ISCA Archive
http://www.isca-speech.org/archive

Speech and Language Technology in Education (SLaTE2011)
Venice, Italy
August 24-26, 2011

Using Speech Synthesis to Simulate an Interlanguage and Learn the Italian Lexicon

Rodolfo Delmonte, Francesco Stiffoni

Department of Language Studies
Ca' Foscari University of Venice
delmont@unive.it

Abstract

We present work carried out to simulate an interlanguage for English speakers learning Italian with a speech synthesizer made available by Apple. The lexicon is made of the most frequent 30,000 word forms of Italian as extracted from corpora and other similar Frequency Lists. The synthesizer has been filtered by a program in C-language that addresses the internal system of rules running on Apple's MacinTalk speech synthesizer that allows to modify phones and rhythm. By applying different sets of rules it is possible to simulate knowledge of the phonotactics, phonology and prosody in such a way as to mimic the condition of different stages of an English learner of Italian. Correct pronunciation is produced by an adequately modified Apple's synthesis for Mexican Spanish.

1. Introduction

Besides spoken and written text comprehension and sound-grapheme correspondences, an important and related goal of second language learning is acquiring the lexicon of the language. Indeed, lexical learning is the first challenge a second language student has to face [1]. A range of interesting problems might be helped by TTS:

- Offering a first-level, linguistically appropriate exposure to the target language by presenting its sound system in context;
- Allowing the student to couple meaning, sound, and sometimes image to the words to be learned, thus making the task simpler;
- Providing an easy authoring facility that allows teachers to increase the size of lexicon by adding new words to the database;
- Allowing students to adapt to varying speaking rate, voice type, and voice quality (difficult to obtain with real life recording);
- Exposing the learner to interlanguage and prompting comparison of word pronunciations according to proficiency level.

The sounds of a language are organized in a phonological system, which children gradually induce upon exposure to their mother tongue. This gradual induction does not usually happen in second language learning. TTS in a CALL lexicon application may help to reduce the gap with native speakers because the sounds of the language are spoken in the context of the lexicon of that language.
TTS in a CALL lexicon application may be helpful simply because a rich, multisensory learning environment is more effective than a lean, one-dimensional one. In particular, we assume that it is both more efficient and more natural to learn words of a second language by associating sound to orthographic image. It would be even better to couple the sound with a still image representing the concept expressed by the word, as usually happens with children.

2. Using TTS to simulate interlanguage

Lexical learning is regarded one of the most interesting applications of TTS because it has the possibility to simulate various stages of interlanguage [2]. It can be assumed that the learner of a second or foreign language (L2) will at first try to adapt the phonological system he/she has already mastered – primarily, the mother tongue (L1) but also any other system already acquired. Interlanguage stages will then vary as follows: from the level of a *full beginner*, in which the learner uses no phonological rules of L2 to translate from graphemes to phonemes in a pronunciation task; to *false beginner*, when the learner knows such rules but has not mastered all the phonemes of the target system; to *intermediate* level, in which the learner has mastered the phonemes of L2 but not all the prosodic (rhythm and intonation) rules; to *advanced* or quasi-native speaker level, where full phonetic and prosodic knowledge is applied [3,4].

To study these phenomena, we experimented by modifying a TTS system for L2 introducing rules of phonetic and prosodic realization that mimic the interlanguage of the L2 learner who speaks a particular L1 - English. Treating Italian as the L2 for an English speaker who wants to learn Italian, we implemented under HyperCard the Spoken Italian Word List (SIWL), a list of 30,000 Italian words with their phonemic and prosodic transcriptions, lemmata, and morpho-syntactic information [5,6]. We then applied these rules to a TTS for American English using SIWL.

The idea to use American English TTS to mimic the interlanguage stages of an American speaker learning Italian could similarly, be applied to mimic a Spanish speaker learning English, in which case we could have taken a lexical database of English with full linguistic annotation – this is the case for the database called SLIM [7,8] - and applied a TTS for Spanish American.
The TTS application implemented by Apple may be fed a pure orthographic file or a phonetically transcribed version of the orthographic file. In the latter case, grapheme-to-phoneme rules as well as prosodic rules may first have to be applied to generate an adequate format for the synthesizer. Table 1 shows the flow of information for a typical TTS system. It can be seen that both phonetic and prosodic knowledge are required to provide sufficient information to interact directly with the synthesizer. To this end, the Italian words in the SIWL were converted from graphemes to phonemes and the position of word-stress was marked in the input string to allow for appropriate internal prosody to apply. This was done by means of a computer program started in the '80s to provide synthesis of spoken Italian [6] – more on this below.

SIWL contains word forms which have more than two graphemes and does not contain any foreign word nor

proper names. Lemmatization of word forms has extended the initial list by an additional 16706 word forms with different lemma. We also isolated words allowing more than one type of pronunciation – either by a different stress position, or because they require a different stressed vowel for open-closed e/o alternations. These homograph word forms amount to 812 unique word forms which become 1652 different phonetic words, and 3123 different lemmatized words.

The procedure for mimicking levels of interlanguage is to manipulate the flow of information typified in Table 1. The interface for doing this is shown in Figure 1, from the synthetic interlanguage application based on the SIWL. The FIND function at the upper right of the screen allows the user to type in a word and see it displayed with related linguistic information. Alternatively, the user may move to and from the current word by pressing on the arrows. As seen at the lower right side of the screen, each word may be synthetically pronounced at three levels of simulated proficiency:

Level 1: NO RULES
Level 2: NO PROSODIC RULES
Level 3: APPLY ALL RULES

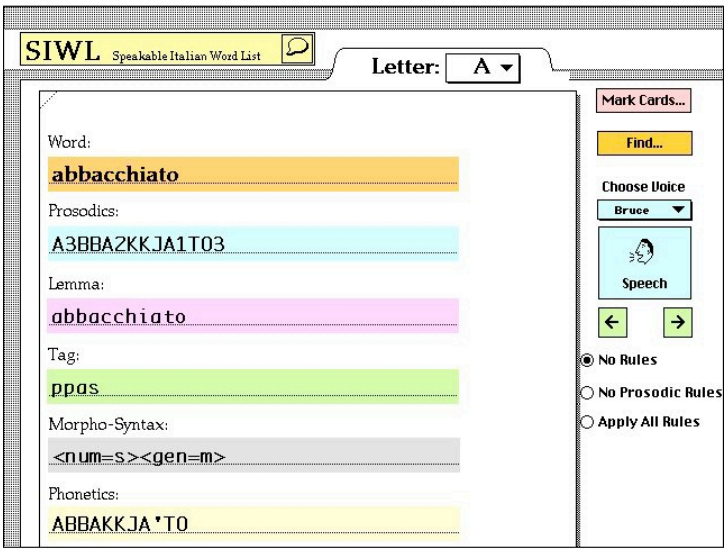

Figure 1: *Activity Window for Spoken Italian Word List Interlanguage Application*

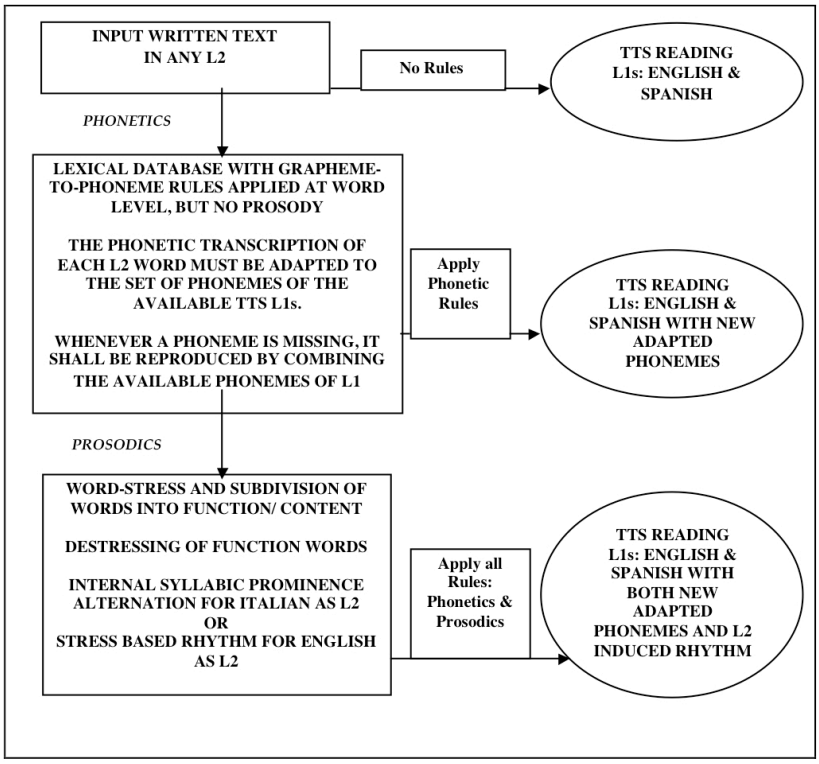

Table 1: *TTS Interaction with Linguistic Rules for L2 to induce Interlanguage Effects*

We can formulate a number of observations on the results obtained applying this procedure:
At *Level 1* the synthesized voice uses its own grapheme-to-phoneme rules and associated set of phonemes to produce a reading of the input word(s). The impression one gets on hearing the TTS is that of a native speaker of American English reading aloud the word list with no training on problematic words. The level of interlanguage competence in Italian is comparable to that of a false beginner. In particular, one notices that all L2 phonemes that coincide with L1 phonemes are pronounced correctly. But whenever a phoneme is lacking in the L1 inventory or a phonological rule is not present in L1, the grapheme-to-phoneme rules convert the graphemes into a wrong phonetic equivalent. This results many times in a non-comprehensible reading because of the distance between the Italian and English phonological systems. From a pedagogical point of view, this observation confirms the need for learners to study phonetics and prosodics of the L2.

At *Level 2* the TTS is activated by a C-language program that filters the input word and assigns it a grapheme-to-phoneme transcription with the appropriate internal phonetic symbols. In this way, the synthesized voice is endowed with the closest possible approximation to the phonetic system of Italian. The reading is now fully intelligible and in some cases also prosodically close to the Italian realization. The impression is of an American who is at an intermediate level of learning Italian but has still not mastered the prosody of the language. One hears typical errors at the word-stress level, where stress is placed on the most predictable position – such as the penultimate syllable – when it should be assigned elsewhere. In addition, depending on the L1, one hears misapplied vowel reductions on unstressed syllables; for example, in case the L1 (TTS speaking) is English, one hears the unstressed syllables in Italian words like *rivendicano* pronounced without reduced vowels.

Finally, at *Level 3*, all rules are applied, both phonetic and prosodic. The latter add an Italian-like syllable-based rhythm to the phonetic reading that markedly improves the quality of the output. The auditory impression one gets, in many cases, is of an advanced student of Italian (a native speaker of American English) reading with native-like pronunciation. The student's detectable accent lies mainly in vowel quality. Table 2. gives the list of symbols used to represent SIWL words input to the synthesizer, together with the Italian phonemes that correspond to each symbol. In addition, to produce valid prosody, each syllable to be synthesized is marked by a duration and amplitude index that captures the alternation of stressed/unstressed syllables at intraword level [10]. We used nine different prosodic markers in the SWIL (Table 3.), which also allow us to differentiate syllable contexts. Thus, *rivendicano (claim_3rd_pers_plur)* would be represented as RI3V&6NDI3KA5NO3.

Eventually there's another possibility, i.e. listening to the Italian-like pronunciation of the current word by switching the synthetic voice to Carlos, which is Apple's Mexican-Spanish TTS module. In order to let the Spanish voice speak Italian, we introduced some simple conversion rules at the level of phonetics in order to have the synthesizer produce the missing phonemes. We also use prosodic transcriptions to reproduce rhythm and word-stress, as in the English counterpart.

B --> /b/	**A --> /a/**
C --> /tʃ/	**E --> /é/**
K --> /k/	**& --> /è/**
D --> /d/	**O --> /o/**
F --> /f/	**@ --> /ò/**
% --> / ʤ/	**I --> /i/**
G --> /g/	**U --> /u/**
< --> / ʎ/	**> --> / ɲ/**
P --> /p/	**M --> /m/**
S --> /s/	**N --> /n/**
X --> /z/	**R --> /r/**
T --> /t/	**J --> /j/**
V --> /v/	**L --> /l/**
Z --> /ts/	**W --> /w/**
$ --> /dz/	

Table 2: *Grapheme-to-phoneme transcription and the corresponding Italian phonemes (within slashes)*

1 --> primary stress in open syllable or in syllable closed by /r/
2 --> secondary stress in open syllable
3 --> unstressed open or final syllable
4 --> semivowel
5 --> unstressed syllable in postonic position may also alternate with two unstressed syllables
6 --> primary stress in syllable closed by sonorants - /r/ excluded
7 --> primary stress in closed syllable and in truncated words
8 --> secondary stress in closed syllable
9 --> unstressed closed syllable

Table 3: *Prosodic markers introduced in the phonemic transcription of Italian words*

3. Prosodic rules for speech synthesis

Prosodic transcription allows us to keep under control the distribution of word-stress, unstressed vowels, secondary stressed vowels and reduced ones. We computed stress patterns for the whole of the SIWL by means of our application for Italian called Proso [11], and found more than 300 different prosodic patterns. Alternating all these different patterns endows words with a variety of associated rhythmic units, which operated in particular in long words to give each word the typical Italian rhythm [12]. Consider for instance some of the patters associated with long words – from 8 to 11 syllable words:

octosyllables
2 3 2 3 2 3 1 3#
3 2 3 2 3 3 1 3#
enneasyllables
2 3 2 3 3 2 3 1 3#
3 2 3 2 3 2 3 1 3#
decasyllables
2 4 3 2 4 3 2 3 1 3#
3 4 2 4 3 3 2 3 1 3#
endecasyllables
2 3 2 3 2 3 2 3 3 6 3#

And consider some words:
2 4 3 3 6 3# (neopresidente, coesisteranno !)
2 4 3 3 7 3# (ideologismi, neoliberiste, autorimessa)
2 4 3 4 1 3 3# (autoveicolo, autoveicoli !)
2 4 4 3 3 1 3# (autoaccensioni, autoadesivi !)
4 2 3 3 1 3# (coagulazione, aerospaziali, preoccupazione)
4 3 2 3 1 3# (riaddormentai, aerocisterna)
2 3 3 2 3 7 # (operatività, originalità)
2 3 4 3 1 3 3# (microelettronici)
2 4 3 2 3 1 3# (videogiocatore, antiamericani, visualizzazione)
2 4 3 4 3 1 3 3# (semiautomatici !)
3 2 3 4 1 3 3# (inesauribile)
3 4 2 3 1 3 3# (meteorologico)
4 3 2 3 3 1 3# (riequilibratura)
2 3 2 3 3 1 3# (cobaltoterapia, minicalcolatori !)
2 3 2 4 3 1 3 3# (radiostereofonico !)
2 3 2 4 3 2 4 1 3# (militareindustriale !)

Prosodic rules underlying our automatic prosodic transcription module can be easily formulated as follows:

1. every word has one single primary stress which is associated to (1) for open syllables, (6) for syllables closed by sonorants, and (7) for syllables closed by /r/;
2. syllables adjacent to the stressed syllables, must be unstressed and marked (3);
3. in words containing more than three syllables there can be other prominent syllables, marked (2), under the following conditions:
   a. they must be in pretonic syllables and be followed by an unstressed syllable;
   b. exceptionally, in words with stress on fourth before last syllables, thus possessing four unstressed syllables, there can be another prominent syllable, alternating unstressed syllables:
4. there cannot be two adjacent prominent syllables;
5. a reduced syllable, marked (5), can be alternated to an unstressed syllable, if it is positioned in post-tonic position and is followed by an unstressed syllable.
6. compound words can undergo internal phenomena of elision, marked by (0), for assimilation effects at word boundary;
7. semivowels are indicated by (4).

In order to apply our rules we computed all syllable structures of our database, where we differentiated open from closed syllables, diphthongs and vowel and consonant clusters. We came up with the following data:

- total number of syllables 96778 over 26828 word forms (we omitted grammatical words)
- 2623 different syllable types

4. Conclusions

We briefly tested the tool with American students of Italian and verified the closeness of the grading of difficulties with interlanguage levels measured previously by means of a grammatical test. Students have found the possibility of hearing mistaken pronunciations highly fruitful and contributing in an important way to their ability to produce correctly sounding Italian words. However, we haven't been able to complete an experiment given the lack of regular attendance by the class in that period.

As a further improvement of the tool we built, we would like to port it on the web and add a feature that collects context from current online Italian newspapers in order to support lexical learning with contextual larger information that is aimed at providing semantic and pragmatic information. We also need to port it from pre-2000 Apple systems no. 9 to 10 and recast the application from HyperCard to other available multimodal programming languages.

References

[1] Heilman, M., K. Collins-Thompson, J. Callan, and Eskenazi, M., "Classroom success of an intelligent tutoring system for lexical practice and reading comprehension", In: *Proc. Interspeech 2006*, Philadelphia, 2006.

[2] Selinker L., *Redescovering Interlanguage*. London: Longman, 1992.

[3] Delmonte R., "L'apprendimento delle regole fonologiche inglesi per studenti italiani", In: *Atti 8° Convegno GFS-AIA*, Pisa, 177-191, 1988.

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

[5] Delmonte R., "Il TTS per simulare l'interlingua in sistemi per l'autoapprendimento delle lingue", In: *Atti XI Giornate di Studio GFS, Multimodalità e Multimedialità nella comunicazione*, Padova, 45-52, 2000.

[6] Delmonte, R., Stiffoni F., "SIWL - Il Database Parlato della lingua Italiana", *Convegno AIA - Gruppo di Fonetica Sperimentale*, Trento, 99-116, 1995.

[7] Delmonte, R. "SLIM Prosodic Automatic Tools for Self-Learning Instruction", *Speech Communication*, Vol. 30, 145-166. 2000.

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

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

[10] Delmonte R. "Speech Synthesis for Language Tutoring Systems, In: V.Melissa Holland & F.Pete Fisher(eds.), *The Path of Speech Technologies in Computer Assisted Language Learning*, Routledge - Taylor and Francis Group, New York. 123-150, 2008.

[11] http://www.isca-speech.org/archive/eurospeech_1991/e91_1291.html

[12] Delmonte R., "Linguistic Tools for Speech Understanding and Recognition", in P.Laface,R.De Mori(eds), *Speech Recognition and Understanding: Recent Advances,* NATO ASI Series, Vol.F 75, Springer -Verlag, 481-485, 1991.