ACOUSTIC-PERCEPTUAL CORRELATES OF SENTENCE PROMINENCE IN ITALIAN*

Mariapaola D'Imperio

Abstract

Research on the acoustic correlates of perceived accentual prominence has generally focused on fundamental frequency (F0) alone, while few studies have attempted to shed light on how other parameters, such as duration and intensity, might interact with F0. A previous study on Italian lexical stress perception shows that duration has a major role. The present work reports on results of an experiment using synthetic speech to test which aspects of the signal, among F0, duration and intensity, are more influential in the perception of prominence structure at the sentence level and whether there are differences between questions and statements. To this end, a series of hybrid LPC-resynthesized stimuli were presented to 22 Italian listeners for forced-choice judgments. The results suggest a bigger impact of the hybridization on interrogative utterances.

1. Introduction

As defined here, prominence is the subjective salience of an element in an utterance. Most recent research on the acoustic correlates of perceived prominence in

*I would like to thank Nick Cipollone for his much needed help in writing the hybridization program used in this study (re-elaborated from an earlier version written by Mary Beckman). Thanks also to Keith Johnson and Jen Muller for comments on a previous version of this paper. Finally, I would like to thank Jose Benki for help with the statistics and for discussing the implications of the results.

speech has focused on F0 or pitch (e.g. Liberman & Pierrehumbert, 1984; 't Hart *et al.*, 1990; Ladd *et al.*, 1994). Comparatively few studies have attempted to shed light on the complex nature of prominence as a result of the interplay of parameters other than F0, e.g. duration and amplitude[1].

From the literature on the topic, it appears that prominence is primarily cued by the presence of a noticeable pitch change or by extreme (either high or low) pitch levels relative to the context (Pierrehumbert, 1980; Pierrehumbert and Beckman, 1988). However, it has been noted that even though pitch variations are not as marked in spontaneous speech as for read speech, clearly perceptible prominences can still be detected, which could be attributed to other physical indices, such as duration and/or amplitude (Boves, ten Have and Vieregge, 1984). More recently, Campbell (1995) has shown that in dialogue speech spectral information can compensate for the lack of tonal cues, when detecting prominence. It also appears that interesting differences exist in the perception of prominence between listeners with different linguistic backgrounds. For instance, Lehiste and Fox (1993) found a stronger effect of duration on Swedish listeners, as opposed to English listeners, in prominence perception.

The present study aims at uncovering the perceptual role of certain acoustic correlates of prominence in Italian, namely duration, amplitude and fundamental frequency. The relative salience of the aforementioned correlates has been already tested for isolated words in this language. Previous experiments (Bertinetto, 1980) aimed at discovering the relative weight of each of those correlates in determining lexical stress pattern, in minimal pairs such as *áncora* "anchor" and *ancóra* "again", but did not study prominence at the sentence level.

Another important difference with previous studies pertains to methodological issues. We are still far from understanding the complex proportional variations due to variables such as position in the utterance or natural occurring combinations of different parameters for such free manipulations to be useful. Hence, the present study attempts to overcome past methodological problems and to examine sentence level phenomena. The stimulus set employed in this work was generated through a technique that is very different from the one used in earlier experiments on prominence perception in Italian. Specifically, the correlates of prominence will not be directly manipulated here. An experiment was then designed in order to assess the weight of each of the acoustic correlates of stress in Italian, by cross-combining the acoustic substance of natural utterances where the focus, broad or narrow, is placed on different elements.

Despite the methodological discrepancies, previous research suggests that Italian subjects are very sensitive to durational differences, both in perception of lexical stress patterns (Bertinetto, 1980) as well as in the perception of unstressed syllable duration (Bertinetto and Fowler, 1989). It is plausible, therefore, that sentential prominence in Italian is cued by duration and intensity, as well as F0. We expect, then, that replacing only one prominence correlate (i.e., duration, or intensity, or F0) of a "donor" utterance with that of a "recipient" utterance will affect perceived prominence. This kind of manipulation was carried out for this study, whose details will be presented below. The results presented here suggest in fact that the role of duration is particularly important in the perception of specific intonation patterns.

[1]In this paper, I shall use the term "amplitude" and "intensity" interchangeably to refer to the physical property of the signal producing the subjective sensation of loudness.

2. Previous studies

The investigation of perceptual cues of stress goes as far back as the 1950s, when the classic experiments described in Fry (1955, 1958) were performed. Those studies found that acoustic prominence is concerned with certain physical correlates of the salient syllable in a word. This aspect of prominence is believed to be associated primarily with high degree of pitch variation, long duration and high amplitude (Fry, 1955,1958; Lieberman, 1960; Lehiste, 1970).

While Fry's studies had determined that F0 was indeed the most important correlate for stress in English, three decades later Beckman (1986) reestablished the role of intensity through the use of a loudness measure[2]. In her perception experiments with Japanese and English, she found in fact that F0 has a much greater role in Japanese than in English for the purpose of signaling stress. English listeners seemed to pay more attention to loudness differences[3].

Recently, most experiments on prominence perception have concentrated on the role of fundamental frequency (Ladd *et al.*, 1994; Terken, 1992; Hermes and Rump, 1994; Bartels and Kingston, 1996). Terken (1992) investigated the relative importance of fundamental frequency change and fundamental frequency maximum in determining prominence judgments in subsequent peaks, finding that the relation is more complex than expected. Hermes and Rump (1994), despite admitting that "the physical attribute underlying prominence perception is multidimensional" (p. 90), investigate perceptual prominence of falling and rising pitch movements while regarding intensity and duration as secondary cues that can only "intensify" an already existing accent. The authors used a method in which subjects had to adjust the pitch of an accented syllable in order to match the prominence of a previously heard accent. As was noticed by the authors, however, since the only adjustable dimension was pitch, it may well be that subjects tended to pay attention only to this cue and not to others.

In Italian, unlike English and Swedish, few perceptual experiments focusing on prominence, especially at the sentence level, have been performed. The only study that has explored the perceptual interaction of the various acoustic correlates in Italian is Bertinetto (1980). This study investigated the relative weight of duration, fundamental frequency and intensity on the perception of stress in the bisyllable [papa]. This segmental sequence can have two different meanings according to the stress pattern, i.e. "Pope" ['papa] or "daddy" [pa'pa]. Bertinetto (1980) argued that the role of duration is markedly greater than that of intensity and F0 for signaling word stress in Italian. F0 was instead found to be the weakest cue. He also found a listener bias in favoring the second syllable of the bisyllable when judging stress. This could have been a result of the

[2]This measure of loudness is actually labeled "total amplitude" in Beckman (1986) and is a measure that combines duration and amplitude.

[3]Beckman, who finds a pattern very similar to Nakatani and Aston (1978), offers an explanation for the difference of her results with Fry's findings. In Fry (1958), F0 overruled amplitude and duration as a correlate of word stress in a dramatic way. Beckman notices that the kind of synthesis used by Fry might have unnaturally reproduced intensity by simply attributing level values to the segments, without preserving naturally occurring contours and thus sounding very unnatural. Conversely, the LPC resynthesized stimuli that Beckman and the present study employ might make for more naturally sounding stimuli and, therefore, for a higher effectiveness of the amplitude parameter.

positional characteristics of the two syllables[4]. Though the results are very interesting, this study had some methodological limitations, which prevent a conclusive interpretation. Those limitations are mainly related to the issue of directly manipulating prosodic cues, which was avoided in the present study.

An additional variable introduced in this study pertains to the influence of modality in prominence perception, in other words whether questions are different from statements in this respect. Ultimately, I would like to discover whether the pitch values alone produce an overriding pattern of prominence responses or if the duration/amplitude values can, as predicted by Bertinetto's results, significantly determine the identification of the prominence pattern. Since we are not at a stage in which we can give an account of the prosodic organization of Italian, it was necessary to validate prominence patterns identified according to standard linguistic theories and to acknowledge observed patterns that do not strictly follow established theoretical beliefs.

For this purpose, a preliminary study (D'Imperio, 1997a) was designed in order to assess the perceptual prominence response of Italian subjects to natural speech stimuli varying in focus placement (early, medial, late) and focus type (broad vs. narrow). This preliminary experiment serves as background to the experiment described here, in which synthetic stimuli were manipulated. The experiment validated the robust recognition of intended focus in narrow focus utterances, while yielding results around chance for broad focus statements (while late focus was always identified as such in broad focus questions). Broad focus seems to be signaled by an accent that is less salient than the narrow focus accent, in that it is downstepped. Also, the lexical item that is associated to it is generally not chosen as the "most prominent" within the utterance (D'Imperio, 1997a). Therefore, we expect that the "weaker" perceptual prominence of broad focus accents will be enhanced when one of the acoustic cues of narrow focus utterances is combined with it. Additionally, narrow focus identification will be less robust when one of the correlates of broad focus is combined with a narrow focus utterance.

The analysis of the intonation contours presented here was carried out within the ToBI framework (Beckman and Ayers, 1994). The melody is basically decomposed into "target levels" (highs and lows), which can be thought of as the "notes" associated to some specific segmental locations.

3. Methods

3.1 Stimuli

A set of stimuli was created by using the hybrid resynthesis technique first developed by Nakatani and Aston (1978) and subsequently adopted by Beckman (1986) and Hirschberg and Ward (1993). The technique consists in, first, sampling RMS amplitude, timing, LPC coefficients and pitch information for each original utterance of each stimulus pair and then synthesizing new files in which one of the sampled features of the original utterances was exchanged for those of another (with synthetic files produced by linearly interpolating between sample points). As a last step, new utterances are resynthesized on the basis of the "hybrid" files using LPC resynthesis.

[4]As it turns out, final stressed syllables appear to be shorter than syllables in other positions in production studies.

Direct manipulation of the stimuli was, as mentioned above, avoided, since it is impossible at this point to estimate parameter intervals that would be equal as to perceptual effect. The hybridization technique allows one to avoid the risk of involuntarily creating discrepancies in step sizes that would make the perception effect of one acoustic dimension seem stronger than it is in reality.

Stimuli consisted of simple Subject-Verb utterances, using the sentence *Mario esce* "Mario goes out". The original utterances were identical from a segmental point of view, while various intonational combinations of modality (questions or statements) and focus type were superimposed on them. The utterances were all produced by a female speaker of the variety of Italian spoken in Naples (the author).

As shown in Table 1, the same sentence was uttered as either a neutral utterance with broad focus (Broad) or as a narrow focused utterance, where the focus occurred on either the Subject (NarrowS) or the Verb (NarrowV). The utterances were all auditorily transcribed to check for intended focus pattern. The recordings were made in the Department of Linguistics Lab, Ohio State University, where they were digitized at 16 kHz on a SUN Sparc Station using ESPS Waves+.

Mario esce "Mario goes out"	broad focus (Broad)
MARIO esce "MARIO goes out"	narrow focus on S (NarrowS)
Mario ESCE "Mario GOES OUT"	narrow focus on V (NarrowV)

Tab. 1 Patterns of sentence stress in the test utterances.

For the hybrid resynthesis, spectral coefficients of the natural utterances were obtained through an 18th-order LPC (Linear Predictive Coding), while amplitude and fundamental frequency values were extracted using an autocorrelation F0-tracking program. The values obtained were used to create hybrid utterances where just one of the acoustic correlates of prominence was exchanged at a time. For instance, the F0 donor utterance could be *MARIO esce?*, with nuclear (i.e., the most prominent accent in the sentence) accent on the subject (see Figure 2, middle), while the duration and (RMS) amplitude donor utterance would be *Mario ESCE?* (see Figure 2, lower), with nuclear accent on the verb. In such a case, the goal is to find which word will be judged the most prominent by the listeners, i.e. whether F0 cues or duration and intensity cues will have a stronger impact in this sense.

Non-hybrid	**F01**+LPC1+RMS1+D1
F0 change	**F02**+LPC1+RMS1+D1
RMS change	**RMS2**+F01+LPC1+D1
Dur. Change	**D2**+F01 +LPC1+RMS1

Tab. 2 Acoustic correlate manipulations used in the Experiment. 1 = donor utterance; 2 = base utterance.

The order of the base utterance/donor utterance combination could be reversed to allow for indirect exchange of original spectral parameters. For example, the fundamental

frequency of a broad focus utterance was in one case combined with spectral, amplitude and duration values of values of a narrow focus utterance (either on the subject or on the verb). In another instance, the fundamental frequency of the narrow focus utterance was combined with spectral, amplitude and duration values of the broad focus utterance. Along the same lines, the amplitude or duration of the donor utterance was in another instance combined with all other acoustic values of the base utterance. For example, as a result of inserting the fundamental frequency of the broad focus *Mario esce* in the narrow focus *MARIO esce*, with focus on *Mario*, we obtain that the stressed syllable *Ma-* (of *Mario*) will be strongly marked by the substantive values of duration and amplitude, but will not be marked by a strong pitch accent. All combinations of broad focus utterance plus one of the features of narrow focus utterances (and viceversa) were obtained. Narrow focus utterances were never combined with each other, since this produced unnatural effects.

Syllabic boundaries were marked in the original utterances, yielding 4 cuts or "anchors" (one for each syllable). Frame numbers were obtained for each cut. When duration was the parameter exchanged, the frame number for each cut in the base utterance was exchanged for the frame numbers of the donor utterance, while a linear interpolation algorithm was used to obtain new spectral, amplitude and F0 values in the hybrid utterance. When amplitude or fundamental frequency values were taken from the donor utterance, those were interpolated to the frame number relative to the anchors in the base utterance.

The original spectral coefficients were recombined with adjusted amplitude and F0 contours or simply readjusted as to frame number. Hybrid utterances were then resynthesized through LPC resynthesis. The spectral coefficients of the hybrid utterance were always derived from the base utterance and the only permissible combination was broad focus plus narrow focus utterance, and neither broad-broad nor narrow-narrow combinations were employed.

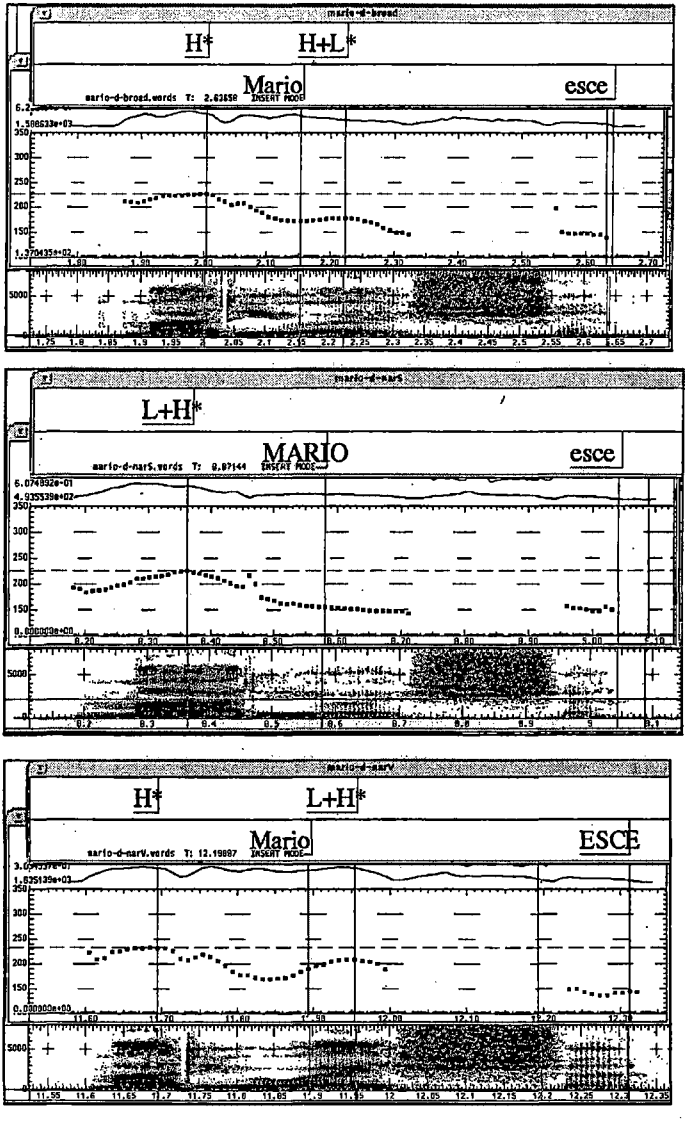

Figure 1. F0 curves and spectrograms for a broad focus declarative (upper), a declarative with narrow focus on the subject (middle) and a declarative with narrow focus on the verb (lower).

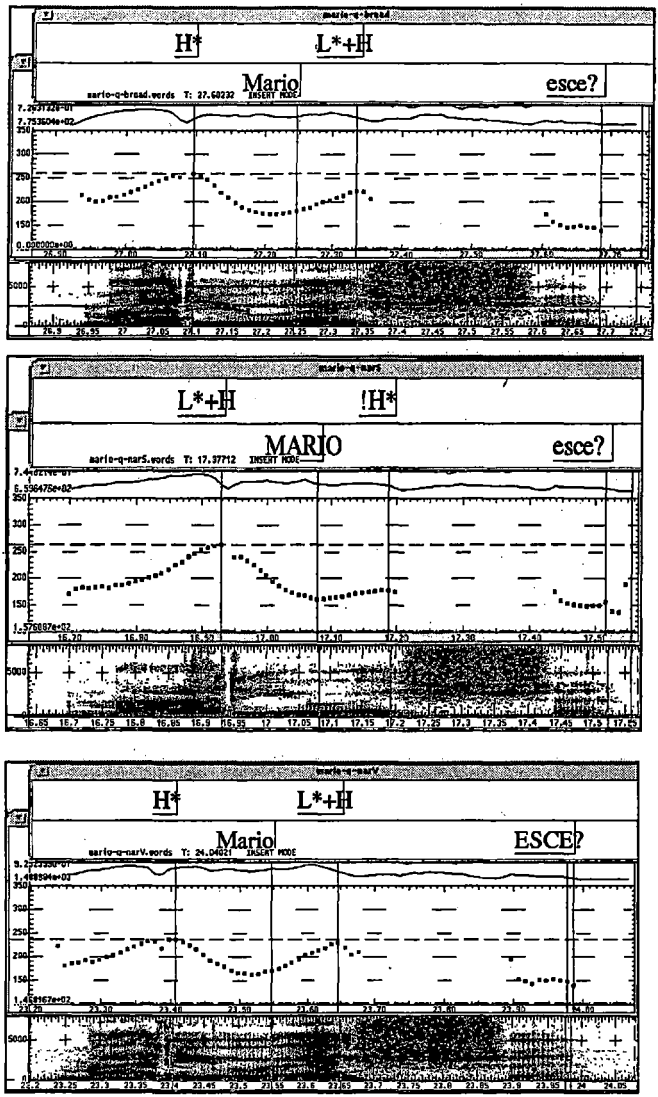

Figure 2. F0 curves and spectrograms for a broad focus question (upper), a question with narrow focus on the subject (middle) and a question with narrow focus on the verb (lower).

3.2 Procedure

The 24 hybrid stimuli plus 6 non-hybrid resynthesized originals, used as controls, were presented to the listeners in random order. The task consisted of choosing the "most important" word in the utterance (forced choice judgment) by clicking on its orthographic version presented on a computer screen. After the choice was made, the computer played the subsequent stimulus, and the following choice was made.

The listeners were instructed to listen carefully to each sentence and to choose the answer as quickly as possible after listening to each stimulus, even when not entirely sure about it. Explicit use of proper linguistic terms such as "prominence" and "focus" was avoided in order to leave linguistic notions outside of the task, so that even naive listeners could perform it without confusion.

A short training session preceded the set of trials, where the experimenter presented examples of utterances with varying intended focus structure (see Tab. 1) and had the subject point at one of the words as being the most important. The experiment was self-paced, and each stimulus was played only after the previous choice was made.

3.3 Listeners

Twenty-two listeners participated in the experiment. All but two of the listeners were undergraduate students at the University Federico II of Naples, with ages varying between 22 and 27. They were all speakers of Neapolitan Italian and hence had the same geolinguistic background of the speaker who produced the stimuli[5]. They all had normal hearing and performed the task without problems. Some of the subjects had attended introductory linguistic courses.

4. Results

The listening test yielded a total of 3300 responses (30 stimuli * 5 repetitions * 22 subjects). Three factors were used in the repeated measure Analysis of Variance (ANOVA), i.e. MODALITY, MANIPULATION and FOCUS TYPE (see Table 3). MODALITY had two levels (question vs. statement intonation), while FOCUS TYPE had 4 levels. These levels are the result of dividing up the hybrid stimuli as follows: broad focus utterance plus one of the correlates of utterances with narrow focus on V (Broad+NarrowV), broad focus utterance plus one of the correlates of utterances with narrow focus on S (Broad+NarrowS), utterance with narrow focus on S plus one of the correlates of utterances with broad focus (NarrowS+Broad) and utterance with narrow focus on V plus one of the correlates of utterances with broad focus (NarrowV+Broad). The natural utterances were grouped with the hybrid ones, according to focus type. MANIPULATION had four levels, according to the parameter that was manipulated (Duration, F0, amplitude, non-hybrid). Therefore, the design was a 2x4x4 factorial. The variables were manipulated within subjects. The number of judgments favoring verb prominence for each stimulus was determined, henceforth NUMBER OF V JUDGMENTS, which was the dependent measure. Planned comparisons were also carried out on relevant scores.

[5]Only three of the subjects were knowledgeable in linguistics, but none was aware of the purpose of the experiment.

Factors	*Levels*
MODALITY	Question, Statement
FOCUS TYPE	Broad+NarrowV, Broad+NarrowS, NarrowS+Broad, NarrowV+Broad
MANIPULATION	Duration, F0, RMS amplitude, non-hybrid

Tab. 3 Factors and levels of the statistical analysis.

In Table 4 the main effects and interactions of MODALITY, MANIPULATION and FOCUS TYPE are given.

Effects	F	P-value
Main effects		
Modality	60.37	<0.01
Focus Type	140.5	<0.01
Manipulation	8.34	<0.01
Two-way interactions		
Modality * Focus Type	0.3	NS
Modality * Manipulation	3.6	.02
Focus Type * Manipulation	50.8	<0.01
Three-way interaction		
Modality * Focus Type * Manipulation	4.49	<0.01

Tab. 4 Main effects and interactions of MODALITY, MANIPULATION and FOCUS TYPE.

The results support the hypothesis that acoustic manipulation can affect the perceived intended focus of the base utterance. A large main effect of both FOCUS TYPE and a main effect of MODALITY were found. Moreover, a significant interaction of MANIPULATION with FOCUS TYPE and a significant three-way interaction were found.

4.1 Statements

Figures 3 and 4 show the mean overall results for the four focus types. The bars in the two figures are the mean for NUMBER OF V JUDGMENTS for hybrid stimuli (duration, F0 and RMS amplitude) vs. non-hybrid stimuli. The manipulations associated with the different labels are shown in Table 5.

HYBRID STIMULUS	BASE	DONOR
Broad+S	broad focus	NarrowS
Broad+V	broad focus	NarrowV
NarrowS+B	NarrowS	broad focus
NarrowV+B	NarrowV	broad focus.

Tab. 5 Combinations of Base and donor utterances used to create the hybrid stimuli.

The results were averaged across subjects. Overall, statements present a mean score that is never greater than 4, while questions have higher values. The effect of modality was nearly significant in the two-way interaction with manipulation.

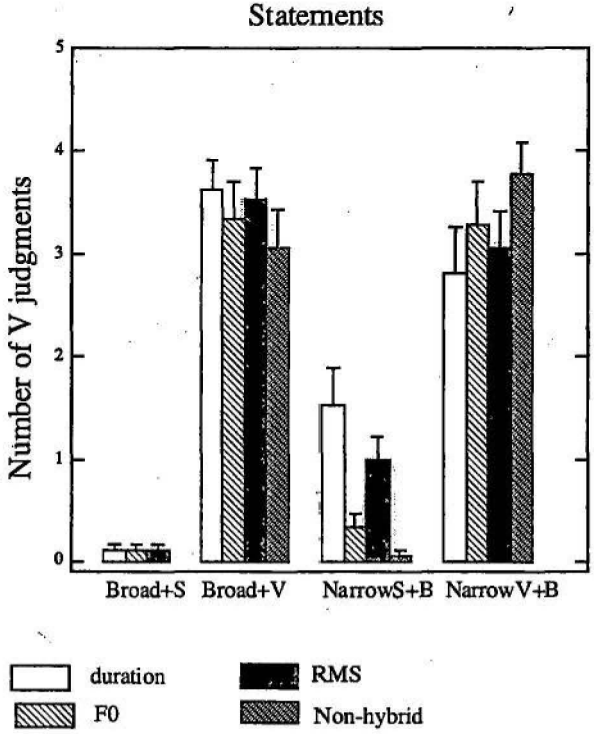

Fig. 3 Mean values for "Number of V judgments" for all speakers across focus types for statements. Manipulation levels are indicated by different bar patterns.

Figure 3 shows the mean values of the dependent variable across different focus types. The three acoustically manipulated patterns scored equally well in the Broad+NarrowS (Broad+S) manipulation, with a mean V judgement score of 0.09. This value was very close to the non-hybrid focus-S pattern, which was 0.05. In other words, all three acoustic correlates successfully displaced perceived prominence from the verb to the subject position. This might be due to the high sensitivity to the beginning of the utterance that has already been found in perception of natural utterances with varying focus position (D'Imperio, 1997a). This triplet must be contrasted to the natural broad focus stimulus in the Broad+NarrowV series. Broad focus stimuli were conventionally grouped with stimuli in which a prominence shift towards the verb was expected. In standard phonological theory, broad focus sentences have late prominence and no naturally occurring broad focus utterances have focus on S.

The Broad+NarrowV manipulation scored in the opposite direction. The three acoustically manipulated patterns successfully reinforced the perceived prominence on the verb for non-hybrid broad focus stimuli. Among the cues, duration scored a slightly greater number of V judgments (3.6), followed in strength by amplitude (3.5) and F0 (3.05). Though all three acoustic correlates seemed to reinforce the perceived prominence on the verb position, only duration did so significantly as a result of planned comparisons with the non-hybrid stimulus score..

The results of the NarrowS+Broad category are particularly interesting in that they show a substantial difference in the patterning of the various non-hybrid stimuli. Only duration and RMS succeeded in displacing perceived prominence from the subject to the verb position. Remarkably, duration is the strongest cue in this manipulation, with a mean of 1.5. Even though the effect of this manipulation does not appear unusual at a first glance, it acquires a different meaning when considering that the maximum value reached by non-hybrid broad focus utterances was only 3.05. The weakest cue appears to be F0, with a 0.3 mean, preceded by amplitude, which scored a mean equal to 1. As expected, it is more difficult to shift perception when the original utterance has prominence on the first element of the utterance. As to the non-hybrid NarrowS stimuli, only in a mean of 0.05 utterances did listeners assign prominence to the verb.

The NarrowV+Broad manipulation appeared to revert the pattern established in the previous category. At a first approximation, we notice that the highest mean score among the hybrid stimuli was found for the F0 manipulated stimuli. However, a successful shift from a narrow focus to a broad focus pattern needs to "lower" the prominence at the verb position. In fact, for natural broad focus stimuli the verb location receives low scores of perceived "importance" (see D'Imperio, 1997a). Also, non-hybrid broad focus utterances showed a modest mean of 3.05, which is barely above chance. Therefore, a stronger effect will translate into a smaller number of utterances with assigned prominence to the verb. Among the correlates, only duration displaced prominence from the verb to the subject in a significant way. In other words, the duration manipulation appears to affect the stimuli in a way that they tend to assume the uncertain prominence pattern already recorded for natural broad focus utterances (D'Imperio, 1997a). The results for stimuli with duration manipulation indeed show a mean score of 2.8, which goes in the direction of a weaker prominence on the verb. The F0 manipulation was the least different from the non-hybrid NarrowV manipulation. Amplitude (RMS) results are intermediate between the other two manipulations.

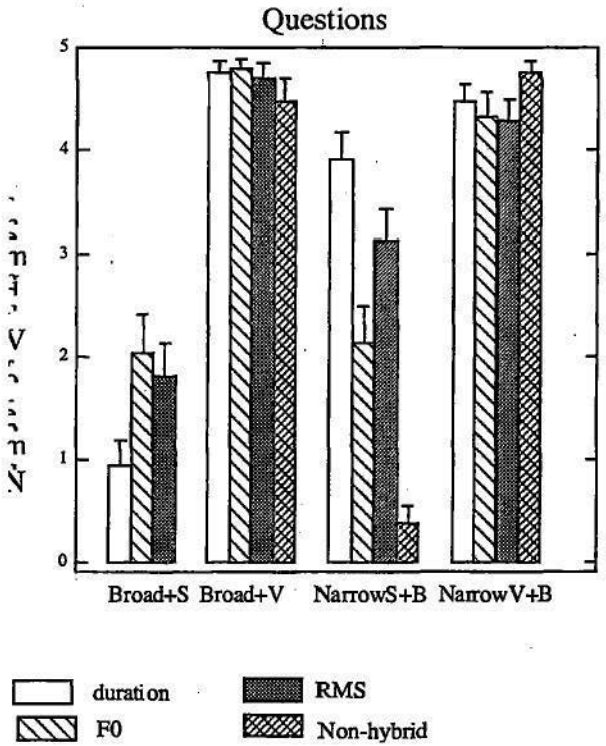

Fig. 4 Mean values for "Number of V judgments" for all speakers across focus types for questions. Manipulation levels are indicated by different bar patterns.

4.2 Questions

As Figure 4 shows, the role of duration and RMS was quite remarkable, at least in some manipulations. Within the Broad+NarrowS hybrid manipulation, duration was strongest in displacing perceived prominence from the verb to the subject position (lower bars indicate low scores of V responses and, as a result, high scores of S responses). After duration, the second highest effect is due to amplitude, followed by F0. Although it was the least effective cue, F0 scored better than chance in shifting prominence perception.

The Broad+NarrowV manipulation presents an interesting "tie" among the acoustic cues. All three hybrid levels seemed to score marginally better than the non-hybrid, broad focus stimulus, as expected, since each cue has the effect of reinforcing the perceived prominence on the verb. However, this was only a non-significant trend. Non-hybrid stimuli scored a mean of 4.5, which was lower than the 4.8 scored by non-hybrid stimuli with NarrowV focus. All hybrid stimuli registered a mean score that is very close

to the non-hybrid NarrowV stimuli, i.e. from 4.71, (RMS manipulation) to 4.81 (F0 manipulation).

For the NarrowS+Broad manipulation, duration again was more salient than the other acoustic cues. In fact, this cue had the biggest effect in displacing prominence from the subject to the verb. This pattern is strikingly similar to the NarrowS+Broad manipulation in statements, even though it is reproduced here on a greater scale. Duration was able to shift prominence perception of the original NarrowS utterances from the subject to the verb, with a score of 3.9. This score compares to 3.1 for the RMS manipulation. The difference that we observe in the magnitude of this effect in questions, as opposed to statements, might be due to a peculiarity of early focus questions. Unlike statements, early focus questions in Neapolitan Italian present a postnuclear pitch peak on the last stressed syllable of the utterance. (!H* in Figure 2, middle and lower panel). We observe here that the conflicting cues represented by a very strong pitch accent on the subject and a weak pitch accent on the verb is resolved by duration. Duration appears to be capable of compensating for the lack of tonal prominence on the part of the nuclear pitch peak in quite an effective way. F0 manipulation was again found to be the poorest correlate in this pattern, yielding a mean score that is near chance (2.1). In the non-hybrid NarrowS manipulation the mean is quite low (0.4), as expected for this pattern, even though it is greater than the mean we found in the same category for statements (0.05). This might be due to a bias for questions to receive late prominence identification due to the conspicuous pitch accent that characterizes them.

Finally, for the NarrowV+Broad category, the results are similar to those of the Broad+NarrowV manipulation. All three correlates appear to weaken perceived prominence on the verb, but none of them did so significatively. However, within this group, duration appears to have a slightly stronger effect. As mentioned above, the non-hybrid manipulation yielded a mean of 4.8.

5. Discussion

The results appear to support the hypothesis that duration is an important correlate of prominence in Italian, not only at the word level (Bertinetto, 1980), but also at the sentence level. At least for two manipulations, i.e. Broad + NarrowS donor and NarrowS + Broad donor, duration is the correlate that has the biggest impact in displacing perceived prominence. In all of these manipulations, F0 is the weakest cue, which parallels Bertinetto's findings[6].

The results provide strong support for the idea of a trading relation among acoustic cues in the perception of prominence. The hybrid Broad+NarrowS manipulation completely reverted the prominence pattern of broad focus base utterances, for instance. Such manipulation had the effect of making listeners assign prominence to the subject most of the time, in both question and statement stimuli. Moreover, when NarrowS base questions where combined with a broad focus question as a donor (NarrowS+Broad manipulation), prominence was significantly shifted to the verb (except for the F0

[6]Bertinetto's view of duration contribution has to be seen in the right perspective, though: "Thus, although D undoubtedly bears the greatest importance in the determination of perceptive responses concerning prominence, this component must not be viewed separately from the others. When certain conditions are met, the combined effects of I and F0 may in fact exceed the weight of D" (Bertinetto 1980, p. 392).

manipulation). The hybrid manipulation also had the effect of reinforcing prominence on the verb in the Broad+NarrowV manipulation and reducing it in the NarrowV+Broad manipulation. This result was true only for statements and was expected from the typical prominence responses to non-synthetic broad and late narrow focus stimuli (D'Imperio, 1997a).

The NarrowS+Broad manipulation had a bigger overall effect in interrogatives. This is probably due to the different postnuclear contour of early focus interrogatives as opposed to early focus declaratives (D'Imperio, 1997b).

The present results can be compared to Beckman's (1986) results for American monolingual subjects. When separately looking at amplitude and duration in the American-monolinguals results, duration was more effective than amplitude. However, in Beckman (1986) the most effective cue overall was F0.

One outcome of the present experiment that cannot be compared to previous studies is the effect of modality. Especially interesting is the comparison between Broad+NarrowS statements and questions. While in the statements all three manipulations produced a very strong effect, in the questions they did not. In fact, F0 and intensity did not succeed in shifting prominence perception in this condition as successfully as duration did. This outcome can be explained by the fact that, unlike statements, NarrowS questions present a late postnuclear pitch-accent (see middle and lower panel of Figure 2; see also D'Imperio, 1997b). In this case, switching the melodic contour of a NarrowS question has the effect of slightly decreasing the percept of a tonal event on the verb, which could account for the weaker effect of F0. In statements, the melodic contour of a NarrowS utterance has no postnuclear tonal markings (see Figure 1, middle and lower panel); therefore, no late tonal event can attract perceptual prominence.

The effect of duration in the NarrowS+Broad questions is even more surprising in the light of what we know about preboundary lengthening (Beckman and Edwards 1990), by which the phrase-final section of an utterance is lengthened. Just as it appears that listeners can factor out the gradual declination of F0 in the course of an utterance, it is also expected that they would perceptually adjust for longer utterance portions in the proximity of a boundary. However, this was not the case in the question results. The percept of a longer verb constituent made it perceptually more prominent than the pitch prominent subject. In this case, the duration of the stressed vowel in the first syllable of *esce* traded for the lack of a perceptually strong pitch accent for the purpose of signaling prominence on that word. The strength of the duration manipulation is further supported by the higher consistency in the results for this manipulation as opposed to the F0 and RMS manipulation (see § 4.2 above).

What these results mean for traditional trading relation hypotheses is difficult to say for a number of reasons. First, most of the literature on the topic of the last decade has concentrated on segmental features[7], like the feature [voice] or manner features such as [fricative] (see Repp 1982 for a review). Auditory integration can be evoked to explain the trading relation by appealing to psychophysical properties of the auditory system

[7]We also know that prominence (or stress) is not a feature, at least not in the sense as [+ voice] or [-velar] are. Since Liberman (1977) our view of metrical strength has changed from being an absolute, categorical, value (as in Chomsky and Halle, 1968) to a relational dimension between terminal elements in a structure. It may well then be that it is not possible to easily generalize from feature perception to prominence perception and that the two fields have to be kept apart.

(Kingston and Diehl, 1995). This position is severely criticized by motor theory supporters such as Repp (1982), who claims that (p. 93) "In most other cases, [however], the cues that participate in a trading relation are simply too diverse or too widely spread out to make auditory integration seem plausible" (brackets inserted by the author). The extreme position represented by Repp is one that simply denies cue integration as a generic auditory process and which, instead, regards it as yet another proof of the existence of a specialized phonetic mode of speech perception. In this perspective, trading relations among acoustic cues could only occur "because listeners perceive speech in terms of the underlying articulation and resolve inconsistencies in the acoustic information by perceiving the most plausible articulatory act" (Repp, 1982: p. 95).

In order to support a speech-specific view of trading relations in the realm of prosody, motor theorists can appeal to the results of works such as Smith (1978), cited in Repp (1982). In this study, relative duration of two subsequent syllables was varied and two types of judgments were elicited from the subjects, one linguistic (stress position) and the other auditory (which syllable was longer). It was found that subjects had a first syllable bias only when they were performing the linguistic task. The explanation given to account for the bias is that, when listeners are in a "speech mode" of perception, they expect the second syllable to be longer because of the speech specific phenomenon known as final lengthening. In other words, when perceiving the stimuli as speech, longer duration in the second syllable is not as strong a cue as in the first syllable, hence a first syllable bias in the responses. Bertinetto's (1980) results point to something similar, though in the opposite direction. In this work, subjects showed a second syllable instead of a first syllable bias. The suggested explanation is that they might have adjusted for the intrinsic shorter length of final stressed syllables reported in Italian production studies.

A speech-specific interaction of prosodic correlates of stress is argued against by Beckman (1986). On one hand, in her results intensity and duration could be seen as being in a trading relationship because of their common articulatory origin: an augmented jaw movement can result in a longer as well as in a more intense acoustic signal[8]. However, drawing from psychoacoustic literature on temporal summation of loudness, Beckman proposes that the special relationship between intensity and duration has an auditory and not simply an articulatory basis. Loudness, in fact, appears to be the result of the combined effect of intensity and duration over a segment (see psychoacoustic literature cited in Beckman 1986). The claim that duration contributes to the loudness percept in speech has been recently opposed by Sluijter *et al.* (1997): "It has only been established for pure tones of a relatively short duration that differences in duration are responsible for differences in the perception of loudness." (p. 511). Sluijter *et al.* argue instead for a relevant effect of intensity manipulations over high frequency regions of the spectrum. Mere intensity level (i.e. affecting the entire spectral range) variations are regarded, instead, as having no "communicative significance" because of their vulnerability to environmental masking. Intensity is, in fact, highly affected by environmental noise, position of the mouth, intervening obstacles, etc. However, the role of intensity level as expressed by RMS amplitude in the present study cannot be entirely

[8]Another complication of duration as a cue to stress is due to its ambiguous articulatory origin. In Articulatory Phonology terms, longer duration can be a result of either reduced stiffness in the gesture or a result of changes in intragestural phasing (Browman and Goldstein, 1990). Our data cannot say anything about this matter, since it is impossible to differentiate between the two hypotheses on

dismissed. Amplitude manipulations appeared, in fact, to have an effect that was even stronger than the F0 manipulation in the majority of cases.

It is interesting that in the present experiment conflicting cues did not give rise to extremely confused results, as direct realism theory would predict (see Fowler, 1996 discussion of the results of Fitch *et al.*, 1980). Since no discrimination task followed the forced identification, no strong argument against this view can be provided at this point. However, the findings presented here appear to speak against a direct realist view of speech perception for additional reasons. If prominence is directly perceived, we would have to postulate a unique articulatory gesture decoded from the acoustic proximal event. The problem is that while it is somehow possible to postulate a common origin of intensity and duration variations, it is more difficult to reconcile the articulatory production of these last two cues with fundamental frequency production. In other words, both increased laryngeal activity and jaw opening, say, should be both translated back to the linguistic category "prominent". Should we favor a more abstract motor theoretic approach, we could hypothesize that what listeners do is decode some kind of speech "effort" localized on the prominent syllables (see de Jong (1995) for articulatory characteristics of stressed syllables). This "effort" can be translated back to either neural commands for jaw opening, subglottal pressure increase or greater laryngeal activity, or, alternatively, to a combination of them.

It seems to me that the best explanation for the data presented here is the "strong auditorist" perspective represented by works such as Kingston and Diehl (1995). This view entails that some acoustic properties cohere not just when sharing a common articulatory origin, but also when producing the same auditory effect. In other words "certain acoustic correlates of a phonological distinction are integrated into perceptual properties that enhance contrasts" (Kingston and Diehl, 1995, p. 24). It may also be that cues are integrated into an *intermediate perceptual property* (IPP), which in this study would be the percept of something being *prosodically stronger*. That the cues enhance each other is proven by the results of the broad focus + late (narrow) focus manipulations. In order to prove the soundness of the theory, we would need to perform a test where synthetic stimuli, sufficiently different from speech, would be used. Moreover, we would still have to account for the language-specific nature of the postulated IPP level. An alternative proposal, as suggested in Nearey's commentary on Kingston and Diehl's paper (Nearey, 1995) is that the IPPs are actually relevant only in the process of language acquisition and that we need not postulate them as independent levels in the representation. The problem is that our knowledge of psychoacoustic cue integration cannot be easily applied to language (for instance, one cannot easily extend the findings on pure tone perception; cf. Sluijter *et al.* 1997 criticism presented above).

6. Conclusion

The hybridization method appears to successfully affect perceived prominence in Italian. Specifically, duration appears to have a dominant role when the "donor" and "recipient" utterance have different accent structure (as in Broad+NarrowS manipulations. Differences in overall accent structure between questions and statements seem to determine differences in the effect of the manipulation. Our results present a

problem for theories where pitch is the primary correlate of prominence[9]. The results support a view by which duration is an active prominence cue in nuclear stress perception in Italian, and, more broadly, represent a crucial step towards understanding the interplay of language-specific acoustic correlates of stress.

REFERENCES

Bartels, C. and Kingston, J. (1995). *Salient pitch cues in the perception of contrastive focus*, in Dickey, M.W. and S. Tunstall (Eds.) UMOP 19, pp. 1-25.

Beckman, M.E. (1986). *Stress and Non-stress Accent.* Dordrecht, Foris Publications.

Beckman, M.E. and Ayers, G. M. (1994). *Guidelines for ToBI Labelling.* Unpublished manuscript, Ohio State University. [Send email to tobi@ling.ohio-state.edu for ordering information, or visit the English ToBI homepage at http://ling.ohio-state.edu/Phonetics/etobi_homepage.html].

Beckman, M.E. and Edwards, J. (1990). *Lengthenings and shortenings and the nature of prosodic constituency*, in J. Kingston and M.E. Beckman (eds.), *Papers in Laboratory Phonology II*, Cambridge, CUP, pp. 152-178.

Bertinetto, P.M. (1980). *The perception of stress by Italian speakers*, J. of Phon., 8, pp. 385-95.

Bertinetto, P.M. and Fowler, C.A. (1989). *On sensitivity to durational modifications in Italian and English*, Rivista di Linguistica, 1, 1, pp. 69-94.

Boves, L., Ten Have, B.L., Vieregge W.H. (1984). *Transcription of Intonation in Dutch*, in Gibbon., D. and H. Richter (eds.), *Intonation, Accent and Rhythm*, Berlin, De Gruyter, pp. 20-45.

Browman, C.P. and Goldstein, L. (1990). *Gestural specification using dynamically-defined articulatory structures*, J. of Phon., 18, pp. 299-320.

Campbell, W.N. (1995). *Loudness, spectral tilt and perceived prominence in dialogues.* Proc. ICPhS 95, vol. 3, pp. 676-679.

D'Imperio, M. (1997a). *Breadth of focus modality and prominence perception in Italian.* OSU Working Papers in Linguistics, 50, pp. 19-39.

D'Imperio, M. (1997b). *Narrow focus and focal accent in the Neapolitan variety of Italian.* Proceedings of an ESCA Workshop on Intonation, Athens, pp. 87-90.

D'Imperio, M. (1998). *Prominenza accentuale, focus e modalità intonativa nella percezione di parlato italiano letto.* In Proceedings of the "VIIIe Giornate di Studio del Gruppo di Fonetica Sperimentale (GFS)", December 18-20, Pisa, Italy.

de Jong, K.J. (1995). *The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation*, JASA, 97, pp. 491-504.

[9]Beckman (1996), p. 38 "For example, accented vowels tend to be longer and articulated closer to the periphery of the vowel space (see de Jong, 1995, for a review and some recent data). However, these are minor variations compared with the qualitative difference between inherently longer full vowels and inherently very short reduced vowels that categorically defines the stress contrast between heavy and light syllables at the lowest level of the stress hierarchy, and could be called ancillary to the tonal markers (Beckman & Edwards, 1994). Thus it is not possible to talk about stress at the two higher levels without explicitly or implicitly assuming an intonational pattern for an actual or imagined utterance of the text" (the boldface is mine).

Farnetani, E. and Kori, S. (1983). *Acoustic manifestation of focus in Italian*, "Quaderni del Centro di Studio per le Ricerche di Fonetica", 2:287-318.

Fowler, C. (1996). *Listeners do hear sounds, not tongues*, JASA, 99 (3), pp. 1730 41.

Fry, D.B. (1955). *Duration and intensity as physical correlates of linguistic stress*. JASA, 23, pp. 765-769.

Fry, D.B. (1958). *Experiments in the perception of stress*. Language and Speech, 1:126-152.

't Hart, J., Collier, R. and Cohen, A. (1990). *A perceptual study of intonation*. Cambridge, England, CUP.

Hermes, D.J. and Rump, H.H. (1994). *Perception of prominence in speech intonation induced by rising and falling pitch movements*. JASA, 96 (1), pp. 83-92.

Hirschberg, J. and Ward, G. (1992). *The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in English*, J. of Phon. 20, pp. 241-51.

Kingston, J. and Diehl, R.L. (1995). *Intermediate properties in the perception of distinctive feature values*, in Connell, B. and A. Arvaniti (eds.) *Papers in Laboratory Phonology IV*, Cambridge, CUP, pp. 7-27.

Ladd, R.D., Verhoeven, J. and Jacobs, K. (1994). *Influence of adjacent pitch accents on each other's perceived prominence: two contradictory effects*, J. of Phonetics, 22:87-99.

Lehiste, I. (1970). *Suprasegmentals*. MIT, Cambridge, MA.

Lehiste, I. and Fox, R.A. (1993). *Influence of duration and amplitude on the perception of prominence by Swedish listeners*, Speech Communication 13, pp. 149-54.

Liberman, M. and Pierrehumbert, J.B. (1984). *Intonational invariance under changes in pitch range and length*, in M. Aronoff & R.T. Oehrle (eds.), *Language Sound Structure: Studies in phonology*, 157-233, Cambridge, MA: MIT Press.

Lieberman, P. (1960). *Some acoustic correlates of word stress in American English*, JASA, 22, pp. 451-454.

Nakatani, L. and Aston C. (1978). *Acoustic and linguistic factors in stress perception*. Unpublished manuscript, Bell Laboratories.

Nearey, T. M. (1995). *A double-weak view of trading relations*, in Connell, B. and A. Arvaniti (eds.) *Papers in Laboratory Phonology IV*, Cambridge, CUP, pp. 28-40.

Pierrehumbert, J.B. (1980). *The phonology and phonetics of English intonation*. Doctoral dissertation, MIT, Indiana University Club.

Pierrehumbert, J.B. and Beckman, M.E. (1988). *Japanese Tone Structure*. Cambridge, MA, MIT Press.

Repp, B.H. (1982). *Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception*, Psychological Bulletin, vol. 92 (1), pp. 81-110.

Sluijter, A.M.C., van Heuven, V.J. and Pacilly, J.J.A. (1997). *Spectral balance as a cue in the perception of linguistic stress*, JASA, 101 (1), pp. 503-13.

Terken, J. (1992). *Fundamental frequency and perceived prominence of accented syllables*. JASA, 89 (4), pp. 1768-76.