A Prosodic Module for Self-Learning Activities

Rodolfo Delmonte

Università Ca' Foscari - Ca' Garzoni-Moro
Laboratorio Linguistico Computazionale
San Marco, 3417 - 30124 Venezia (Italy)
delmont@unive.it
Website:project.cgm.unive.it

Abstract

We created an application specialized in prosodic tutoring, called the Prosodic Module(PM). The PM is composed of two different sets of Learning Activities, the first one dealing with prosodic problems at word syllabic level, the second one dealing with prosodic problems at phonological phrase and utterance level. The PM is able to detect significant deviations from a master's word/ phrase/ utterance and offers visual aids and a written diagnosis of the problem as well as indications on how to overcome and correct the error. This is achieved by means of a comparison between the two signals, the master and the student one; elements of comparison are constituted by the acoustic correlates of well-known prosodic elements such as intonational contour, sentence accent and word stress, duration at syllable, word and sentence level. We argue that the use of Automatic Speech Recognition as Teaching Aid should be targeted to narrowly focussed spoken exercises, for intermediate or higher level students, disallowing open-ended dialogues, in order to ensure consistency of evaluation. In addition, we support the conjoined use of ASR technology and prosodic tools to gauge Goodness of Pronunciation for linguistically consistent feedback.

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 [1]. The application we produced is able to detect significant deviation from a master's word/ phrase/ utterance production and offers visual aids and a written diagnosis of the problem as well as indications on how to overcome and correct the error.

The basic idea which lead to the development of the Prosodic Module was this: a master signal (pronounced with a high accuracy by a native speaker), eventually labelled with phonological information, is presented to the student learning that language. In turn the student, while working on oral activities, will record and listen to his voice, in order to compare it to the master's voice. In a self-learning scenario, he will be in need of appropriate and consistent feedback from the automatic tutor incorporated in the system, to be told whether his performance was good or not. This is accomplished by means of a comparison between the master and the student signals.

Elements of comparison are constituted by the acoustic correlates of well-known prosodic elements such as intonational contour, sentence and word accent, rhythm and duration at syllable, word and sentence level.

In order to tackle with the task at hand, special procedures have been implemented for silence detection, fricatives detection, ***FØ*** tracking, noise cutting, and for the detection of boundaries delimiting speech units with the aid of cepstral coefficients. The alignment procedure is based on the branch-and-bound method in which branches are generated using ***FØ*** traces already detected in a many-to-many correspondence type and "the best branch" is established heuristically by means of duration and energy variation criteria [2].

In learning to speak a foreign language like a native, it is precisely the prosodic features that generally prove the hardest to acquire [3]. Current ASR systems are insensitive to fundamental frequency, to amplitude and to details of vocal cord activity. They are also tolerant to differences in the durations and amplitude of speech sounds since they contribute little to determining the phonetic identity of speech sounds which is the primary source of information to recover their lexical correspondence.

Present-day speech recognizers are sensitive exclusively to phonetic information concerning the words spoken - their contents in terms of single phones. Even though cues to other types of information, such as the syntactic structure of the sentence, would contribute indirectly to knowledge of the word sequence, they are ignored because it is difficult to integrate them into the decision process effectively.

In addition, ASR systems are intended for native speakers. The systems developed are, for the most part, statistical pattern matching systems trained on a corpus of native speakers. In contrast, language learners are by definition non-native speakers. To work well for non-native speakers, the models should accommodate non-native speech, a need which

is satisfied by building a second model to be used for evaluation.

In order to be efficient it is necessary that the feedback be consistent, complete, concise and ready with respect to the exercise being performed. Learning through self-teaching must ensure the student with complete feedback to enable him to evaluate his own results and direct his own educational process.

Speech technology needs to be suited for use by non-native speakers, and suited for the teaching of a homogeneous variety of the language. However the challenge is how to get sufficient and consistent information for adequate evaluation of the student's performance: only in case such information is available, accurate and adequate feedback can be produced.

1.1 Organization of the paper

The paper consists of three main sections. In the first section we have the Introduction, of which this is a subsection, where we outline the main topics of the paper, which we list here below:

1) the importance of ASR as teaching aid for phonetic discrimination and intelligibility tasks as well as for other language production activities, limited though to closed dialogues - Sections 2 and 3;
2) the relevance of addressing prosody as the most appropriate linguistic level for building effective automatic language teaching aids - Section 4;
3) the importance of appropriate and consistent feedback and pronunciation scoring which is addressed in our case by putting forward a theoretically based contrastive analysis of the two languages in contact, L1 - Italian a syllable-timed/based language vs L2 - English a stress-timed/based language

2. Speech Recognition and Acoustic Models

In all systems based on HMMs[4,5] 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 [6] the authors discuss the procedure followed to generate them: as expected, they are trained on the native speakers database where dynamic time warping has applied in order to eliminate the dependency of scoring for each phone model on actual segment duration. Duration is then recovered for each phone from each frame measurements and normalized in order to compensate for rate of speech. Phonetic time alignment is then automatically generated for the student's speech. Working with bigram or trigram models, HMMs are unsuitable to encode duration seen that this parameter cannot be treated as an independent variable.

HMModels are inherently inadequate to cope with prosodic learning activities since statistical methods can only produce distorted results in a teaching environment. In general, 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.

We assume 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" for the in order to overcome the danger of teaching bad linguistic habits!

Another well-know problem is the quantity of training data to be used to account for both inter-speaker and intra-speaker variability. In particular, 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 a learning environment.

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 [8]. We use ASR only in a very controlled linguistic context in which the student has one of the following tasks:

1) 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.
2) repeat in a sequence "minimal pairs" presented on the screen and which the student may listen to previously;
3) another possibility is speaking 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;
4) do roleplay, i.e. intervene in a dialogue of a videoclip 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. The interaction with the system may be both in real time or in slow-down motion.

3. Prosodic Exercises

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. 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 interlanguage 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.

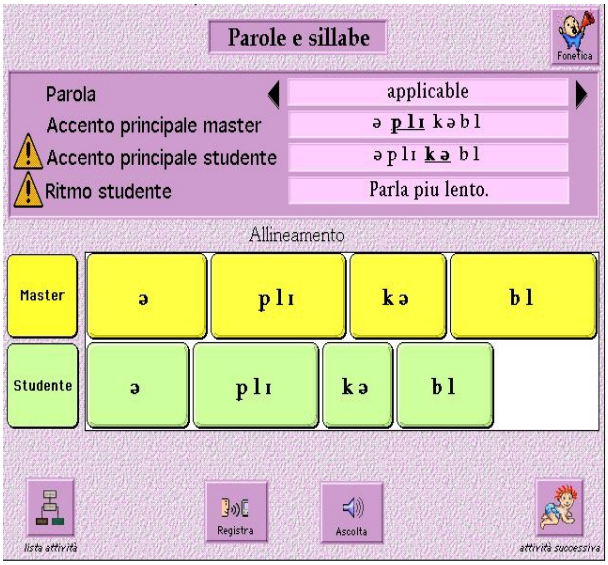
Contrastive studies have clearly pointed out the relevance of phonetic and prosodic exercises both for comprehension and perception. In particular, whereas the prosodic structure of Italian is usually regarded as belonging to the syllable-timed type of languages, that of English is assumed to belong to stress-timed type of languages[9,10]. 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 which tend to preserve the original phonetic features of interstress intervals [9].

Word-level exercises 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, which is not part of the inventory of phonemes and allophones of the source language, must be learned.

The main Activity Window for "Parole e Sillabe"/Words and Syllables is divided into three main sections: in the higher portion of the screen the student is presented with the orthographic and phonetic transcription 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 Interlanguage the student belongs to. We use six

Tab.1 Word Level Prosodic Activities

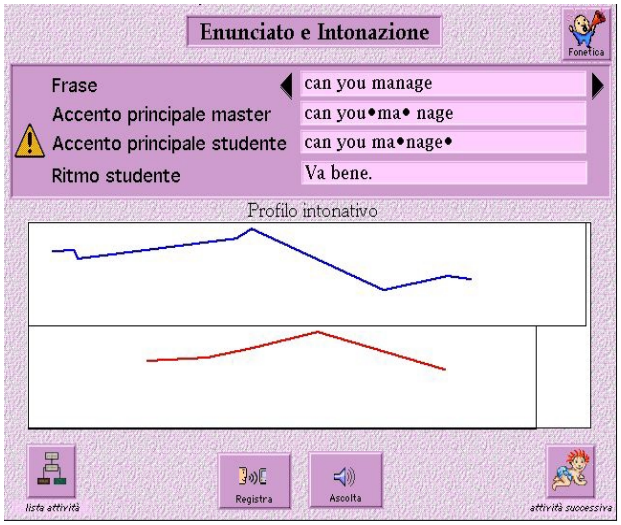


different levels [11,12]. In particular, the stressed syllable is highlighted between a pair of dots. The main central portion of the screen contains the buttons corresponding to each single syllable which the student may click on. The system then waits for the student performance which is dynamically 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. At the same time, in the section below the central one, two warnings will be activated in yellow and red, informing the student that the performance was wrong: prosodic information concerns the placement of word stress on a given syllable, as well as the overall duration. In case of error, the student practicing at word level will hear at first an unpleasant sound which is then followed by the visual indication of the error by means of a red blinking syllable button, the one in which he/she wrongly assigned word stress. This is followed by the rehearsal of the right syllable which always appears in green. A companion exercise takes care of the unstressed portion/s of the word: in this case, the student will focus on unstressed syllables and errors will be highlighted consequently in that/those portion/s of the word. Finally the bottom portion of the window contains buttons for listening and recording on the left, arrows for choosing a new item on the right; at the extreme right side a button to continue with a new Prosodic Activity, and at the extreme left side a button to quit Prosodic Activities.

In Utterance Level Prosodic Activities the student is presented with one of the utterances chosen from the course he is following. Rather than concentrating on types of intonation

contours in the two languages where performance-related differences might result in remarkable intraspeaker variations, we decided to adopt a different perspective. Our approach is basically communicative and focuses on a restricted number of communicative functions from the ones the student is practising in the course he is following (for a different approach see [13] on Japanese-English). Contrastive differences are thus related to pragmatic as well as performance factors.

Tab.2 Utterance Level Prosodic Activities



References

[1] Bagshaw, Paul, 1994. Automatic Prosodic Analysis for Computer Aided Pronunciation Teaching, Unpublished PhD Dissertation, Univ. of Edinburgh, UK.

[2] Kittler, J.; A.E. Lucas, A New Method for Dynamic Time Alignment of Speech Waveforms, 1990. in *Speech Recognition and Understanding, Recent Advances, Trends and Applications*, Pietro Laface, Renato de Mori (eds), NATO ASI Series, , vol. 75, 1990.

[3] Mennen I., 1998. Can language learners ever acquire the intonation of a second language*?*, in *Proc. STiLL '98, ESCA, Sweden*, 17-19.

[4] 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, 657-660.

[5] Ronen, O.; Neumeyer, L.; Franco, H., 1997. Automatic Detection of Mispronunciation for Language Instruction, in *Proc. Eurospeech97*, Vol.2, 649-652.

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

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

[8] Meador J.; Ehsani, F; Egan, K.; Stokowski, S., 1998. An Interactive Dialog System for Learning Japanese, in *Proc. STiLL '98, ESCA, Sweden*, 65-69.

[9] Bertinetto, Pier Marco, 1980. The Perception of Stress by Italian Speakers, *Journal of Phonetics*, 8, 385-395.

[10] Lehiste I., 1977. Isochrony reconsidered, in *Journal of Phonetics* 3:253-263.

[11] Delmonte R.; Cristea, Dan; Petrea, Mirela; Bacalu, Ciprian; Stiffoni, Francesco, 1996. Modelli Fonetici e Prosodici per SLIM, *Convegno GFS-AIA*, Roma, 47-58.

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

[13] Ueyama M., 1997. The Phonology and Phonetics of Second Language Intonation: The Case of "Japanese English, in *Proc ESCA'97*, Vol.5, 2411-2414.

[14] Hiller, S.; Rooney, E.; Laver J.; Jack, M, 1993. SPELL: An automated system for computer-aided pronunciation teaching, *Speech Communication*, 13:463-473.

Acknowledgements

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