SLIM

Prosodic Module for Learning Activities in a Foreign Language

Rodolfo Delmonte, Mirela Petrea, Ciprian Bacalu

Università Ca' Foscari - Ca' Garzoni-Moro
Laboratorio Linguistico Computazionale
San Marco, 3417 - 30124 Venezia (Italy)
Tel.:041-2578464/52/19 E-mail:delmont@unive.it
WebSite:byron.cgm.unive.it

published in Proc.ESCA , EuroSpeech'97, Rhodes, Vol.2, pagg.669-672

ABSTRACT

The Prosodic Module of SLIM has been created in order to solve problems related to segmental and suprasegmental features of spoken English in a courseware for computer-assisted foreign language learning called SLIM - an acronym for Multimedia Interactive Linguistic Software, developed at the University of Venice. It is composed of two different sets of Learning Activities, the first one dealing with phonetic and prosodic problems at word segmental level, the second one dealing with prosodic problems at utterance suprasegmental level. The main goal of Prosodic Activities is to ensure feed-back to the student intending to improve his/her pronunciation in a foreign language.
The programme works by comparing two signals, the master and the student ones, where the master has been previously edited by a human tutor inserting orthographic syllabic information at segmentation marks automatically computed by the underlying acoustic segmenter called Prosodics(see 1). When a student, after listening and evaluating the master signal tries to mimic the original utterance or word the system assigns a score and, if needed spots a mistake and indicates what it consists of. The elements of comparison are constituted by the acoustic correlates of prosodic features such as intonational contour, sentence accent and word stress, rhythm and duration at word and sentence level.

1. INTRODUCTION

The Prosodic Module has been created in order to deal with the problem of improving a student's performance both in the perception and production of prosodic aspects of spoken language activities in the courseware for computer-assisted foreign language learning called **SLIM**(see 2), developed at the University of Venice.
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 some feedback from the automatic tutor incorporated in the system, to be told whether his performance was good or not.
The main goal is to ensure feedback to the student, by means of a comparison between the two signals, the master and the student one; to assign a score to the student and, if the score is bad, to tell him where or what the mistake is.
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 word and sentence level.
In phonological terms, phonetic exercises based on suprasegmental elements are related to the following cues: the position of stress at word level; the position of main accent at sentence level; the overall intonational curve computed according to absolute parameters like for instance those referred to a generic subdivision of speakers into two types, males and females.
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. The alignment procedure is essentially based on the branch-and-bound method in which the branches are generated using the ***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. METHOD PROPOSED

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 lerning activities could be decomposed at least into the following aspects:
i. phonemic aspects, which include the correct pronunciation of single phonemes and the co-articulation of phonemes into higher phonological units such as syllables and feet where phonological rules might modify their phonemic and prosodic nature;
ii. prosodic aspects which include
-- the correct position of stress at word level;

-- the relative effect of stress at levels higher than syllable and word level in terms of compensation and vowel reduction in a language like English;
-- the correct position of sentence accent which interacts both with word-stress and intonational contour at utterance level;
-- the generation of the adequate rhymth from the interleaving of stress, accent, and overall phonological behaviour of remaining syllables;
-- the generation of adequate intonational pattern for each utterance related to communicative function and semantic content.

For a student to communicate intelligibly and as close as possible to native-speaker's pronunciation, prosody is the most important factor. The application we produced is able to detect significant deviation from a master's word/ phrase/ utterance production and offer a visual aid and a written diagnosis of the problem as well as indications on how to overcome and correct the mistake.

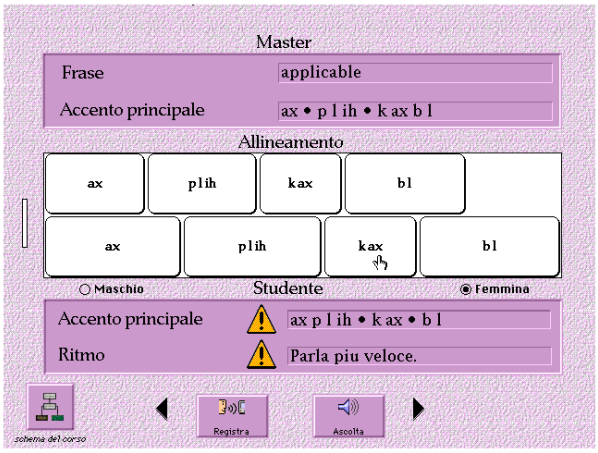
We assume together with other researchers that all prosodic aspects of speech are better described at syllable level [5]: so we perform a segmentation and associate it with the corresponding orthographic transcript. In turn this transcript is used by the programme to highlight portions of word or utterance to which the student should pay more attention, as well as to individuate some performance errors and generate appropriate diagnostic messages.

We assume that word stress is predominantly marked by variation in duration and energy at syllable level [6, 7]: we also acknowledge the fact that vowel quality at syllable nucleus is also subject to variation when stress intervenes, so we decided to compute spectral distances and use that information only when needed. 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.

Consider the problem of the correct position of stress at word level and the corresponding phenomena that affect the remaining unstressed syllables of words in English: prominence at word level is achieved by increased duration and intensity and/or is accompanied by variations in pitch and vowel quality (like for instance vowel reduction or even deletion, in presence of syllabifiable consonant like "n, d"). To detect this information, the system produces a detailed measurement of stressed and unstressed syllables at all acoustic-phonetic levels both in the master and the student signals. However, such measurements are known to be very hard to obtain in a consistent way(see 3, 4): 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 then issued.

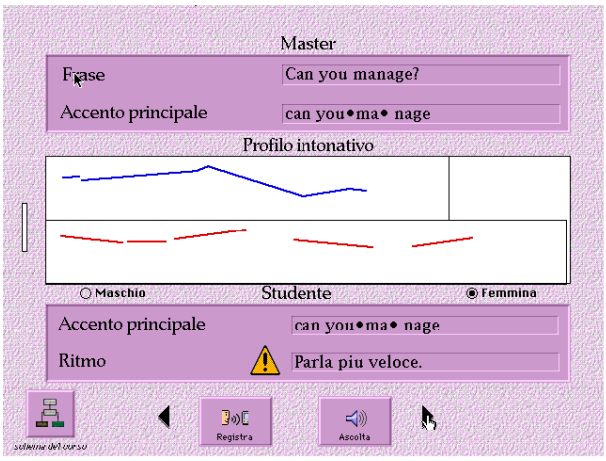
As mentioned above, the student is presented with a master version of an utterance or a word in the language he is currently practising and he is asked to repeat the linguistic item trying to produce a performance as close as possible to the original native speaker version. This is asked in order to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

Table 1. Word Level Prosodic Activities



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.

Table 1. Utterance Level Prosodic Activities



2. 1 The Supervisor

SLIM is an interactive multimedia system for self-learning of foreign languages and is currently addressed to Italian speakers. 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[8,9,10]. 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.
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 modyfing the presumed phonemic nature of their vocalic nucleus preserving the sound occurring in stressed position: so that 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. Thus, duration measurements should detect this fact in the acoustic phonetic representation. 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 [8].

3. PHONOLOGICAL MODEL

Differently from what happens in Speech Recognition Modules, in the application of acoustic and prosodic analysis for computer aided pronunciation teaching, the orthographic/phonemic transcript of an utterance is known in advance and may undergo a phase of pre-editing. In addition, a foreign language learner is asked to read a given word/sentence from the course material he has already practiced at some level of complexity. Reliable segmental information is therefore available. Following Bagshaw we describe here below the overall system for prosodic analysis. However, in our case, since the input data are known in advance, they can be used as background knowledge model onto which to match input acoustic data coming from the speech waveform. The segmentation and alignment process 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.

3.1 Preprocessing Phase

1. Each digitalized word, phonological phrase or sentence is automatically segmented and aligned with its phonetic transcript provided by the human tutor; with the following sequence of modules:

• Compute acoustic events for silence detection, silence detection, fricatives detection, noise elimination;
• Extract Cepstral coefficient from the input speech waveform sampled at 16 Mhz, every 5 ms for 30 ms frames;
• Follow a finite-state automaton for phone-like segmentation of speech in terms of phonological features;
• Match predicted phone with actual acoustic data
• Build syllable-like nuclei and apply further restrictions.

According to Umeda(1977) the effect of context in determining consonant duration can be schematized in the following conditions:

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

The system is organized into three different layers: phonemic level, syllabic level, word level. Phonemic level phonological rules are used to restrict acoustic analysis and to produce a limited set of probable phone candidates; this is matched against the input phonemic description and a first choice is made. The sequence of phones produces a syllable which is processed by the next layer where a set of other restrictions is taken into account and applied to produce the most adequate segmentation. Finally, the syllable is passed onto the higher layer in order to decide whether the syllabic segment corresponds to a complete word or not. In case it does, word level restrictions are applied and segmentation decisions are taken. Every such decision is followed by another run of segmentation processes at phoneme level until the input unit is completely segmented.

TABLE 1. Phonemes and Features
For each phoneme we use the following information:
- Computed Duration
- Voicing
- Frication Level (Three values)
- Energy Level (Three values)
- General Phonological Feature
Duration is computed as in most speech analysis/synthesis systems starting from a base duration and then applying a number of phonological rules take into account their right or left context. Classes of sounds that undergo similar treatment are the following:

• DIPHTHONGS AND VOWELS
A. STRESSED DIPHTHONGS
B. STRESSED SHORT VOWELS
C. STRESSED LONG VOWELS
D. UNSTRESSED VOWELS
Tables for consonant contextual conditions apply
when the single consonant is either preceded or followed by a vowel. Base values apply in word initial stressed position.
RULE 1. DiphVDur * 2 --> Vow/__Prepausal
RULE 2. DiphVDur*1.50--> Vow {ContentWord}
RULE 3. DiphVDur*0.50--> Vow {Unstressed}
RULE 4. DiphVDur * 2 --> Vow /__VoiceFricat
RULE 5. DiphVDur * 1.50 --> Vow {EndofWord}
RULE 6. DiphVDur * 1.25 --> Vow /__VoiceStop
RULE 7. DiphVDur*0.50--> Vow /__UnvoicStop

E. FRICATIVE CONSONANTS
RULE 8. FCons* 1.50 --> Cons {EndofWord}
RULE 9. FCons*0.50-->Cons{/__Obstruent, Unstressed}
RULE 10.FCons*1.25-->Cons+V {StressSyllable}

F. STOP CONSONANTS
RULE 11.SCons*1.50--> Cons + V {Initial, Medial}
RULE12.SCons*0.50-->Cons{WordFinal, MorphemeFinal}
RULE 13.SCons*1.25 -->Cons {Preceded by Nasal}

G. SONORANT CONSONANTS
RULE 14.NCons * 1.50 --> Cons + V {Init, Med}
RULE 15.NCons * 2 --> Cons {WordFin, MorphFin}
RULE 16.NCons*1.50-->Cons {+ VoiceObstr | Sonor}
RULE 17.NCons * 1.25 --> Cons {= VoiceObstr}

3.2 Phonological Rules at Word Boundaries

1. The process of homorganic stop deletion is activated whenever a stop is preceded by a nasal or a liquid with the same place of articulation and is followed by another consonant
• In front of voiced/unvoiced fricative
- you want some chocolate / juwonsem /
- and this is my colleague / anthIs /
• Homorganic Stop Deletion with Glottalization
what can I do for you / wa?ken /
• Homorganic Liquid and Voiced Stop Deletion in Consonant Cluster
you should be more careful / jushubbi /

2. Palatalization Rules affect all alveolar obstruents: /t, d, s, z/
• Palatalization of Alveolar Fricative
can I use your phone / kenajushor /
• Palatalization of Alveolar Nasal
may I join you? / meiaidg'oinju /
• Palatalization of Alveolar Stop
nice to meet you / mitchju /

3. Degemination
just take a seat / jastak /

4. CONCLUSIONS

We presented a fully implemented Module for Prosodic Activities integrated in a system for self-learning of foreign languages called SLIM. The system is currently undergoing its first evaluation phase with students of English for Economics, in self-access modality. After a period of six months it should be turned into a full-fledged product. More information is needed on the efficiency and feasibility of computer-based self-instruction in order to be able to assess its impact in a real University course.

5. REFERENCES

[1] Delmonte R., Dan Cristea, Mirela Petrea, Ciprian Bacalu, Francesco Stiffoni, *Modelli Fonetici e Prosodici per SLIM*, Convegno GFS-AIA, Roma, 47-58.
[2] Delmonte R., Andrea Cacco, Luisella Romeo, Monica Dan, Max Mangilli-Climpson, Francesco Stiffoni, SLIM - A Model for Automatic Tutoring of Language Skills, Ed-Media 96, AACE, Boston, 326-333.
[3] J. Kittler, A.E. Lucas *"A New Method for Dynamic Time Alignment of Speech Waveforms"*, in Speech Recognition and Understanding, Recent Advances, Trends and Applications, Pietro Laface, Renato de Mori(eds), NATO ASI Series, , vol. 75, 1990.
[4] Paul Bagshaw, *"Automatic Prosodic Analysis for Computer Aided Pronunciation Teaching"*, Unpublished PhD Dissertation, Univ. of Edinburgh, UK, 1994.
[5] Alex Waibel, "*Recognition of Lexical Stress in a Continuous Speech Understanding System - A Pattern Recognition Approach*", in Proc. ICASSP '86, Tokio, IEEE, 2287-2290.
[6] Domokos Vékás, Piermarco Bertinetto, "*Controllo vs. compensazione: sui due tipi di isocronia*", in E.Magno Caldognetto e P.Benincà(a cura di*), L'interfaccia tra fonologia e fonetica*, Padova, 1991.
[7] Pier Marco Bertinetto, *Strutture prosodiche dell'italiano*, Accademia della Crusca, Firenze, 1981.
[8] Pier Marco Bertinetto, "*The Perception of Stress by Italian Speakers*", Journal of Phonetics, 8, 1980, 385-395.
[9] Cinzia Avesani, *"Indici prosodici e segmentazione del segnale vocale nel riconoscimento del parlato continuo"*, Elaborazione dell'informazione vocale, FUB 1991-1992.
[10] P.C.Bagshaw, S.M.Hiller, M.A.Jack(1993), *"Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching"*, Proc.Eurospeech93, 1003-1006, Berlin,
[11] Y.Medan, E. Yair, & D.Chazan(1991), *"Super Resolution Pitch Determination of Speech Signals"*, IEEE Trans.Signal Processing, ASSP-39(1):40-48.
[12] N.Umeda(1977), *"Consonant Duration in American English"*, JASA , 61, 846-58.