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

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Contribution of Prosody to the Perception of Spanish/Italian *accents*

Philippe Boula de Mareüil[1], *Giovanna Marotta*[2] *& Martine Adda-Decker*[1]

[1] LIMSI-CNRS, Orsay, France
{mareuil;madda}@limsi.fr
[2] University of Pisa, Pisa, Italy
marotta@ling.unipi.it

Abstract

Advantage is taken of new technologies, and in particular speech synthesis, to clear up the relative importance of prosody (melody and rhythm) in the identification of a foreign language or *accent*. The methodology we propose, based on the prosody transplantation paradigm, can be applied to different languages or language varieties. Here, it is applied to Spanish and Italian. We built up a dozen sentences which are spoken in almost the same way in both languages. And we wanted to study what is perceived when the segmental and suprasegmental characteristics of these two languages are crossed. Results obtained with French, Italian and Spanish listeners converge and suggest that prosody plays a greater role than the articulation of phonemes to identify the Spanish/Italian language and *accent*.

10.21437/SpeechProsody.2004-157

1. Introduction

The goal of this paper is to disentangle the influence of segmental vs. suprasegmental features (i.e. the phoneme string vs. prosody) in what is perceived as a foreign language or *accent*. Is it by chance that the term *accent* has been chosen to designate a foreign way of speaking? This question also applies to regional *accents*. Invariants probably exist, beyond regional differences, between a Neapolitan and a Milanese speaking English, but it is not proved. Some authors suggest a more important plasticity of melody [5], which we can observe in infants in phase of language acquisition. It has been claimed that the characteristic categories of the mother tongue prosody are extracted very early — prior to the lexicon acquisition [6].

It is noteworthy that, in linguistic atlases, little space is usually devoted to prosody, somewhat of an unexplored field even in dialectology. Though, studies especially within the framework of the IViE project [3] revealed how certain intonational patterns are specific to some English *varieties*. For lack of space, let us only cite one dissertation dedicated to the contribution of intonation to the impression of German *accent* in English and English *accent* in German [4]. Some studies were carried out using procedures put forth in [6], to discriminate between different rhythmic classes. Others attempt to elucidate the role of intonation rather than rhythm. Some of these studies (concerned with English, German, Italian or Spanish) are in keeping with the guiding principles of the autosegmental-metrical model of intonation [5, 4, 3]. This theoretical framework assumes autonomous segmental and suprasegmental tiers, which is interesting for our issue.

Perceptually too, prosodic clues facilitate human language acquisition. On the whole, however (with maybe the exception of stress and phoneme duration), the question of the role of prosody in the perception of a foreign *accent* has barely been tackled. The Speech Learning Model (SLM) developed by Flege, for instance [2], primarily addresses the notion of phonetic similarity and **phoneme** acquisition — particularly in relatively experienced speakers of a second language (L2). Prosody has often been neglected, perhaps owing to experimental difficulties linked to equipment problems.

Speech synthesis now allows us to sort out the role of segmental and suprasegmental contents in the perception of a foreign *accent.* It is a good tool to make allowances, since it enables us to monitor a number of parameters: that is the reason why it has been used for delexicalisation and monotonisation purposes [6]. It has been used, together with simulated or altered speech in research on foreign *accent* [4]. To obliterate (most of) the segmental structure, low-pass filtered speech is often used.

In this study, which in return may find applications in automatic language identification, text-to-speech (TTS) synthesis is regarded as a tool. But it is also of practical and theoretical interest to understand why a synthesis voice is perceived with this or that *accent*, which can be caricatured.

The present study investigates the identification of two neighbouring *accents*: Italian and Spanish *accents*. To study prosody independently of segmental properties, the methodology which is proposed here could apply to the issue of regional or social *accent* (e.g. rural vs. urban). But when two forms of the same language are mutually intelligible, the need to be understood is not motivated in the same way. We come up against the matter of the prestige attached to this or that variety, this or that dialect. Working on languages of comparable status such as Spanish and Italian enables us to factor out a whole range of stylistic, historical and social parameters (social class, levelling, etc.).

For the purpose of this study, we built a dozen sentences, which are spoken in almost the same way in Spanish and in Italian. And through various experiments, we studied what is perceived when we cross the segmental and suprasegmental features of these two languages.

This paper is organised as follows: next section presents the conducted experiment (material design, methodology and protocol). Listening tests were administered to different populations. Results, which suggest a major role of prosody, are provided in section 3, before concluding (section 4).

2. Experiment: corpus and protocol

To analyse the perception of Spanish/Italian *accent,* we designed a corpus of 14 sentences of 15 syllables on average, which are (almost) spoken in the same way in Spanish and Italian: e.g. *ha visto la casa del presidente americano* ("you/(s)he saw the American president's house"). And we intended to examine what is perceived when we cross the segmental and suprasegmental cues of these two languages.

The experiment described in the remainder of this paper makes use of diphone speech synthesis, a method which relies on the concatenation of pre-stored units stemming from natural voices. The Italian and Spanish voices used here are those of the Elan PSOLA-based multilingual TTS system [1]: independent of this study, they come from native speakers who were recorded in France, where they had lived for a short

while. The pitch and duration parameters are then handled thanks to the PSOLA algorithm. Energy is not processed: it is only normalised. As for pitch, it is defined for each phoneme by an initial target, a final target and possibly an intermediate target: one or two linear pitch movements are thereby associated to each phoneme. The pitch value of supposed unvoiced segments is equated to zero; and the initial pitch of each phoneme is connected to the final pitch of the preceding one, if any.

It is possible to replicate the experiment with different types of stimuli: modified natural speech, possibly filtered, duration and pitch values predicted by a full TTS synthesis system, etc. Freely available speech processing software such as PRAAT, permitting prosody manipulation, now yield good results. But modifying fundamental frequency (F_0) and duration also creates artefacts (audible voice quality dissimilarities with respect to unaffected utterances); and it has the disadvantage of being more time consuming than using diphone synthesis. With diphone synthesis, one may even play on the allophone inventory of Spanish, the typical lengthenings of the Italian language, etc. Here, we use prosody transplantation and cross-language phonemic transcoding. Native speakers of Spanish and Italian were recorded; their prosodic parameters were extracted, checked manually with the help of native speakers and applied on the diphone bases.

2.1. Text preparation

Fourteen sentences were created, while trying to maintain a certain semantic coherence in order to select different modalities (exclamatory, assertive, interrogative), varied grammatical structures (with prepositional phrases, conjunctive or relative subordinate clauses), in different tenses (present, perfect, imperfect, preterit, future), and as many function words as possible. In the phonetic field, we watched over the pronunciation of phonemes such as /ʎ/ and over the diversity of stress patterns: oxytone (e.g. *autobús*), paroxytone (e.g. *perdono*), proparoxytone (e.g. *crédito*). 80% of polysyllabic words in the corpus are paroxytone (i.e. stressed on the penultimate syllable). This figure is consistent with the language. Of course, the matching is not perfect between the phonemes of Italian and Spanish, first because in the variety which serves as the cultural prestige norm for Italian (Toscan), the phonological inventory is made up of 7 vowels (/a ɛ e i u o ɔ/), compared to 5 for Spanish (/a e i u o/); second because the spirantised allophones of the Spanish language do not merge exactly with the Italian fricatives ([β] with /v/ for instance). But we can retort to the first point that variability is prevalent within Italian mid vowels; in addition, the question of "close" but not identical phonemes is far from being solved. In our experiment, the Italian synthetic voice speaking Spanish closes mid vowels, but does not apply spirantisation rules to voiced plosives such as /d/ → [ð] / V_V. As for the Spanish synthetic voice speaking Italian, it does not apply these spirantisation rules, but nor does it open mid vowels.

2.2. Recordings

One Spanish male from Madrid (SM), one Spanish female from Barcelona (SF, also of Castilian mother tongue), one Italian female from Milan (IF) and one Italian male from Naples (IM) freely volunteered to read the obtained sentences. The tape-recording took place in Paris, in a soundproof booth with a high-quality microphone located about 20 cm from the mouth, using a DAT (input sampling rate of 48 kHz). The data (three repetitions of each utterance per speaker, on average) were then transferred onto the computer with a sampling frequency of 22.05 kHz and a 16 bit resolution, mono, for further processing. Only one repetition per sentence was retained for each speaker.

The speakers, who were all under 40 years old and graduated, had nothing to do with the experiment. They were instructed to articulate properly, without however making too many pauses other than the ones punctuation marked. Moreover, the Spanish speakers were asked to pronounce the digram 'll' as /ʎ/ and not /j/ (*lleísmo*, which distinguishes the lateral liquid and the palatal fricative), and the Italian speakers were asked to utter the intervocalic 's' as /s/ and not /z/, in words such as *casa* (the Toscan norm).

Speech rates ranged from 12.5 for IF to 15.5 phonemes/sec. for SM — not including pauses. IF's speech rate is relatively slow (with respect to SM especially); but if we only look at the phonemes of unstressed syllables, we can see that the mean duration difference diminishes. This restriction is justified by the fact that Italian and Spanish are traditionally regarded as syllable-timed languages, tending to have isochronous unstressed syllables. Interestingly, the duration ratio between stressed and unstressed phonemes is 1.5 for Italian and 1.1 for Spanish. The Italian lengthening substantially contributes to a slower speech flow.

The Italian lengthenings may be at the origin of a wider pitch range in this language, which is often portrayed as sing-song, impressionistically. Defined in semitones with respect to the maximum and minimum F_0 targets of voiced segments as $12 \log_2(F_0\text{max}/F_0\text{min})$, pitch range is 14 semitones for the two Italian speakers, whereas it is only 12 and less than 11 semitones for SM and SF. F_0 standard deviation — where non-null F_0 values are expressed in semitones with a reference of 1 Hz — is greater for Italian speakers (> 2 semitones) than for Spanish speakers (<2 semitones).

2.3. Methodology

The prosodic parameters extracted from SF, SM, IF and IM were grafted onto a diphone base, by using a prosody transplantation tool [1]: given an audio file and the text corresponding to what is pronounced, the system generates a file in "prosodic writing" and an audio file including the computed prosodic characteristics copied from the original — in terms of phoneme by phoneme pitch and duration values. The mean pitch of voiced segments was 177 Hz for IF, 202 Hz for SF, 106 Hz for IM and 107 Hz for SM. IF's (resp. SF's) mean pitch value was multiplied by 1.05 (resp. 0.95) to better fit the intrinsic pitch of the diphone voice. This way too, we avoid the bias of a strategy consisting for the listeners to rely on the voice height to discriminate the speakers.

Each sentence of our corpus thus allowed us to generate stimuli with comparable mean pitch, speech rate and intensity: 2 languages (Spanish and Italian) × 2 types of prosody (native and non-native) × 2 genders (male and female) = 8 stimuli. To the 80 stimuli corresponding to the first ten sentences, 4 stimuli were added to give a sample of the 4 voices (Spanish and Italian male and female) with native or crossed prosody. These 4 stimuli were presented at the beginning of the test to the listeners, and were not counted in the results. They were preceded by instructions and a "learning" phase of another 4 natural utterances of Spanish and Italian (6 second long, from 2 males and 2 females), without any link with the experimental material, in order to refresh listeners' memory. In the very test, the trial order was randomised, and the order changed between the subjects.

The experiment took place in a quiet room, through headphones. It lasted about 15 minutes. The listeners, all with normal hearing, were not paid for this task. They were not

urged to answer; but they could listen to each stimulus only once. They were instructed that they would listen to acoustically modified speech, coming from native speakers of Spanish and Italian, who could speak both languages. They were informed that the test sentences which were read in either language by the speakers could be spoken almost in the same way in Spanish and Italian. And they were asked to judge what they would listen to through a user-friendly interface programmed with the E-prime software. We believe that the results were not biased by the experimenter, the first author, who is francophone, but also speaks Italian and Spanish — certainly with a French *accent*.

3. Results

3.1. Results of the perceptual test with Italian subjects

A first perceptual test was conducted in Pisa (Toscany), with students in linguistics and staff from the University and Scuola Normale Superiore. The task consisted of judging whether what was displayed was Spanish, Spanish with an Italian accent, Italian with a Spanish accent or Italian. Twenty subjects (6 male, 14 female) participated in the experiment. Half of them were of Toscan origin, the other half from other regions of Italy. Most listeners self-rated their familiarity with the Spanish language 1 or 2, on a 10-point scale.

In Table 1 and below, S_vS_p refers to "Spanish voice with a Spanish prosody", S_vI_p to "Spanish voice with an Italian prosody", I_vS_p to "Italian voice with a Spanish prosody" and I_vI_p to "Italian voice with an Italian prosody". As for the following significance *p*-values, they were all computed by performing χ^2 tests. We can see an overweight number of answers "Italian with a Spanish accent". For S_vS_p sentences, this can be explained by an annexationist tendency which would incline Italian listeners to answer "Italian (with a Spanish accent)" as soon as they understand the meaning. For I_vI_p sentences, this overrated answer "Italian with a Spanish accent" may be due to the artificiality of the stimuli, a strangeness which can be interpreted as foreign-accented. But in the majority of cases, S_vS_p sentences are recognised as Spanish and I_vI_p sentences are recognised as Italian. In crossed sentences (I_vS_p and S_vI_p), the most frequent answers are "Italian with a Spanish accent". If this answer is slightly more represented in the case of I_vS_p sentences, the 1% difference with S_vI_p scores is not significant ($p>0.05$). It is the same if we consider the number of answers "Spanish with an Italian accent" for these I_vS_p and S_vI_p sentences: the difference is not significant. Therefore, from the analysis of these categories "Italian with a Spanish accent" and "Spanish with an Italian accent", we cannot conclude that prosody is the more influential feature. But if we examine "Spanish" and "Italian" answers, it turns out that S_vI_p is more recognised as Italian than as Spanish; and reciprocally, I_vS_p is more recognised as Spanish than as Italian. These results are highly significant ($p<0.01$), and suggest a more important role of prosody.

Table 1: *answers obtained with 20 Italian listeners (see text).*

	Spanish	Spanish with an Italian accent	Italian with a Spanish accent	Italian
S_vS_p	**181**	42	166	11
I_vS_p	90	75	**174**	61
S_vI_p	69	72	**170**	89
I_vI_p	14	53	125	**208**

During informal conversations after the trial, some subjects reported that they had relied on rhythmic cues or on the pronunciation of phonemes such as /r/ or /s/ (which typically tends to be apical [s̺] in the North and the Centre of Spain). We had no feedback concerning the identification of region-specific features in Italian, which rules out a feared bias.

3.2. Results of the perceptual test with Spanish subjects

The same task was completed in Barcelona with students in psychology (2 male, 18 female) of Castilian mother tongue. Both parents of each subject were Castilian-speaking, and all the subjects self-reported no or very poor familiarity with Italian. The results achieved are summarised in Table 2.

Table 2: *answers obtained with 20 Spanish listeners (see text).*

	Spanish	Spanish with an Italian accent	Italian with a Spanish accent	Italian
S_vS_p	**209**	95	81	15
I_vS_p	67	**148**	143	42
S_vI_p	67	**165**	92	76
I_vI_p	18	145	80	**157**

In the absolute or relative majority of cases, S_vS_p and I_vI_p, stimuli are identified as Spanish and Italian respectively. As expected, S_vS_p stimuli are better recognised than are I_vI_p stimuli, whereas it was the opposite with Italian listeners. I_vS_p stimuli are perceived as Spanish as many times as S_vI_p stimuli are; but these ones are more often perceived as Italian: the differences (76 vs. 42) is highly significant ($p<0.01$). A large number of answers "Spanish with an Italian accent" can be noticed — especially with S_vI_p sentences, but the difference in this respect with I_vS_p sentences is not significant ($p>0.05$). This trend is symmetric with the one observed with Italian listeners, even though the number of answers "Italian with a Spanish accent" is still sizeable.

If we add up Italian and Spanish listeners' answers, we can bring to light the fact that prosody detracts more from acceptability than segmental information. In particular, the impression of Spanish ***accent*** in Italian is given by I_vS_p stimuli more than by S_vI_p stimuli, and the difference is highly significant ($\boldsymbol{p}<0.01$). Also, the impression of Italian *accent* in Spanish comes from S_vI_p stimuli more than from I_vS_p stimuli, even though it is not so clear: the difference is not significant ($p>0.05$). This can be accounted for by arguing that prosody is more "marked" in Italian than it is in Spanish, while a kind of harsh or creaky voice quality is more reminiscent of Spanish.

3.3. Results of the perceptual test with French subjects

We found it too difficult to ask French listeners to judge whether what they heard was Italian-accented Spanish or Spanish-accented Italian. Nonetheless, we kept a forced choice between 4 possibilities, by introducing a confidence measure: the French subjects were requested to tell if the **mother tongue** of the speakers they heard was probably or very probably Spanish or Italian. Since they listened to the same stimuli as above, we expected their answers to be facilitated when the voice and the type of prosody matched.

The 20 French subjects (13 male, 7 female) who took part in the follow-up experiment were of various origins and backgrounds, but all of French mother tongue. They were half more familiar with Spanish and half more familiar with Italian (6 subjects) or with no language (4 subjects), somewhat in accord with the statistics of a recent demographic study (www.ined.fr/publications/pop_et_soc/pes376/PES3762.html).

As is apparent in Table 3, the French listeners recognised the stimuli with Spanish prosody as Spanish and the stimuli with Italian prosody as Italian, in 2/3 of cases, which is highly significantly above chance level ($p<0.01$). Other χ^2 tests were run to compare the performance on non-crossed stimuli on the one hand (67.25% for S_vS_p vs. 68.5% for I_vI_p), on crossed stimuli on the other hand (59% for I_vS_p vs. 59.5% for S_vI_p). The recognition scores turned out to be comparable between the two languages, or in other terms not significantly different ($p>0.5$). Nor do overall Spanish/Italian answers differ significantly across the group of listeners who are more familiar with Spanish and the other group ($p>0.05$). We can see that these scores are higher when the voice and the type of prosody match, and the differences are highly significant.

Table 3: *answers obtained with 20 French listeners (see text).*

	Very likely Spanish	Probably Spanish	Probably Italian	Very likely Italian
S_vS_p	**115**	154	95	36
I_vS_p	78	**158**	124	40
S_vI_p	54	108	**136**	102
I_vI_p	33	93	**142**	132

Let us now examine the results in more detail, taking advantage of the listeners' confidence in their 4-choice answers. Interestingly, the most frequent answers are probably or very probably Spanish for S_vS_p stimuli and probably or very probably Italian for I_vI_p stimuli. In hybrid stimuli, where the voice and the superimposed prosody are in conflict, the most frequent answers are "probably Spanish" and "probably Italian". I_vS_p stimuli exhibit fewer answers "very probably Spanish" than do S_vS_p stimuli (78 vs.115); and likewise S_vI_p stimuli exhibit fewer answers "very probably Italian" than do I_vI_p stimuli (102 vs. 132). These differences are highly significant ($p<0.01$).

Overall results can be interpreted as follows: the articulation of phonemes helps in identifying the speakers' origin, but prosody is the more reliable clue. This outcome confirms the one obtained with Italian and Spanish listeners, but it is surprising insofar as listeners were prompted to focus on speakers' mother tongue rather than target language. This precision was brought to the French listeners, who did not understand the meaning of the sentences as Italian and Spanish listeners could, in case they would have detected a mismatch between the two languages. As a matter of fact, very few listeners recognised that there were only 4 voices. Ramus imagined that rhythm might not be sufficient to discriminate Spanish from Italian — two syllable-timed languages [6]. Prosody is sufficient, under similar conditions of diphone-based synthetic speech.

4. Discussion and conclusion

At the term of this analysis, one can wonder if the latter enables us to assess the contribution of segmental and suprasegmental factors to the perception of foreign *accent*, or to language identification. Indeed, we very partially treated the phonological transfer phenomena, which may occur during L2 acquisition, and we did not especially attend to the way these "negotiations", language-specific approximations or "mappings" are taught. In particular, the Spaniards' tendency to spirantise L2 plosives is unclear. Following the SLM [2] (according to which production and perception remain subject to adaptation across the life span), we can even imagine that, in a longitudinal study, the place of prosody increases as long as the *accent* fades. Indeed, segmental errors may mask prosodic errors — as serious phonological errors can draw attention away from mere phonetic "errors". Prosody should arguably be implemented in a model such as SLM.

Here, we chose to consider an extreme case (that of a strong foreign *accent*), where the speaker would attune his/her segmentals as a minimum, and would adopt a perfect prosody in L2. In this framework, what we try to identify is the very foreign *accent*. In language identification (LId), listeners can base their strategies on the recognition of discriminating segments such as the Spanish /x/ or Italian geminates. These phonemes are absent from the sentences of the current experiment, which are by construction not phonetically balanced; this possibility is therefore excluded.

In the framework of this experiment, prosody happened to play an important role, and even the stronger role as compared to segments, all other things being equal. The tendency was observed with Italian, Spanish and French listeners, which defies commonly held beliefs and intuitive assertions in LId, in dialectology and in the field of foreign *accent* that prosodic differences are secondary. Our results may originate in Italian's wider pitch range and lengthenings. Though, they should be taken with caution, since they may be put down to the use of diphone synthesis, which behaves as a bottleneck. Segmental quality is not necessarily better preserved by using modified natural speech. But if it is the case, it is interesting to replicate this experiment with less degraded signals. With this end in view, we recorded Italian/Spanish bilingual speakers. They read the same sentences in the Spanish manner and in the Italian manner; and we plan to study if our results are confirmed with this new material.

As in a classical LId experiment, the problem of individual characteristics is no nearer solution. Is an *accent* detected or is it the speaker? This may also depend on who judges. In the case of French listeners of Spanish/Italian utterances, what is assessed is the linguistic *representation*, one's image of a foreign language (here from the same Romance group). And the results of this pilot study will be used in future investigation on Spaniards and Italians speaking French.

5. Acknowledgements

We gratefully acknowledge I. Vasilescu and N. Sebastian Gallés for their useful help in the design of this experiment.

6. References

[1] Boula de Mareüil, P. *et al*., 2001. Elan Text-To-Speech : un système multilingue de synthèse de la parole à partir du texte. *Traitement Automatique des Langues* 42(1), 223-252.

[2] Flege, J.E., 2003. Assessing constraints on second-language segmental production and perception. In *Phonetics and Phonology in Language Comprehension and Production*: *Differences and Similarities,* A. Meyer & N. Schiller (eds.). Berlin: Mouton de Gruyter.

[3] Grabe, E. & Post, B., 2002. Intonational Variation in English. *Speech Prosody,* Aix-en-Provence, 343-346.

[4] Jilka, M., 2000. *The contribution of intonation to the perception of foreign accent. Identifying intonational deviations by means of F0 generation and resynthesis.* PhD thesis, University of Stutgart.

[5] Ladd, D.R., 1996. *Intonational phonology*. Cambridge: Cambridge University Press.

[6] Ramus, F., 1999. *Rythme des langues et acquisition du langage*. PhD thesis, Paris: EHESS.