PERCEIVED TONE "TARGETS" AND PITCH ACCENT IDENTIFICATION IN ITALIAN

Mariapaola D'Imperio†[1], Jacques Terken‡ and Michel Pitermann*[2]
† Department of Linguistics, The Ohio State University – USA
and LORIA – France
‡ IPO, Center for User-System Interaction – Eindhoven, The Netherlands
*Queen's University, Canada
dimperio@loria.fr, j.m.b.terken@tue.nl, piterman@loria.fr

ABSTRACT: This study investigates the role of temporal alignment, f_0 and peak shape in determining perceived tonal target values in Neapolitan Italian. In this variety, the alignment of the accent peak appears to be a strong perceptual cue to the question/statement identification (D'Imperio and House, 1997), everything else being equal. In the present study, the f_0 contour of a question, uttered by a female speaker of Neapolitan Italian, was stylized and resynthesized by means of PSOLA. A set of stimuli was created in which either tonal alignment was varied, while f_0 height was kept constant, or f_0 height was varied orthogonally to alignment. For the alignment manipulation, an additional variable was the shape of the accent peak, which could be either flat (creating a short plateau) or sharp. Thirty Neapolitan subjects listened to the stimuli and identified each as a question or a statement. The results suggest that the contribution of f_0 peak height to the question/statement identification is much less important than that of target alignment. Moreover, peak shape affects the perceived alignment of the target tone, in that flat peak stimuli cause the perceived target to be displaced towards the end of the plateau.

INTRODUCTION

In Italian, a question can be signaled by intonational means alone (Avesani, 1990). Also, the shape of both question and narrow focus statement pitch accents of the Neapolitan variety is very similar, being characterized by a rise-falling contour(D'Imperio, 2000a), which is usually realized as a sequence of a L, a H and a L tone (LHL). However, the alignment of the accent peak appears to be a strong perceptual cue to the question/statement identification (D'Imperio and House, 1997) and is systematically manipulated in production (D'Imperio, 2000b). Another plausible cue to the question/statement distinction might be the fundamental frequency (f_0) level of the accent peak. For instance, Gósy and Terken (1994) found that higher accent peak cue more questions in the perception of Hungarian synthetic stimuli. Hence, the target(s) measured in the fundamental frequency curve, and that generally corresponds to Highs and Lows, might be mapped in specific ways to the target(s) perceived as such by the listener.

In a previous study (D'Imperio and House, 1997), the f_0 contours of a declarative and an interrogative utterance produced by a female Neapolitan speaker were stylized and then resynthesized by means of PSOLA (Moulines and Charpentier, 1990). Four stimulus series were created from the two original utterances. The two interrogative base series differed as to the shape of the final rise-fall, which had a sharp peak in one series and a flat peak in the other. Within the sharp peak series (inter-peak), the timing of the peak was shifted by 35 ms steps within the stressed vowel, while for the flat peak series (inter-rise) it was the rise that was shifted throughout the vowel. Nineteen Neapolitan listeners identified the stimuli as either questions or statements. The results for the inter-peak series showed that earlier peak alignment cues more statement responses. For the flat peak stimuli, the implicit assumption was that the peak of flat stimuli would correspond to the time coordinate of the end of the rise. Consequently, we would expect statements to be identified at early peak locations, just as for peak stimuli. Surprisingly,

[1]Partial support for this study came from special funds from the University of Eindhoven (UniversiteitFunds Eindhoven) and an Ameritech Fellowship to the first author.
[2]This author was supported by NIH Grant No. DC–00594 from the National Institute of Deafness and other Communication Disorders and NSERC.

though, these stimuli scored a majority of question responses already at early continuum locations. However, peak target location for flat peak stimuli was neither explicitly defined nor tested. Also, the effect of another potential cue to the question/statement contrast, i.e., f_0 level, was excluded from the study by normalizing f_0 level in the original statement and question utterances. One of the goal of this study was to gauge temporal localization of perceived targets in such tough cases. Note also that, within the psychoacoustic literature, it has been found that if a tone glissando is followed or preceded by a plateau, the perceived pitch value will correspond to the pitch of the plateau Nábeläk et al. (1970). Thus, we need to first replicate the results of D'Imperio and House (1997) regarding the timing manipulation for the inter-peak series, in which the alignment of peak stimuli was shifted through the stressed vowel. Then, we will use those results as a reference for further comparisons, such as a comparison with flat peak (plateau) stimuli[3] and a comparison with stimuli with the same peak shape but different f_0 level.

First, we tested the hypothesis (hypothesis 1, see upper left panel of Figure 1) that when the LHL configuration is moved backwards within the stressed vowel, a higher percentage of statement responses should be obtained. Then we tested the hypothesis (hypothesis 2, see upper right panel of Figure 1), that the number of question responses will be the same for plateau stimuli and peak stimuli whose peak is timed at plateau onset. If the hypothesis is not verified, then we will test the alternative hypothesis that the perceptual target location for plateau stimuli corresponds to the the offset of the plateau, i.e., the beginning of the HL fall. This will be done by comparing scores for peak stimuli with peak timed at plateau offset (hypothesis 2b, lower left panel of Figure 1). But targets are also specified melodically, and not only temporally. Hence, it is plausible that the observed differences are also due to changes in target specification relative to the pitch domain (Gósy and Terken, 1994). Hence, an additional hypothesis was tested (hypothesis 3) stating that perceived differences in pitch target level will induce different scores of question/statement identification, independent of timing (see lower right panel of Figure 1). If the prediction is borne out, this will suggest that the potential difference in score between plateau and peak stimuli cannot be attributed to a pitch level difference for the perceived target. Our aim is to discover the mechanism of perceived target mapping in all conditions. That is, we aim to assess perceived tonal target timing indirectly, through the question/statement answer paradigm. In sum, we first, we need to show that questions are indeed identified with the percept of later target timing in those configurations where the peak has an obvious manifestation, such as sharp peak contours. Then, once the timing relationship is defined, we can use it to determine if accent configurations with less clearly definable peaks, and with a specific timing relationship to the sharp peaks, will be associated to different perceived targets.

METHODS

The base stimulus used to create the stimulus continua was selected from a set of read sentences produced by a native speaker of Neapolitan Italian (the first author). This was the utterance *Vedrai il nono?* "Will you see the ninth?", produced with a L*+H nuclear accent on *nono*. The natural base question stimulus was also included in the perception experiment. The results of the natural stimulus identification were then employed to set a criterion for listener inclusion in the data analysis, in that listeners showing less than 80% question responses for the natural stimulus were assumed to be unable to perform the task in a reliable fashion. The f_0 contour of the base utterance was extracted and its contour modified with the help of PIOLA, which has characteristics that are similar to PSOLA, within the GIPOS software developed at IPO, Netherlands. Once the f_0 shape was altered with new target values (both timing values and/or f_0 values), the utterance was resynthesized.

In order to test hypothesis 1, a continuum was created in which the timing of the three tonal targets within the LHL sequence, i.e., L1, H and L2, was simultaneously manipulated. Seven stimuli were thus created, which will be referred to as the "primary continuum" (see Figure 2, left). The fundamental frequency and timing values of this continuum are shown in Table 1. The f_0 values were taken from actual target values within the natural question used as a base stimulus. The choice of the specific timing region for the construction of the continuum was based on timing observations in production (D'Imperio, 2000b). The three target locations (L1, H and L2) were simultaneously shifted backwards, in 15 ms steps, from the original values of the natural utterance. By stretching the region between such values at both extrema we expect to obtain an S-shaped response curve, with near 100% question or statement responses at either of the extrema, and highest response ambiguity around the center of the continuum.

[3]Note that in the production of Neapolitan Italian utterances, instances of plateaus are found for both statement and question utterances (D'Imperio, 2000b), therefore we do not assume that such a phenomenon can cue a specific modality *per se*. Also, we assume that short plateaus are perceived in terms of a single H target, and that parsing two separate targets is not possible.

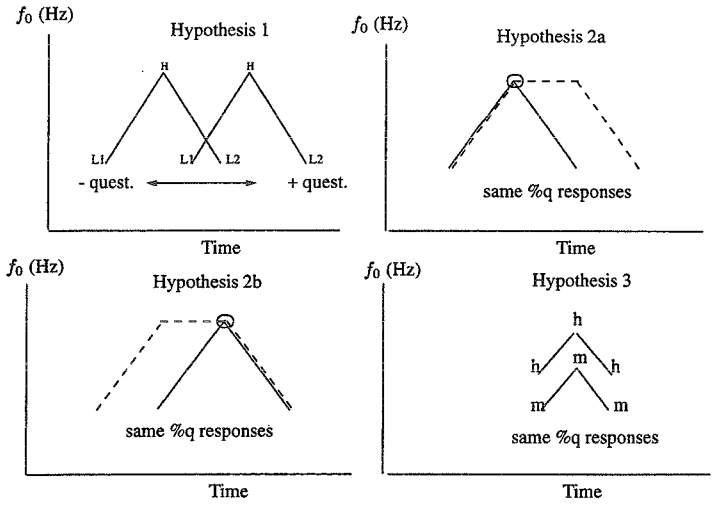

Figure 1: Schematic representation of the hypotheses. The dotted line represents plateau stimuli.

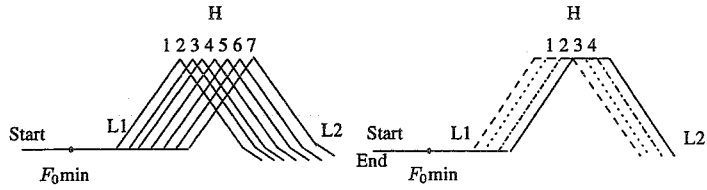

Figure 2: Structure of primary continuum stimuli (left) and of plateau continuum stimuli (right).

To test hypothesis 2, we created a "plateau" continuum, which consisted of a series of 4 stimuli whose shape was characterized by a 45 ms f_0 plateau (see Figure 2, right). Plateau duration was therefore equal to three steps of the primary continuum. This allowed us to verify hypotheses 2a and 2b stating, respectively, that perceived peak target location corresponds with either LH rise offset (i.e., the plateau beginning) or HL fall onset (plateau end). Such tests were performed by comparing, first, the results for plateau stimuli and stimuli with peaks timed at plateau onset (i.e., primary continuum stimuli T1, T2, T3 and T4). Then, to test hypothesis 2b, we compared results for plateau stimuli with stimuli with peaks timed at plateau offset (i.e. primary continuum stimuli T4, T5, T6 and T7). In order to test hypothesis 3, we created a continuum in which f_0 height of L1, H and L2 targets was varied between medium values ("mmm"), high values ("hhh") and low values ("lll"). f_0 level was simultaneously varied for all 3 targets (L1, H and L2) at 3 timing locations corresponding to T1, T4 and T7 of the primary continuum. I will refer to this continuum as the "f_0 continuum". The precise f_0 values for each combination are reported in Table 2.

The stimuli were presented binaurally through headphones in a studio at the University "Federico II" of Naples. The listeners, who were thirty Neapolitan speakers, were instructed to perform a two-alternative forced choice task, in which they had to identify the stimulus heard as either a question or statement. They were also told to report the answer by crossing a "d" (*domanda* "question") or "a" (*affermazione* "statement") box on an answer sheet. After ten practice trials, the stimulus group was played 5 times in 5 differently randomized blocks with each stimulus occurring once per block.

Tone target	f_0 value (Hz)	Latency from vO (ms)
L1	210	-60,-45,-30,-15,0,+15,+30
H	280	+60,+75,+90,+105,+120,+135,+150
L2	180	+180,+195,+210,+225,+240,+255,+270

Table 1: f_0 values and latency from v0 (vowel onset) for L1, H and L2 in the primary continuum (minimum and maximum values are given).

	L1 f_0(Hz)	H f_0(Hz)	L2 f_0(Hz)
mmm	210	280	180
hhh	230	300	200
lll	190	260	160

Table 2: f_0 values for L1, H and L2 within the f_0 continuum.

RESULTS

Primary continuum

Scores were calculated as mean values for each listener. The left panel of Figure 3 shows mean question scores pooled for all listeners (y axis) across stimulus Time Step (x axis). As expected from the results for the "inter-peak" continuum reported in D'Imperio and House (1997), shifting the L1-H-L2 configuration backwards within the accented vowel decreased the number of question responses. Since the variance was not constant throughout the timing continuum (as the standard error bars in Figure 3, left, show), we performed a one-way Analysis of Variance on the arcsine transformed data, with Timing as a factor.[4] The timing manipulation yielded a significant result, as expected [F(6,175) = 61.11; $p < 0.01$].

Plateau continuum

Once we have confirmed the impact of LHL timing on the perception of a specific target, which is in turn associated with question or statement meaning, let us look at the results for the plateau continuum. Those data will be employed to test hypothesis 2 regarding the impact of peak shape on perceived target location. The right panel of Figure 3 presents mean question results for the plateau continuum. These results appear immediately strikingly different from those of the primary continuum (Figure 3, left). What we notice is that a great number of question responses was already obtained early in the continuum. Stimulus 2, for instance, presents a score that is already well above chance (0.75). Then, in order to test hypothesis 2a, stating that T1 peak stimuli would score the same percentage of question responses as T1 plateau stimuli, we performed a two-way ANOVA on the arcsine transformed results for plateau stimuli and peak stimuli from T1 to T4, with Timing and Shape (peak or plateau) as independent variables. The results were significant for both Timing and Shape manipulation [Timing: $F(3, 200) = 31.73; p < 0.01$; Shape: $F(1, 200) = 132.7; p < 0.01$]. The interaction resulted not to the be significant, though in this case the p value was quite close to the cutoff point [$F(3, 200) = 2.98; p = 0.033$]. Hence, the results made us reject hypothesis 2a.

We then tested hypothesis 2b, i.e., that peak stimuli with peaks timed at the plateau offset would receive

[4]The transformed scores (W) were obtained through $W = arcsin\sqrt{\frac{X}{100}}$, where X is the original score.

Plateau Stimulus	Mean score	Primary Stimulus	Mean Score
T1	0.47	T4	0.58
T2	0.75	T5	0.82
T3	0.90	T6	0.95
T4	0.88	T7	0.97

Table 3: Mean scores for the plateau continuum (pooled results, left) and mean primary continuum scores for stimuli with peaks timed at plateau offset (right).

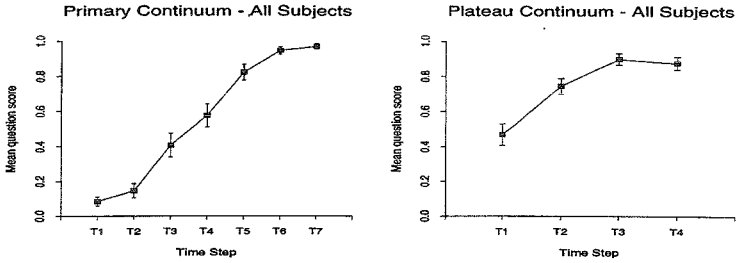

Figure 3: Primary continuum mean scores (left) and plateau continuum scores (right) for all listeners. Standard error is indicated by vertical bars.

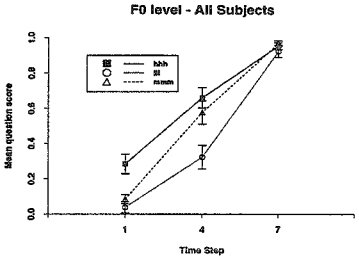

Figure 4: Mean scores for the f_0 continuum (pooled results). Standard error is indicated by vertical bars.

the same question scores as plateau stimuli. However, the results of a two-way ANOVA on the arcsine transformed results for all plateau stimuli and peak stimuli from T4 to T7 were still significant [Timing: $F(3,200) = 32.97; p < 0.01$; Shape: $F(1,200) = 8.98; p < 0.01$]. The interaction of the main effects was not significant, instead $[F(3,200) = 0.18; p = 0.9]$. The results therefore show that the perceived target of a plateau stimulus corresponds to neither that of a peak stimulus with peak timed at the plateau onset nor to that of a peak stimulus timed at the plateau offset. Therefore, also hypothesis 2b is to be rejected. The results, however, still show a visible trend for a greater similarity between plateau scores and scores for peak stimuli timed at plateau offset (see Table 3). The next question is if plateau stimuli are perceived as simply having a higher pitch than the corresponding peak stimuli.

f_0 continuum

In Figure 4 mean question scores are plotted against Time step, and separately for each of the target f_0 height combinations. Note that, apart from the scores at T1, the mean scores for hhh and mmm do not seem very different from each other. A two-way ANOVA was then run on the arcsine data for the three continua, with Time Step and F0 level as independent variables. The results show a significant overall effect of Time Step $[F(2,225) = 222.5; p < 0.01]$ as well as an effect of F0 level $[F(2,225) = 9.4; p < 0.01]$. The interaction was not significant $[F(4,225) = 2.7; p = 0.033]$. However, though the f_0 level manipulation was significant, the results of a post-hoc analysis (Tukey, confidence interval = 0.01) showed that the only significant difference was the one between lll and hhh, while neither the hhh/mmm nor the mmm/lll comparison were crucially different from each other. Therefore, the results suggest that the f_0 level effect is relatively small and that bigger differences seem to be needed in order to show such an effect.

DISCUSSION

The perception experiment presented here aimed at testing some hypotheses about tonal target perception in Neapolitan Italian using a question/statement answer paradigm. First, we examined the responses of the listeners to a manipulation of the three main targets of the rise-fall configuration in the time domain, through the use of resynthesized stimuli. We found that such a manipulation can indeed shift the perception of a question to a statement and took this to mean that the perceived location of the rise-fall peak is shifted in the time domain from a "late" to an "early" value. This aspect of the data serve to reconfirm the results of D'Imperio and House (1997), regarding the inter-peak series, in which the rise-fall created on the basis of an interrogative utterance was shifted in time.

We then went on to test the hypothesis that the shape of the peak can affect the perception of target location, again by assuming that this perceptual difference would translate into a greater or smaller percentage of question responses on the part of the listeners. Here, we found a difference between the results of primary continuum stimuli timed at plateau onset and plateau stimuli. A much smaller difference was found when a comparison was made between the results of primary continuum stimuli timed at plateau offset and plateau stimuli. We interpret this result to show that shape of the accent peak does indeed affect the perception of target location, and that the perceived target for plateau stimuli must be displaced somewhat towards plateau offset, though not exactly timed with this location. Above all, we take this result to mean that the perceived target of plateau stimuli cannot be identified with the end of the LH rise, as implicitly assumed in D'Imperio and House (1997).

One might object that such a difference between plateau and peak stimuli scores is due to an effect of f_0 level height within the LHL configuration. However, f_0 had relatively little effect when stimuli with peaks (and lows) characterized by a 20 Hz difference were compared. Especially, what we take to be an important result is that primary continuum stimuli, which had an average f_0 value, did not differ significantly either from the hhh stimuli (whose contour was globally 20 Hz higher) or from the lll stimuli (whose contour was globally 20 Hz lower). A difference was found, however, between lll and hhh stimuli, suggesting that there must be an f_0 effect, but it is a small one that is apparent only for extreme contrasts. Note also that the f_0 difference between the lll and the hhh stimuli is equal to 40 Hz, which represents 40.4% of the speaker's range within the accent (D'Imperio, 2000b) Hence, such a sensible difference does not produce a score modification that is as great as the one produced by the timing manipulation, especially at T1 and T7. This is taken as supporting evidence for a greater role of timing in defining tonal target perception.

REFERENCES

AVESANI, C. (1990). A contribution to the synthesis of Italian intonation. Proceedings of the International Conference on Spoken Language Processing. Kobe, Japan.

D'IMPERIO, M. (2000a). Focus and tonal structure in Neapolitan Italian. Speech Communication .

— (2000b). The role of perception in defining tonal targets and their alignment. Ph.D. thesis, The Ohio State University.

D'IMPERIO, M. and HOUSE, D. (1997). Perception of questions and statements in Neapolitan Italian. Proceedings of Eurospeech'97, edited by G. Kokkinakis, N. Fakotakis, and E. Dermatas, vol. 1. Rhodes, Greece.

GÓSY, M. and TERKEN, J. (1994). Question marking in Hungarian: timing and height of pitch peaks. Journal of Phonetics 22:269–281.

MOULINES, E. and CHARPENTIER, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communication 9:453–467.

NÁBELĚK, I. V., NÁBELĚK, A. K, and HIRSH, I. J. (1970). Pitch of tone bursts with changing frequency. The Journal of the Acoustical Society of America 48:536–553.