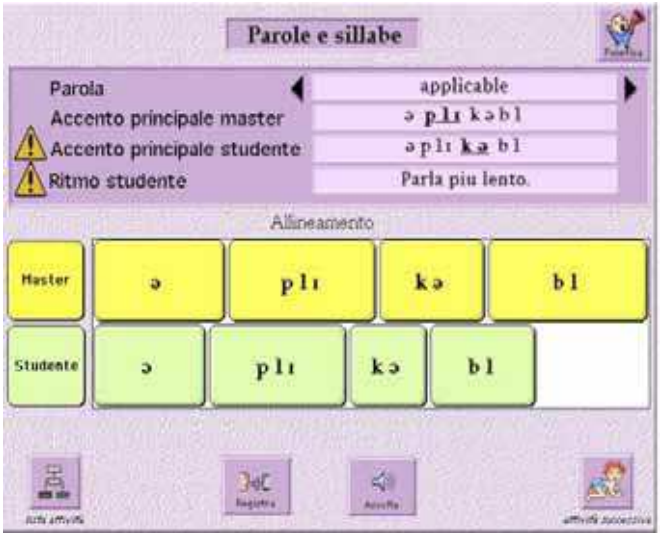

Fig. 7. Syllable Level Prosodic Activities
syll.jpg

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 R., Cristea D. et al. 1996; Delmonte R., et al. 1996). 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 (see Bannert 1987; Batliner et al., 1998).
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.

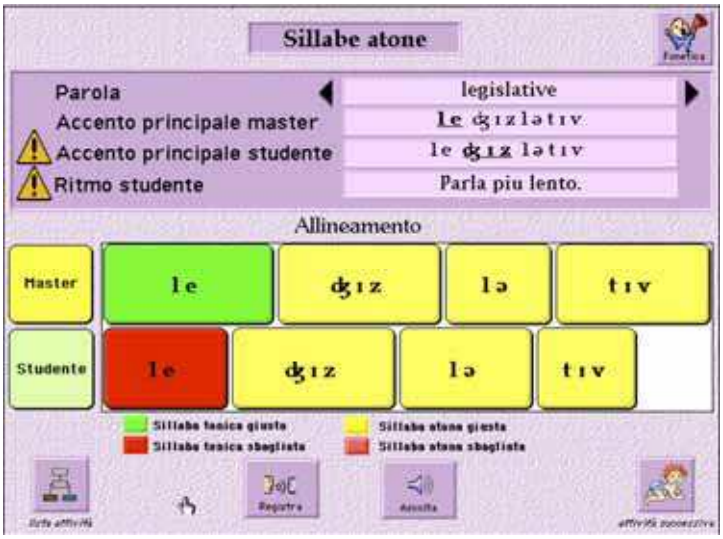

Fig. 8. Word Stress Prosodic Activities
stress.doc

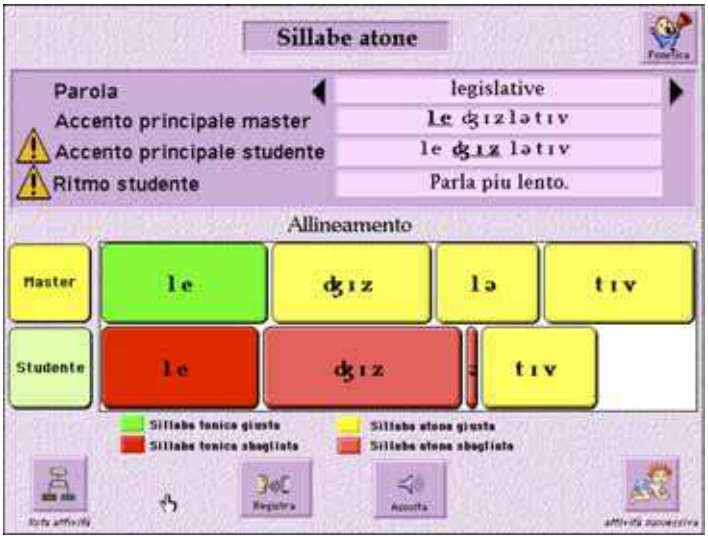

Fig. 9. Unstressed Syllables Prosodic Activities

2.3.3 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. As a result of the opposition between weak and strong syllables at word level (Eskénazi, M., 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 R., 2000c).

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
- Homorganic Stop Deletion with Glottalization
- Homorganic Liquid and Voiced Stop Deletion in Consonant Cluster
- Palatalization Rules affect all alveolar obstruents: /t, d, s, z/
- Palatalization of Alveolar Fricative
- Palatalization of Alveolar Nasal
- Palatalization of Alveolar Stop
- Degemination
- Velarization

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.

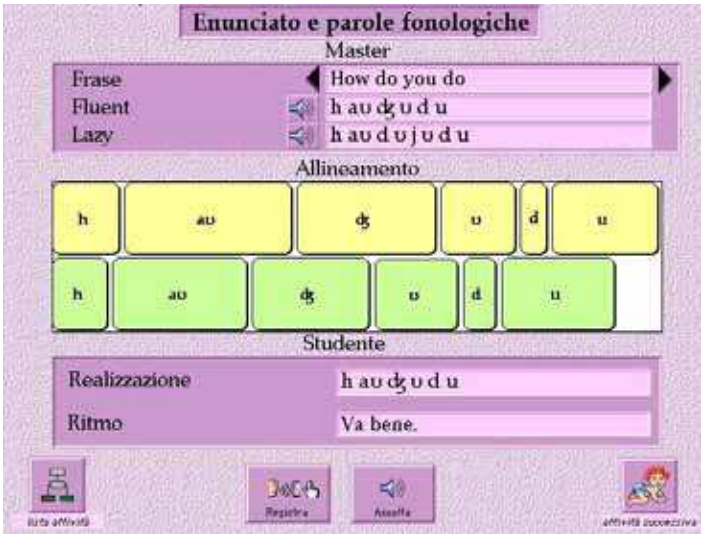

Fig. 10. Phonological Phrase Level Prosodic Activities

3. Section II: 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 M. &Möhler, G., 1998; Jilka M., 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:Chapter 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.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 [46], 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%.

	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%

Table 1. Tone inventories of American English and Italian

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 C., 1995).The deviations are summarized in an inventory of nine major differences in the

productions of the Dutch speakers (Willems, N., 1983). The listed deviations, which correspond to distinct instances of intonational foreign accent, include what Willems terms:

- the direction of the pitch movement (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 [13.] 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, D. S., 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 (ibid. :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" (ibid. :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" (ibid. :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.

3.2 Intonation practice and visualizatio: 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 E., 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 R. 1983, 1985, 1987, 1988) 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 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 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 FØ(t) = 1/T(t). However, by this method dissimmetries will eventually result: on rising portions of T(t), FØ(t) is normally compressed, while on falling portions of T(t), FØ(t) is stretched. As the displayed pitch is intended to put in evidence the rising portions of FØ(t) where accent appears, we prefer to simply compute a symmetric function of the T(t) slope instead of calculating the 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 [22.] Delmonte 2010. 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.

3.3 State of the art in prosodic CALL tools: intonation

As we did in the previous section, we report here below a select choice of commercial products and prototypes documented in the literature as being concerned with self-learning tools in the field of prosody, in particular tackling the problem of intonation. In some cases, the same product presented in the previous section reappears here, without repeating the comment, though.

3.3.1 Visi-Pitch visualization

One of the first examples of a program that displays visual pitch curves is a product from Kay Elemetrics called Visi-Pitch that has been available for a number of years for DOS-based personal computers (PCs). With Visi-Pitch, students are able to see both a native speaker's and their own pitch curve simultaneously. Students first speak a sentence into a microphone; their utterance is then digitized and pitch-tracked, and they can see a display of their pitch curve directly under a native speaker's pitch curve of the same sentence. Fig. 11 from Fischer (1986) shows the pitch contours of the French question Qu'est-ce qu'il fait? (What is he doing?) as spoken by a native speaker in the top half, and the same question produced by an American learner in the bottom half.

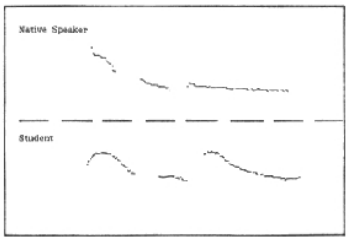

Fig. 11. Visi-Pitch Visualization of Pitch Curve

3.3.2 Auralog's TeLL me More

With the launch of the new version of **TeLL me More** in 2000 (see the website at URL7), Auralog allowed consumers to have easy access to resources that would enhance their language learning.As well as the scoring system, the software also allows the student to accurately visualise not only **pronunciation**, **but also intonation**. Two types of display mode (waveform and pitch curve) are provided. The student can display them at the same time, or individually. The **waveform** indicates the amplitude of the voice as a function of time **(the notion of energy)**. It represents the sound intensity of the voice and gives a view of the structure of the pronunciation. The **pitch curve** represents frequency variations in the voice. In tandem with the waveform, this curve enables students to make precise comparisons of his or her own intonation with that of the model **(high-pitched/deep)**. This unique display mode is an innovation developed by **Auralog**. **Auralog** is the only software publisher to offer applications which evaluate pronunciation and intonation of both complete sentences and words, and which allows them to be visualised.

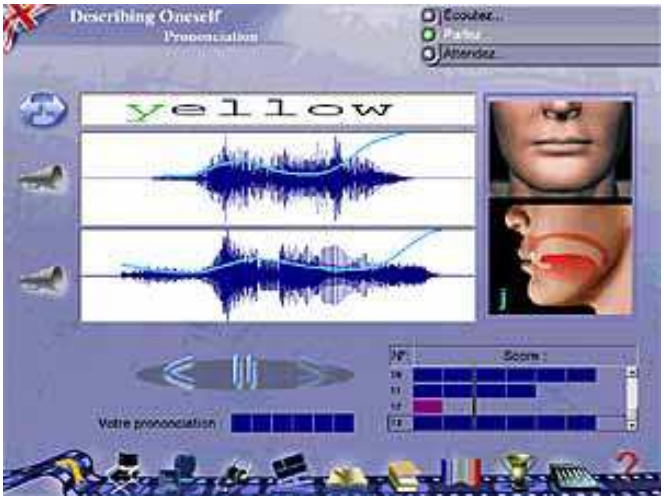

Fig. 12. Auralog's Visualization of Pitch Curve on top of waveform

3.3.3 BetterAccent tutor for english

We repropose here below the visualization of the minimal pair "a present"/"to present"(Fig. 13) where however pitch is used to mark differences between the two phrases. Notice that also the explanation which accompanies the exercise uses information related to intonational curve (Fig. 13.1). The presentation of an utterance is carried out along the same lines: "He said what" (Fig. 14; 14.1). The utterance is an echo question which requires a steep rising tone in coincidence with the wh- word which has been positioned in situ.

3.4 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.

Fig. 13. BetterAccent Visualization of stylized word stress example

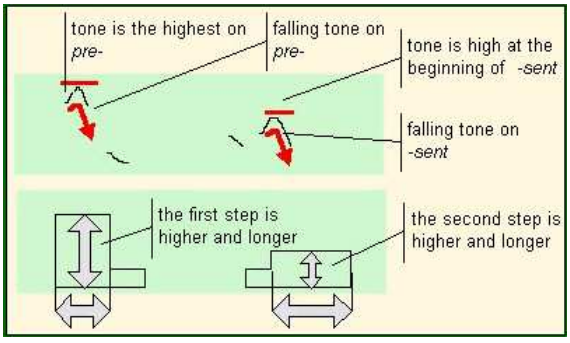

Fig. 13.1 BetterAccent evaluation of word stress example

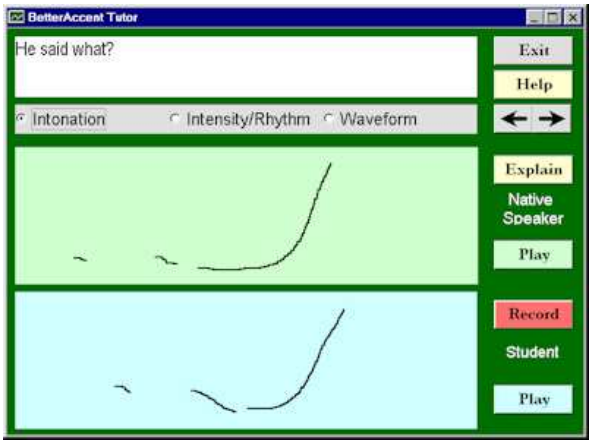

Fig. 14. BetterAccent Visualization of stylized utterance example

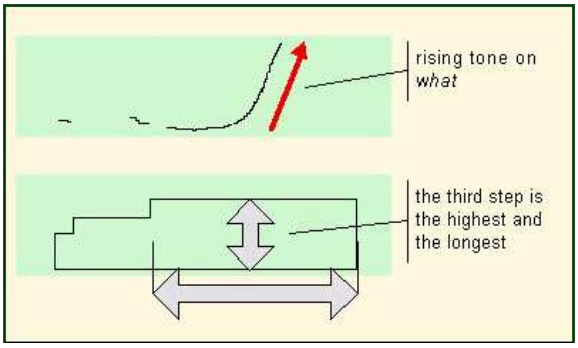

Fig. 14.1 BetterAccent evaluation of utterance example

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; Veridiction; 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:

- ASK; GIVE, OFFER, CONSENT; DESCRIBE; INFORM; SOCIALIZE; ASSERT, SAY, REPLY; EXPRESS EMOTIONS, MODALITIES; 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 R., Cristea D. et al. 1996; Delmonte R. et al., 1996). 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. 21 below, 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”.

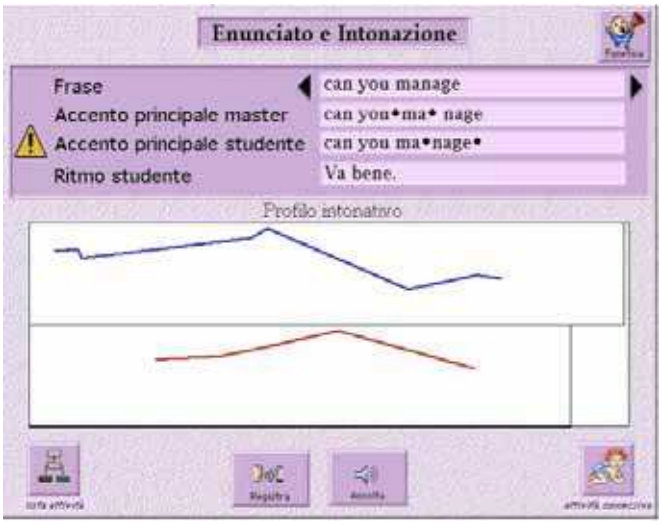

Fig. 15. Utterance Level Prosodic Activities: 1
fig15.jpg

In the second example, 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.

Fig. 16. Utterance Level Prosodic Activities: 2
fig16.doc

The third and final example 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 it correctly.

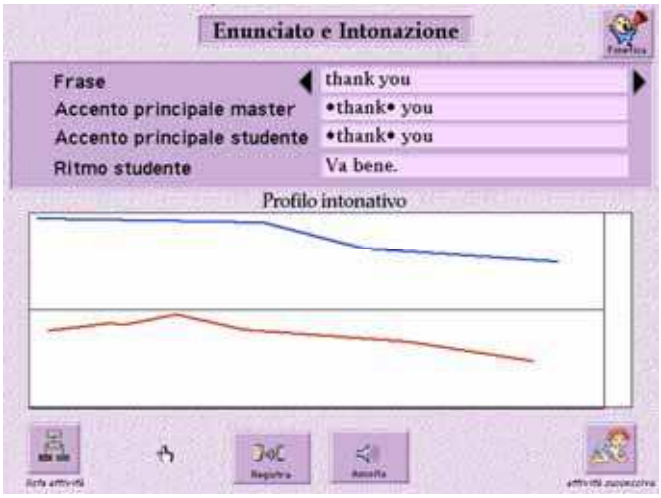

Fig. 17. Utterance Level Prosodic Activities: 3

4. Two systems with animated tutors

Eventually we present two systems that use animated agents or characters to provide feedback to the student and also to guide their activities. The first one has been produced at the Swedish Center for Speech Technology (CTT) at KTH and the second is the result of research work of more than one center, the CSLR.

The Swedish system is called VILLE and is a virtual tutor for Swedish language learners that uses knowledge of phonetics/phonology to help students learn pronunciation (see Engwall and Balter, 2007; Wik et al., 2009). As the authors comment, "the use of embodied conversational agents (ECAs) in computer assisted language learning (CALL) is seen as one way to address feedback issues. Ville guides, encourages and gives corrective feedback to students who wish to develop or improve their Swedish language skills. A first version of Ville was offered in the fall of 2008 to all foreign students at KTH who wanted to learn Swedish. The first version focused on helping students with vocabulary training, providing a model pronunciation of new words and drilling students in memorization exercises… The most serious errors with respect to intelligibility were found to be: lexical stress (insufficient stress marking, or stress on the wrong syllable), consonant deletion in a cluster before a stressed vowel, vowel insertion (epenthesis) in, or before a consonant cluster, vowel and consonant duration errors, vowel quality (difficulties with Swedish vowels not present in L1), and prosodic errors."

The animated tutor has been expanded in its abilities to offer feedback for addressing prosodic errors, in particular in the perception exercises. The result of the implementation of the new 8 Ville capabilities has been studied by means of a questionnaire and students have shown not to care too much to suggestions coming from Ville. In fact, only less skilled students seemed to take advantage of it.

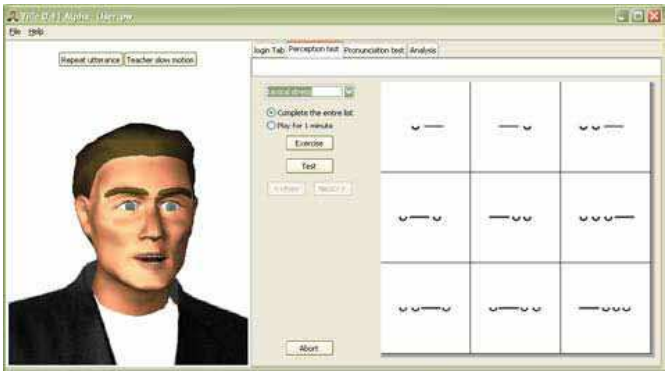

Fig. 18. VILLE animated tutor giving feedback on prosodic exercises

The second system we will comment on is ILT (Italian Literacy Tutor). ILT is a fully comprehensive system for language tutoring expressedly realized for children, the Colorado Literacy Tutor and its companion the Italian Literacy Tutor. Interactive Books, such as that illustrated in Fig. 25 below, incorporate leading edge speech recognition and generation technology, natural language processing tools, computer vision and character animation technologies which provide engaging and immersive learning experiences.The Italian Literacy Tutor is the Italian counterpart of the "Colorado Literacy Tutor" (CLT), a project developed at CSLR (Center for Spoken Language Research, Colorado University Boulder) for English and currently in use in American schools. As its English companion, the ILT integrates two sets of literacy tools, the first one based on speech and animation technology, and the second based on language comprehension technology. These programs are critically useful for children with special needs, in the following four populations: 1) students with reading disabilities, 2) foreign-speaking students with limited Italian/English language proficiency, 3) students with autism spectrum disorder, and 4) students with hearing impairments.

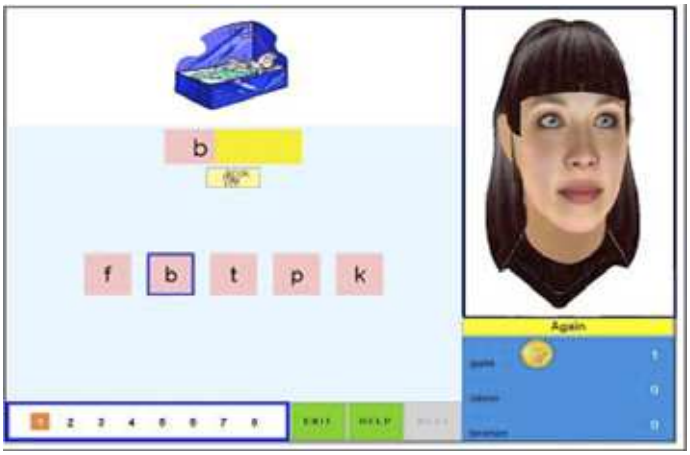

Fig. 19. A phonological and linguistic interactive exercise with an animated agent (LUCIA)

Tutors follow a default sequence from phonological awareness and decoding and encoding of simple consonant-vowel-consonant (CVC) words to more complex orthographic patterns into multisyllable words. Tutors are divided in fact into:

- Phonological Awareness (word, syllable, rhyme, phonemes), with all practice identifying, matching, blending, segmenting, and manipulating these units of spoken language. (see Fig. 19)

Alphabet and Letter-Sound Knowledge ; Reading of Regular Words, from CVC to complex words; Spelling of Regular Words; Reading Sight Words; Spelling Sight Words; Vocabulary

- Comprehension strategies, come into place whenever the children are not successful with the comprehension support and practice within the Books. Word reading, vocabulary, fluency, and comprehension are taught and practiced in Interactive Books, which also assess needs and assign Tutors based on those needs.

Reading Comprehension activites contemplates two types of exercises which requires NLP tools to be used: the first activity is Question/Answering on the contents of the text just read; the other activity, which is more complex to evaluate, is Summarization again of the text just read, which however is no longer visible to the student. In this case, the system activates a Summary evaluation tool which analyses the student text and compares it to a version of the chapter or long paragraph read in a semantic format called Discourse Model (see Delmonte R. 2004, 2007, 2009).

Fig. 20. An Interactive Book of the Italian version of the CLT with Animated Agent

4.1 Animated speech

Three dimensional animated computer characters associate production of natural or synthetic speech, to a wide variety of facial expressions and emotions, and natural body movements.

The characters' heads can be rotated and made semi or fully transparent, so children can watch how sounds are made to improve their own speech clarity and to detect errors. If a child has, for instance, left out the "l" in spelling "sled," the coach can direct him to watch

the tongue movement right after the /s/ in "sled" to discover the missing sound. Children can also compare video capture of their own mouths, in speaking a sound or a word, to the articulation of the coach's mouth. This encourages active and clear speech in the exercises, to improve both the clarity of the child's speech and the underlying precision of his phonological representations for words. They can narrate the book or engage the user in conversational interaction or dialogues to train and test comprehension.
In addition to producing accurate visible speech with associated facial expressions and gestures, animated characters can provide visual feedback to students during learning and conversational interaction. The character can also provide visual feedback and reinforcement, in the form of a head nod, smile, "thumbs up" or other gestures when the student provides correct answers; or look puzzled if the system does not recognize what the student is saying (Cosi et al., 2004a; Cosi et al.,2004b).

4.2 Conclusions and future directions

From what we have shown above, it is possible to make a number of concluding remarks and observations. From what we have shown, it is possible to safely draw a positive conclusion on the introduction of speech technologies in language learning tools. We have also shown that the use of speech technologies is by itself very fruitful in language learning environments but must be complemented by a whole lot of sophisticated tools which take care of pedagogical issues involved in any learning scenarios. In addition to that, speech technologies require empirical research to properly assess the adequateness of its architecture and curriculum for the intended domain and pedagogical objectives, which do not coincide directly with human directed teaching activities. It is still hard to think in terms of linguistic issues when providing feedback to students: as we saw, only the identity of the sounds or syllable or word involved in speech recognition can be addressed by feedback in currently available ASR. As to prosodic issues, only a few of the problems involved in prosodic learning can be detected and properly addressed when producing feedback. So there is still a long way to go to teach using CALL systems (Delmonte R. 2002b, 2003a).
The most challenging scenario is certainlyrepresented by the system at the end of the paper, where animated tutors are incorporated in a full-fledged system for literacy tutoring for children or for the teaching of pronunciation.Animated characters incorporate the technology of the future of language tutoring and constitute the test-bed for interactive activities where both speech synthesis and recognition are used and require implementation of modules for emotional speech. Here we would like to go back to the statements reported at the beginning of this chapter, by Sproat and van Santen, where the complexity of the task facing the use of speech technology is clearly outlined and covers the whole set of scientific domains associated to human language sciences. Animated tutors will certainly become a realilty in a near future, but a lot of work is still needed to address emotional issues both in the visual and in the speech domain.

5. References

[1] 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.
[2] Bagshaw, P. (1994). *Automatic Prosodic Analysis for Computer Aided Pronunciation Teaching*, Unpublished PhD Dissertation, Univ. of Edinburgh, UK.

[3] Bagshaw, P., Hiller, S., & Jack, M. (1993). Computer aided intonation teaching, In: *Proceedings of Eurospeech*, pp. 1003-1006.

[4] 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, 22.4.

[5] Batliner, A., R.Kompe, A.Kiessling, M.Mast, H.Niemann, NoethE. (1998). M - Syntax + Prosody: A syntactic-prosodic labelling scheme for large spontaneous speech databases, in Speech Communication, Vol. 25, No. 4, pp. 193-222.

[6] Bernstein, J. (1998). New uses for speech technology in language education, In: *Proceedings ISCA Workshop on Speech Technology in Language Learning, STiLL 98*, Marhollmen, pp. 173-176.

[7] Bernstein, J. (1999), "PhonePass testing: Structure and construct," Ordinate Corporation, Menlo Park, CA May 1999.

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

[9] Bertinetto, P.M. (1980). The Perception of Stress by Italian Speakers, *Journal of Phonetics*, Vol. 8, pp. 385-395.

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

[11] Campbell, W. (1993). Predicting Segmental Durations for Accomodation within a Syllable-Level Timing Framework, In: *Proc. Eurospeech '93*, pp. 1081-1085.

[12] Campbell, W., IsardS. (1991). Segment durations in a syllable frame, In: *Journal of Phonetics*, Vol. 19, pp. 37-47.

[13] Chun, D.M. (1998). "Signal Analysis Software For Teaching Discourse Intonation", LLTJ, Language Learning & Technology, Vol. 2, No. 1, pp. 61-77.

[14] Cosi, P., R. Delmonte, S. Biscetti, R. A. Cole, B. Pellom, van VurenS. (2004). ITALIAN LITERACY TUTOR: tools and technologies for individuals with cognitive disabilities, in R.Delmonte & S.Tonelli(eds), In: *Proc.INSTIL/ICALL2004*, Venezia, pp. 207-215.

[15] Cosi, P., R. Delmonte, S. Biscetti, ColeR. A. (2004). ITALIAN LITERACY TUTOR: un adattamento all'italiano del "Colorado Literacy Tutor", A. Andronico, P. Frignani, G. Poletti (a cura di), *Atti DIDAMATICA 2004*, Ferrara, pp. 249-253.

[16] Cucchiarini, C., H. Strik, and BovesL. (1997a). Using speech recognition technology to assess foreign speakers' pronunciation of Dutch, *In: Proc. Third international symposium on the acquisition of second language speech: NEW SOUNDS 97*, Klagenfurt, Austria.

[17] Cucchiarini, C., S. Strik, and BovesL. (1997b). Automatic evaluation of Dutch pronunciation by using speech recognition technology, In: Proc. *IEEE Automatic Speech Recognition and Understanding Workshop*, Santa Barbara, CA, 1997.

[18] Delmonte, R. (1983). A Phonological Processor for Italian, *Proceedings of the 2nd Conference of the European Chapter of ACL*, Pisa, pp. 26-34.

[19] Delmonte, R. (1985). Parsing Difficulties & Phonological Processing in Italian, *Proceedings of the 2nd Conference of the European Chapter of ACL*, Geneva, pp. 136-145.

[20] Delmonte, R. (1987). The Realization of Semantic Focus and Language Modeling, In:*Proceeding of the International Congress of Phonetic Sciences*, Tallinn (URSS), pp. 100-104.

[21] Delmonte, R. (1988). Focus and the Semantic Component, In:*Rivista di Grammatica Generativa*, pp. 81-121.

[22] Delmonte R., M. Petrea, C. Bacalu (1997). *SLIM Prosodic Module for Learning Activities in a Foreign Language*, Proc.ESCA, *Eurospeech'97*, Rhodes, Vol.2, pp.669-672.

[23] Delmonte, R. (1999). Prosodic Variability: from Syllables to Syntax through Phonology, in *Atti IX Convegno GFS-AIA*, Venezia, pp. 133-146.

[24] Delmonte, R. (2000). SLIM Prosodic Automatic Tools for Self-Learning Instruction, *Speech Communication*, Vol. 30, pp. 145-166.

[25] Delmonte R.(2002). Feedback generation and linguistic knowledge in 'SLIM' automatic tutor, *ReCALL*, Vol. 14, No. 1, Cambridge University Press, pp. 209-234.

[26] Delmonte R.(2003). Linguistic Knowledge and Reasoning for Error Diagnosis and Feedback Generation, In: Trude Heift and Mathias Schulze(eds.), Error Analysis and Error Correction in Computer-Assisted Language Learning, *CALICO Spring 2003 special issue*, *CALICO JOURNAL*, Southwest Texas State University, pp.513-532.

[27] Delmonte, R. (2004). Evaluating Students' Summaries with GETARUNS, *Proc.INSTIL/ICALL2004*, Unipress, Padova, pp. 91-98.

[28] Delmonte R., (2007), *Computational Linguistic Text Processing – Logical Form, Semantic Interpretation, Discourse Relations and Question Answering*, Nova Science Publishers, New York.

[29] Delmonte R., 2009. *Computational Linguistic Text Processing – Lexicon, Grammar, Parsing and Anaphora Resolution*, Nova Science Publishers, New York.

[30] Delmonte, R. (2010). Prosodic tools for language learning, *International Journal of Speech Technology*, Vol. 12, No. 4, pp.161–184.

[31] Delmonte, R., Andrea Cacco, Luisella Romeo, Monica Dan, Max Mangilli-Climpson, Stiffoni F.(1996). SLIM - A Model for Automatic Tutoring of Language Skills, *Ed-Media 96*, *AACE*, Boston, pp. 326-333.

[32] Delmonte, R., Dan Cristea, Mirela Petrea, Ciprian Bacalu, Stiffoni F. (1996). Modelli Fonetici e Prosodici per SLIM, *Atti 6° Convegno GFS-AIA*, Roma, pp. 47-58.

[33] Engwall O. and Balter O. (2007). Pronunciation feedback from real and virtual language teachers," *Computer Assisted Language Learning*, vol. 20, pp. 235-262.

[34] Eskénazi, M. (2009). An overview of spoken language technology for education, *JSC (Journal of Speech Communication)*, Vol. 51, No. 10.

[35] Eskénazi, M. (1999). "Using Automatic Speech Processing for Foreign Language Pronunciation Tutoring: Some Issues and a Prototype", *Language Learning & Technology*, Vol. 2, No. 2, pp. 62-76.

[36] Eskénazi, M., Yan Ke, Jordi Albornoz, Probst, K. (2000). Update on the Fluency Pronunciation Trainer, Dundee, Scotland, pp. 73-76.

[37] Eskénazi, Maxine, Angela Kennedy, Carlton Ketchum, Robert Olszewski, and Pelton, G.(2007). The NativeAccentTM pronunciation tutor: measuring success in the real world, in *Proc. SLaTE 2007*.

[38] Fischer, L. B. (1986). *The use of audio/visual aids in the teaching and learning of French.* Pine Brook, NJ: Kay Elemetrics Corporation.

[39] Franco, H., Victor Abrash, Kristin Precoda, Harry Bratt, Ramana Rao, and Butzberger, J. (2000). The SRI EduSpeak System: Recognition and Pronunciation Scoring for Language Learning, In: *Proc. InSTiLL*, Dundee, Scotland, pp.123-128.

[40] Grover, C., J. Fackrell, H. Vereecken, J.-P. Martens, Van Coile, B. (1998). Designing Prosodic Databases for Automatic Modelling in 6 Languages, *Proceedings of ESCA/COCOSDA Workshop on Speech Synthesis*, Australia, pp. 93-98.

[41] Havranek, G., "When is corrective feedback most likely to succeed?, (2002).*International Journal of Educational Research*, vol. 37, pp. 255-270.

[42] Hiller, S., E.Rooney, J.Laver and Jack M. (1993). SPELL: An automated system for computer-aided pronunciation teaching, *Speech Communication*, Vol. 13, pp. 463-473.

[43] Hurley, D. S. (1992). Issues in teaching pragmatics, prosody, and non-verbal communication. *Applied Linguistics*, Vol. 13, No. 3, pp. 259-281.

[44] 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

[45] Jilka, M., (2000). The Contribution of Intonation to the Perception of Foreign Accent. Doctoral Dissertation, Arbeiten des Instituts für Maschinelle Sprachverarbeitung (AIMS) Vol. 6, No. 3, University of Stuttgart.

[46] Jilka, M., PhD. Dissertation, available at http://www.ims.uni-stuttg art/phonetik/matthias/

[47] Kanters, Sandra, Catia Cucchiarini, and Strik, H. (2009). The Goodness of Pronunciation Algorithm: a Detailed Performance Study, In:*Proc. SLaTE 2009*.

[48] 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.

[49] Kelm, O. R. (1987). An acoustic study on the differences of contrastive emphasis between native and non-native Spanish speakers. *Hispania*, No. 70, pp. 627-633.

[50] Kim, Y., H.Franco, Neumeyer L. (1997). Automatic Pronunciation Scoring of Specific Phone Segments for Language Instruction, in *Proc. Eurospeech97*, Vol.2, pp. 645-648.

[51] Klatt, D. (1987). Review of text-to-speech conversion for English, *JA.S.A*, No. 82, pp. 737-797.

[52] Komissarchik, E., Komissarchik J. (2000a). Application of Knowledge-Based Speech Analysis to Suprasegmental Pronunciation Training, *AVIOS 2000 Proceedings*, San Jose, California.

[53] Komissarchik, E., Komissarchik J. (2000b). BetterAccent Tutor – Analysis and Visualization of Speech Prosody, In: *Proc. Speech Technology in Language Learning*, Dundee, Scotland.

[54] Lehiste, I. (1977). Isochrony reconsidered, In:*Journal of Phonetics*, No. 3, pp. 253-263.

[55] Loveday, L. (1981). Pitch, politeness and sexual role: An exploratory investigation into the pitch correlates of English and Japanese politeness formulae. *Language and Speech*, No. 24, pp. 71-89.

[56] Luthy, M. J. (1983). Nonnative speakers' perceptions of English "nonlexical" intonation signals. *Language Learning*, Vol. 33, No. 1, pp. 19-36.

[57] Meador, J., F.Ehsani, K.Egan, and Stokowski S. (1998). An Interactive Dialog System for Learning Japanese, In:*Proc. STiLL '98*, pp. 65-69.

[58] Neumeyer, L., Franco, H., Abrash, V., Julia, L., Ronen, O., Bratt, H., Bing, J., Digalakis, V., Rypa, M. (1998). "WebGrader™: A multilingual pronunciation practice tool", In:*Proceedings ESCA Workshop on Speech Technology in Language Learning (STiLL)' 98*, pp. 61-64.

[59] Price, P. (1998). How can Speech Technology Replicate and Complement Good Language Teachers to Help People Learn Language?, in *Proc. STiLL '98*, pp. 103-106.

[60] Ramus, F. and Mehler J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. In: *Journal of the Acoustic Society of America*, Vol. 105, No. 1, pp. 512-521.

[61] Ramus, F., Nespor M., Mehler J.(1999). Correlates of linguistic rhythm in the speech signal, *Cognition*, Vol. 26, No. 2, pp. 145-171.

[62] Roach, P., (1982). On the distinction between stress-timed and syllable-timed languages, In: Linguistic Controversies, D.Crystal (ed.), Edward Arnold, London, pp. 73-79.

[63] Ronen, O., L.Neumeyer, Franco H.(1997). Automatic Detection of Mispronunciation for Language Instruction, in Proc. Eurospeech97, Vol.2, pp. 649-652.

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

[65] Shriberg, E., R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, Van Ess-Dykema C. (1998). Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?, In:*LanguageandSpeech-Special Issue on Prosody and Conversation*, Vol. 41, No. 3-4,pp.439-487.

[66] Umeda, N. (1977). "Consonant Duration in American English", JASA , Vol. 61, pp. 846-58.

[67] URL7=http://www.tellmemore.com

[68] van Santen, J. (1997). Prosodic Modeling in Text-to-Speech Synthesis, In:*Proc. Eurospeech97*, Vol.1, pp. 19-28.

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

[70] van Son, R., van Santen J. (1997). Strong Interaction between Factors Influencing Consonant Duration, In:*Proc. Eurospeech97*, Vol.1, pp. 319-322.

[71] Wik, P., Hincks, R. &Hirschberg J. (2009). Responses to Ville: A virtual language teacher for Swedish, In: *Proc. SLaTE2009*.

[72] Willems, N. (1983). English Intonation from a Dutch Point of View. Doctoral Dissertation, University of Utrecht.

[73] Zechner, K., Derrick Higgins, Xiaoming Xi, (2009). SpeechRater™: A Construct-Driven Approach to Scoring Spontaneous Non-Native Speech, *Proc. SLaTE 2009*.

[74] Zechner, K., I.I. Bejar, and Hemat, R. (2007). Towards an understanding of the role of speech recognition in non-native speech assessment, Educational Testing Service, Princeton, 2007.

Speech and Language Technologies

Edited by Prof. Ivo Ipsic

ISBN 978-953-307-322-4
Hard cover, 344 pages
Publisher InTech
Published online 21, June, 2011
Published in print edition June, 2011

This book addresses state-of-the-art systems and achievements in various topics in the research field of speech and language technologies. Book chapters are organized in different sections covering diverse problems, which have to be solved in speech recognition and language understanding systems. In the first section machine translation systems based on large parallel corpora using rule-based and statistical-based translation methods are presented. The third chapter presents work on real time two way speech-to-speech translation systems. In the second section two papers explore the use of speech technologies in language learning. The third section presents a work on language modeling used for speech recognition. The chapters in section Text-to-speech systems and emotional speech describe corpus-based speech synthesis and highlight the importance of speech prosody in speech recognition. In the fifth section the problem of speaker diarization is addressed. The last section presents various topics in speech technology applications like audio-visual speech recognition and lip reading systems.