THE ROLE OF PERCEPTION IN DEFINING TONAL TARGETS AND THEIR ALIGNMENT

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

the Degree Doctor of Philosophy in the Graduate

School of the Ohio State University

By

Mariapaola D'Imperio, M.A.

* * * * *

The Ohio State University
2000

Dissertation Committee:

Mary E. Beckman, co-Adviser

Keith Johnson, co-Adviser

Elizabeth Hume

Approved by

co-Adviser
Department of Linguistics

co-Adviser
Department of Linguistics

ABSTRACT

Tonal targets can be defined in terms of two-dimensions, i.e., “alignment” and “scaling”, where alignment specifies the exact temporal implementation of tonal highs (H) and lows (L) relative to structural elements (such as syllables and morae) and their segments. Alignment patterns might be constrained by various linguistic factors, such as phonological as well as phonetic factors. Among the phonological factors, the grammar of stress-accent languages specifies that the tones of a pitch accent must be aligned with those syllables that are marked as stressed in the lexicon. Among the phonetic constraints, one finds facts about the perception of pitch and time, both for speech and for non-speech stimuli.

This thesis investigates the role of alignment in determining tonal target perception for yes/no question and (narrow focus) statement contours in Neapolitan Italian. These contours are characterized by a melodic rise-fall, analyzed here as a sequence of a LH pitch accent plus a HL phrase tone. The separation of the rise and the fall is clear in the case of long focus constituents containing at least two words with independently stressed syllables. In more typical cases, however, this configuration is acoustically realized as a sequence of three tonal targets, LHL, due to “merging” of the H tone sequence in nuclear position. This thesis shows that the precise alignment of each of those tonal events influences the perception of the question/statement contrast.

A read speech corpora, produced by two speakers of Neapolitan Italian, was first analyzed to acoustically characterize tonal targets in both yes/no questions and narrow focus statements, with target words differing in syllable structure and segmental environment. Later, a set of resynthesized stimuli was created, which constituted the basis for the perception experiments. Results show that, when tonal targets for the entire rise-fall are displaced later in time, more questions are identified. Also, the sole alignment modification of the first or the second L determines different question responses. The results also suggest that f_0 height has a minor role in signaling pitch accent differences, while rise and fall slope have no impact.

Additionally, when the shape of the peak in the rise-fall is modified, so that a high plateau is created, more questions are perceived. This phenomenon cannot be accounted for in terms of a parsing difference between the question and the statement phonological tone structures, since those structures are the same. Moreover, the effect was also found for non-native listeners. Namely, American English listeners showed an effect of peak shape, as well as a similar use of the alignment contrast as a consequence of alignment modifications, when identifying questions vs. statements of Neapolitan. This result suggests a universal use of alignment and a psychoacoustic effect of perceived target displacement due to peak shape. Hence, despite acoustic and pragmatic differences between their rise-fall contrasts, American and Neapolitan listeners appear to employ similar perceptual strategies.

The Neapolitan results also show that syllable structure manipulations are not able to shift the categorical boundary location between a perceived question and a statement. This result suggests that no look-ahead mechanism is employed when computing perceived target location. On the other hand, a category boundary shift was found when stimuli were

resynthesized from either a question base or a declarative base utterance. This suggests that cues other than target alignment are employed when computing perceived pitch accent contrast. In sum, this thesis proposes that temporal alignment, both as a production and a perception mechanism, must shape phonological systems of intonational contrast, both within and across languages.

In memoria di mia nonna, Rosaria,

e alla mia famiglia

ACKNOWLEDGMENTS

This thesis would not exist, were it not for the inspiration and support of so many people. I feel primarily indebted for my intellectual growth to my adviser, Mary Beckman, who put so much into making it all possible. Thanks for welcoming me to the OSU family and for initiating me to the perils and the fascination of scientific research. Thanks also for inspirational guidance and for teaching me how to look at data with the fresh curiosity of a child.

I am also very grateful to Keith Johnson, who has followed my research since my first year at OSU. Thanks for help in interpreting and analyzing data, setting up perception experiments and for supporting me all along. In addition, I would like to thank Beth Hume for her insightful comments and especially for playing a crucial role in stirring my interest in phonology.

Outside the boundaries of OSU, thanks to Jacques Terken for crucial help in designing the main perception experiment of this dissertation and for his warm hospitality during my visiting month at IPO, in May/June of 1999. At IPO, thanks also to Esther Klabbers for friendship and support, to Leo Vogten for allowing me to use the IPO resynthesis software, and to Raymond Veldhuis. A warm thanks goes also to David House, who got me started on the intonation perception path. I have also greatly benefitted from stimulating discussions on alignment and intonation with Jan van Santen during my various stays at Bell Labs.

Special thanks also to Kevin Munhall, for his generous hospitality at the Speech Production Laboratory at Queen's University during this last year, and to the warm friendship of my colleagues and friends there, Jeff Jones and Claire Mc Duffy.

For getting me started on the linguistics path, thanks to Federico Albano Leoni, my mentor in my years as an undergraduate in Naples. I would like to especially thank him for giving such fascinating lectures that made me run to school even on Saturday mornings! Thanks also for introducing me to experimental phonetics, for teaching me to let the data talk for themselves and for continued hospitality at the CIRASS labs (University of Naples). A very special thanks goes to Loredana Cerrato for always listening and instilling confidence in me since my undergraduate years. Thanks also to all the other people that have been so influential in my first years as a linguist: Pietro Maturi, Franco Cutugno, Elda Morlicchio, Rosanna Sornicola, Renata Savy, Mary Caputo and Francesca Dovetto. Special thanks to all of my Neapolitan experiment participants throughout the past years.

I would also like to thank Alan Cruttenden, for first introducing me to intonation (thanks also for pointing me OSU as the place to do my graduate studies!) and Martin Barry for individual discussions and help at the University of Manchester, England, during my ERASMUS year.

A big thanks goes also to all the people that have discussed various aspect of my work, either in person or through email, and/or for generously providing me with papers and manuscripts. In particular, I would like to thank Bob Ladd for fruitful discussions on alignment over the last years. My greatest thanks go also to: Amalia Arvaniti, Cinzia Avesani, Albert Di Cristo, Sonia Frota, Martine Grice, Carlos Gussenhoven, Dick Hermes, Daniel Hirst, Sun-Ah Jun and Elina Savino. Additionally, for the warm hospitality during my stays at Bell Labs, I wish to thank Joe Olive, Bernd Möbius, Chilin Shih and Richard Sproat.

I am also grateful, for discussion and huge support, to a number of people at OSU. I owe a special debt to Rebecca Herman and Jennifer Venditti, for their friendship and for reading and commenting on parts of my earlier work. I would also like to thank Gayle Ayers Elam, Stefanie Jannedy, Barbara Gili Fivela and Sam Rosenthall for being such generous colleagues and friends. And for support, friendship, and/or mentoring, and much more, I need to thank various past and present members of the OSU Linguistics Department and Phonetics Lab. In particular, thanks to Pauline Welby, Kiyoko Yoneyama, Amanda Miller-Ockhuizen, Grazia Russo, Shu-hui Peng, Hélène Loevenbruck, Svetlana Godjevac, Laurie Maynell, Julie McGory, Liz Strand, Matt Makashay, Allison Blodgett, Steve Hartman Keiser, Ben Munson and Craige Hilts. I also owe thanks to Jose Benki, Julie Boland, Mike Calcagno, Peter Culicover, Kim Ainsworth-Darnell, Mary Bradshaw, Nick Cipollone, Paul Davis, Jamie Green, Martin Jansche, Mari Jones, Andreas Kathol, Ilse Lehiste, Joyce McDonough, Bettina Migge, Jen Muller, Dave Odden, Jignesh Patel, Frederick Parkinson, Robert Poletto, Craige Roberts, Jenny Vannest and Steve Winters. Thanks for making my years at OSU so enjoyable and special.

Financial support for this dissertation came from the OSU Department of Linguistics, a Graduate Student Alumni Research Award and an Ameritech/Presidential Fellowship from the Ohio State University Graduate School. Part of the financial support for designing and preparing the perception experiments came also from the University of Eindhoven, The Netherlands (UniversiteitsFonds Eindhoven).

For always being close to me, even from the distance, I wish to thank my parents, Rosa and Vittorio D'Imperio, my brother Luca, my Aunt Maria, and my cousins Raffaele and Michele. Warm thanks also to Paul, Jeanine and Marion Pitermann, for continuous support and encouragement. Finally, the most special thanks goes to Michel Pitermann, my husband. His unending patience, his methodological support for the statistical analysis of the data and his precious suggestions regarding the conceptual structure of Chapter 3 of this thesis have been crucial. But, above all, thanks for always standing by me, giving me strength and motivation, and much more.

VITA

1969	Born in Naples, Italy
1993	B.A. in Modern Foreign Languages and Literatures, University ”Federico II” of Naples, Italy
1994 – 1999	Fulbright Scholar
1995	M.A. in Linguistics, The Ohio State University
1995 – 1996, 1997-1999	Teaching Associate, The Ohio State University
1996 – 1997, spring/summer 1999	Research Associate, The Ohio State University
summer 1997, winter 1999	Language Consultant, Lucent Technologies Bell Laboratories
2000 – present	Presidential Fellow, The Ohio State University (Ameritech Fellowship)
2000	Ph.D. in Linguistics, The Ohio State University

PUBLICATIONS

Grice, M., D'Imperio, M., Savino, M., and Avesani, C. (in press) Towards a strategy for ToBI labelling varieties of Italian. In S.-A. Jun (ed.) *Prosodic Typology and Transcription: A Unified Approach* (Collection of papers from the ICPhS 1999 satellite workshop on "Intonation: Models and ToBI Labeling". San Francisco, California).

D'Imperio, M. and Gili Fivela, B. (in press) How many levels of phrasing? Evidence from two varieties of Italian. In J. Local, R. Ogden, R. Temple (eds.) *Papers in Laboratory Phonology VI*, Cambridge, Cambridge University Press.

D'Imperio, M. (in press). Focus and tonal structure in Neapolitan Italian. In *Speech Communication.*

D'Imperio, M., Terken, J. and Pitermann, M. 2000. Perceived tone "targets" and pitch accent identification in Italian. In *Proceedings of the 8th Australian Speech Science and Technology Conference (SST2000)*. Canberra, Australia.

D'Imperio, M. 2000. Acoustic-perceptual correlates of sentence prominence in Italian. In *Papers from the Linguistics Laboratory. OSU Working Papers in Linguistics*, vol. 52, pp. 59–77.

D'Imperio, M. and Rosenthall, S. 1999. Phonetics and Phonology of Main Stress in Italian. In *Phonology*, vol. 16, n. 1, pp. 1–27.

D'Imperio, M. 1999. Tonal structure and pitch targets in Italian focus constituents. In J. Ohala (ed.) *Proceedings of the 14th International Congress of Phonetic Sciences*, vol. 3. San Francisco, USA, pp. 1757–1760.

Johnson, K. and Strand, E. and D'Imperio, M. 1999. Auditory-visual integration of talker gender in vowel perception. In *Journal of Phonetics*, vol. 27, n. 4, pp. 359–384.

van Santen, J. and D'Imperio, M. 1999. Positional effects on stressed vowel duration in Italian. In J. Ohala (ed.) *Proceedings of the 14th International Congress of Phonetic Sciences*, vol. 1. San Francisco, USA, pp. 241–244.

D'Imperio, M. 1998. Sul ruolo dei correlati acustici della prominenza nella percezione della varietà napoletana di italiano. In *Proceedings of the "VIIIe Giornate di Fonetica Sperimentale"*. Pisa, Italy: G.F.S. (A.I.A.), pp. 43–54.

Gili Fivela, B. and D'Imperio, M. 1998. Focalizzazione e "phrasing" nel parlato letto. Analisi di due varietà di italiano. In *Proceedings of the "VIIIe Giornate di Fonetica Sperimentale"*. Pisa, Italy: G.F.S. (A.I.A.), pp. 55–66.

D'Imperio, M. 1997. Breadth of focus, modality and prominence perception in Neapolitan Italian. In K. Ainsworth-Darnell and M. D'Imperio (eds.) *Papers from the Linguistics Laboratory. Ohio State University Working Papers in Linguistics*, vol. 50, pp. 19–39.

D'Imperio, M. 1997. Narrow focus and focal accent in the Neapolitan variety of Italian. In A. Botinis, G. Koroupetroglou, G. Carayiannis (eds.) *Intonation: Theory, Models and Applications. Proceedings of an ESCA Workshop*. Athens, Greece, pp. 87–90.

D'Imperio, M. and Gili Fivela, B. 1997. Prosodic boundary phenomena in Italian read speech. In A. Botinis, G. Koroupetroglou, G. Carayiannis (eds.) *Intonation: Theory, Models and Applications. Proceedings of an ESCA Workshop*. Athens, Greece, pp. 91–94.

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

Ainsworth-Darnell, K. and D'Imperio, M. 1997. *Papers from the Linguistics Laboratory*. Ohio State University Working Papers in Linguistics, vol. 50.

D'Imperio, M. 1996. Caratteristiche di timing degli accenti nucleari in italiano parlato letto. In *Atti del XXIV Convegno Nazionale dell'Associazione Italiana di Acustica*. Trento, Italy, pp. 55–60.

D'Imperio, M. 1995. Timing differences between prenuclear and nuclear pitch accents in Neapolitan Italian. In *J*ASA, 98(5), p. 2894.

D'Imperio, M. 1995. Prosodie dell'inglese: analisi uditiva e analisi strumentale. In *Rivista Italiana di Acustica*, vol. 19, n. 3, pp. 41–53.

Caputo, M. R. and D'Imperio, M. 1995. Verso un possibile sistema di trascrizione prosodica dell'italiano: cenni preliminari. In *Atti delle Ve Giornate del Gruppo di Fonetica Sperimentale*. Trento, Italy, pp. 71–83.

D'Imperio, M. 1994. Analisi soprasegmentale di parlato inglese letto e spontaneo: correlati acustici della prominenza. In *Atti delle IVe Giornate di Studio del Gruppo di Fonetica Sperimentale*, Torino, Italy, pp. 39–48.

Cutugno, F., D'Imperio, M. and Savy, R. 1994. Un procedimento di analisi e sintesi parametrica di segnali vocali. In *Atti del XXII Convegno Nazionale della Associazione Italiana di Acustica (AIA)*, Lecce, Italy, pp. 437–442.

FIELDS OF STUDY

Major field: Linguistics

Specialization: Phonetics

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Tune and target

In spoken utterances, the choice of words and word order are not the only things that determine the meaning. Rather, the accompanying melody varies too, as illustrated in Figure 1.1. This is a graph of the fundamental frequency (f_0 henceforth) over time during the production of two closely related utterance types. Here the two different word orders, *Roy nominated 'em* versus *Did Roy nominate 'em?*, signal the contrast between a statement and a question, but the different texts are accompanied also by different melodies. Different f_0 values typically cue different pitch sensations. Note, for instance, that there is a sharp fall aligned to the word *nominated* in the statement as opposed to a sharp rise aligned to the same location in the question. We can describe this difference as a glissando from a *High tone target* (a high-pitched "note") to a *Low tone target* (a low-pitched "note") in the statement versus a Low tone target to a High tone target in the question.

It is essential, though, that the tone glissando be aligned in a well-defined way with the text. That is, the phonology of languages like English places constraints on what can be a well-formed association between tune and text. The main constraint (Liberman 1975;

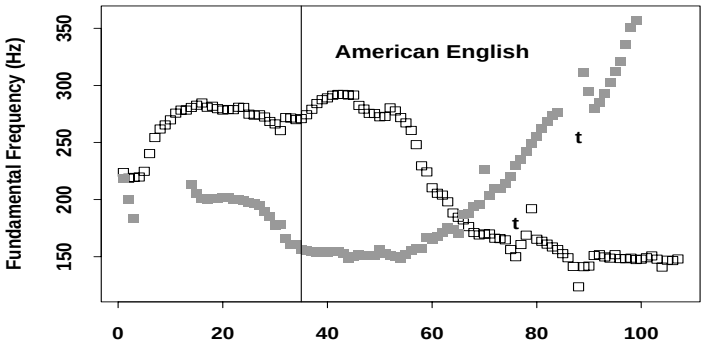

Figure 1.1: f_0 traces for a statement utterance (open squares) *Roy nominated 'em* and a question utterance (filled squares) *Did Roy nominate 'em?* in American English. The vertical bar marks the onset of the stressed syllable *no-*. The "t" marks a perturbation in the contour.

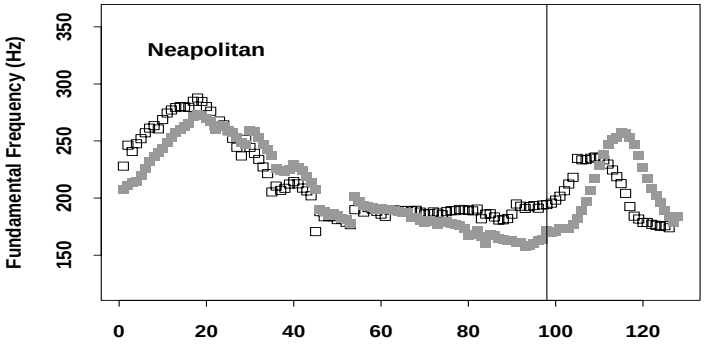

Figure 1.2: f_0 traces for a statement (open squares) and a question (filled squares) *Mamma andava a ballare da Lalla./?* "Mom used to go dancing at Lalla's" in Neapolitan Italian. The vertical bar marks the onset of the stressed syllable *Lal-*.

Beckman and Edwards 1994; Beckman 1996) is that the glissando should occur on a syllable that is metrically strong or "heavy" (that is, the syllable should contain a tense vowel or diphthong or an "unreduced" lax vowel. [1] That is why the HL glissando of the statement in Figure 1.1 cooccurs temporally with the syllable *n*o- of *n*ominated, and not, say, with the last syllable of the same word. Hence, any heavy syllable is a potential site for a tone or a glissando to be produced.

In Italian, the analogous contrast between statement and question is typically signalled by intonation alone. Moreover, in Neapolitan Italian, the aspect of the intonation pattern that is different between the two sentence types is not the choice of tone sequence, but rather a more subtle difference in the *alignment* between the tune and the text. By alignment I mean the temporal cooccurrence of tune and some specific segments within the text. This is illustrated in Figure 1.2, where the utterance is *Mamma andava a ballare da Lalla* "Mom used to go dancing at Lalla's". Note how the peak corresponding to the last High tone (the one on *Lalla*) is later in the question (circles). In fact, despite all the other differences that one might notice in accent shape, it has been suggested that listeners rely upon the H target (peak) alignment as one of the strongest cues when deciding whether the interlocutor has asked something or simply stated something (D'Imperio and House (1997)).

Phonetic alignment is therefore an important vehicle of linguistic information, and it will be the central topic of this dissertation. Since alignment has usually been considered a quantitative phenomenon, and confined to phonetics, it has been neglected by phonological theorists (while other dimensions of tone targets, such as their f_0 height, has been

[1] An example of reduced vowel is the first vowel in the words *p*otato or *p*olice.

extensively studied) until very recently. Hence, I will also be concerned with the relationship between alignment and the phonological concept of (tune-text) association, by which I mean a set of constraints specifying to the speaker that a tone (or glissando) should be produced (and perceived) as temporally cooccurring with specific consonants or vowels in the text. By saying that the tune is associated to the text, I do not imply a serial architecture in which the segmental structure is computed first and then the tonal structure is "attached" to it in a derivational fashion. Rather, we could as well speak about text-tune association. In other words, I will embrace the current view that speech production and perception is the result of constraints of different nature, which act in synergy. A brief discussion of alternative views is offered in §1.3. Finally, the relationship between the notion of tonal alignment and the notion of phonological association are explored, with the aim to describe how the contrast between the question pitch accent and the statement pitch accent might be linguistically represented.

Hence, the main goal of this dissertation was to investigate the notion of temporal "alignment" of tonal targets. But in order to do this, one needs to explicitly define such targets. Intuitively, one can think of a target in the f_0 curve as a maximum or minimum value within the same, as in the obvious contrast in peak location shown in Figure 1.2. Specifically, tonal targets can be defined along two axes, which are the temporal location (i.e., *alignment*) and the melodic value within the f_0 range of the speaker (i.e., *scaling*).

However, it is quite hard to discern a tone target in many occasions. It is possible, in fact, to hear a tune contrast without clearly discerning the peak or valley in the f_0 contour. This is the case shown in Figure 1.3, which shows the f_0 contour for a broad focus declarative in Neapolitan *Mamma andava a ballare da Lalla* "Mom used to go dancing at Lalla's".

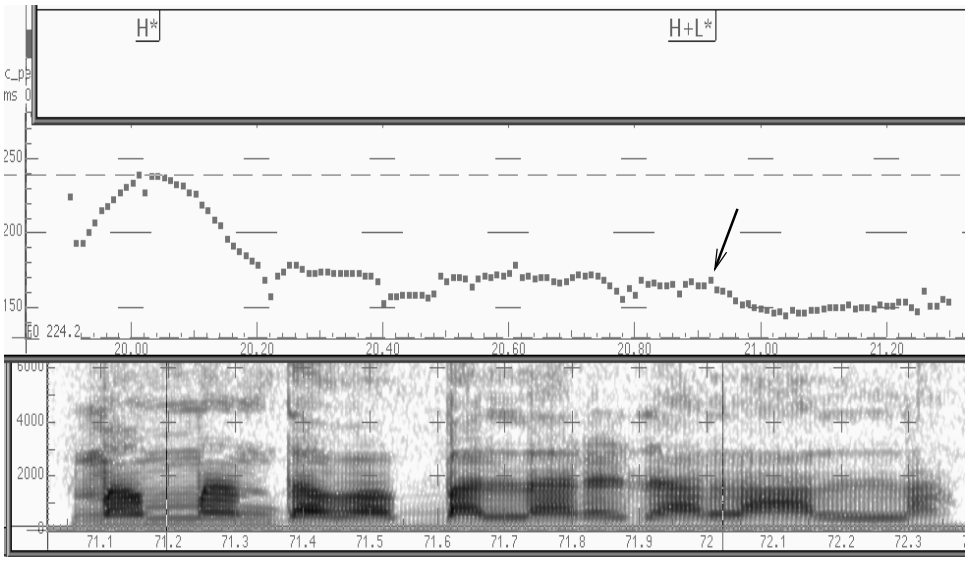

Figure 1.3: f_0 curve and spectrogram for the broad focus declarative *Mamma andava a ballare da Lalla*.

Speakers of Italian hear a clear falling accent on the stressed syllable *Lal-* of *Lalla* at the location marked by H+L* and the arrow. But how can one exactly locate the accent peak within an f_0 plateau?

Alternatively, one could discern an event in the f_0 trace which is not clearly associated with an audible contrast. This case is shown in Figure 1.1, where we can notice a perturbation of the f_0 contour in correspondence of the segment [t] in the word “nominated” in the question rendition and in “nominate” of the statement rendition. Note a sudden rise around the [t] in the question: is such a rise to be identified with a target peak or is it just “noise”? These kinds of perturbations are indeed very frequent and cause tonal target measurement to be quite difficult.

In my view, some of the problems with the notion of tonal target come from defining it as the actual maximum or minimum value that is extracted by the computer and that is displayed in the form of a peak or a valley in the f_0 contour. In this thesis I propose instead to define tonal targets also in perceptual terms, by investigating auditory properties of certain pitch accent configurations and timing. The goal is to arrive at some principled definition of the notion of tonal target by investigating tonal alignment. Specifically, I will propose that the notion of phonological association should also relate to the notion of "perceptual target", as well as to other physically measurable properties. This task was accomplished in a twofold way. On one side, alignment was investigated in production, in order to give a precise characterization of some targets in the acoustic domain. On the other, the mapping between the acoustically defined targets and perceptually determined targets was studied.

The investigation of tonal alignment in perception can also offer a window on some peculiarities of tonal alignment in production. For instance, many languages of the world present a High accent whose alignment is late relative to the syllable with which it is associated (e.g., the prenuclear H* accent of American English, Italian, Spanish, Greek)[2]. It is plausible to hypothesize that the psychophysics of tonal perception might be guiding tonal alignment in production. In other words, from the point of view of psychophysics, such apparent misalignments between accent peak and associated tone bearing unit might be accounted for in terms of perceived target. By performing some perception studies on

[2]The nuclear accent is generally defined as the last accent of an intermediate phrase, hence prenuclear accents are all possible preceding accents

the role of tonal alignment, one can discover whether such peculiarities of alignment patterns in production are due to specific constraints on our perceptual system. Some of these constraints might be psychoacoustic in nature. In fact, work conducted in the seventies suggests that complex pitch configurations (such as a rising or a falling pitch movement) are not perceived in their entirety, and that the perceived pitch target corresponds to the f_0 value at roughly two thirds of the length of the rise or fall (Nábeläk and Hirsh 1969; Nábeläk, Nábeläk, and Hirsh 1970; Rossi 1971; Rossi 1978). This issue is discussed in §1.6 and is experimentally investigated in Chapter 3.

Despite a long tradition of psychophysical studies making use of non-speech stimuli (such as pure tones), only a handful of perception studies on intonation have been carried out with speech stimuli. Furthermore, those studies almost never employed entire words or utterances. That is why we still do not know which part of tonal perception is language-specific and which can be safely considered "universal". Recent findings seem to reflect the contribution of language-specific intonation patterns in our ability to perceive specific tonal configurations. Hence, the effect of native language on the perception of alignment was also explored here. The assumption made is that, though there must be some universal psychophysical basis that allows people to learn a specific linguistic contrast, the symbolic structure reshapes attentional structure in important ways. In other words, universal psychophysical principles cannot be fully explanatory when it comes to the perception of intonation contours. The second goal of this study was therefore to investigate how the psychophysics of alignment and the phonology of alignment interplay in perception. Perceptual experiments were in fact performed with more than one language, namely Neapolitan Italian and American English. Also, American listeners were asked to perform both a

linguistic and a pseudo-psychoacoustic perceptual test. Thus, through the use of the same set of speech stimuli, the contribution of language-specific information to tonal perception could be accounted for.

Accents do not differ only in terms of alignment of peaks and valleys. A production study on Neapolitan Italian tried to capture some of the additional acoustic cues for the question/statement contrast. Such acoustic cues were then employed in order to perceptually test a preliminary theory of tonal target localization. The stimuli employed were resynthesized from natural speech. The use of acoustic resynthesis allows, in fact, for independent manipulation of acoustic parameters, while keeping the speech signal perceptually natural.

Also, tonal alignment appears to be quite sensitive to syllable structure and syntactic environment. We saw in Figure 1.2 that the peak of Neapolitan questions is later than that of statements, everything else being equal. Specifically, question peaks tend to occur close to the offset of the stressed vowel (which, in open syllables, corresponds to the offset of the stressed syllable). Syllable structure effects can, however, render the relative alignment of statement peaks dangerously similar to that of question peaks. Compare the peak location in the question utterance *Vedrai il NANO dopo?* "Will you see the DWARF afterwards?" (capital letters indicate accent placement) and the peak location in the declarative *Mangerai il MANGO dopo* in Figure 1.4. Both peaks are aligned very close to the end of the stressed vowel (indicated by the label "v1" and marked by the dashed line). Note, though, that the stressed vowel is shorter in the declarative, as a result of occurring in the closed syllable *man-*.

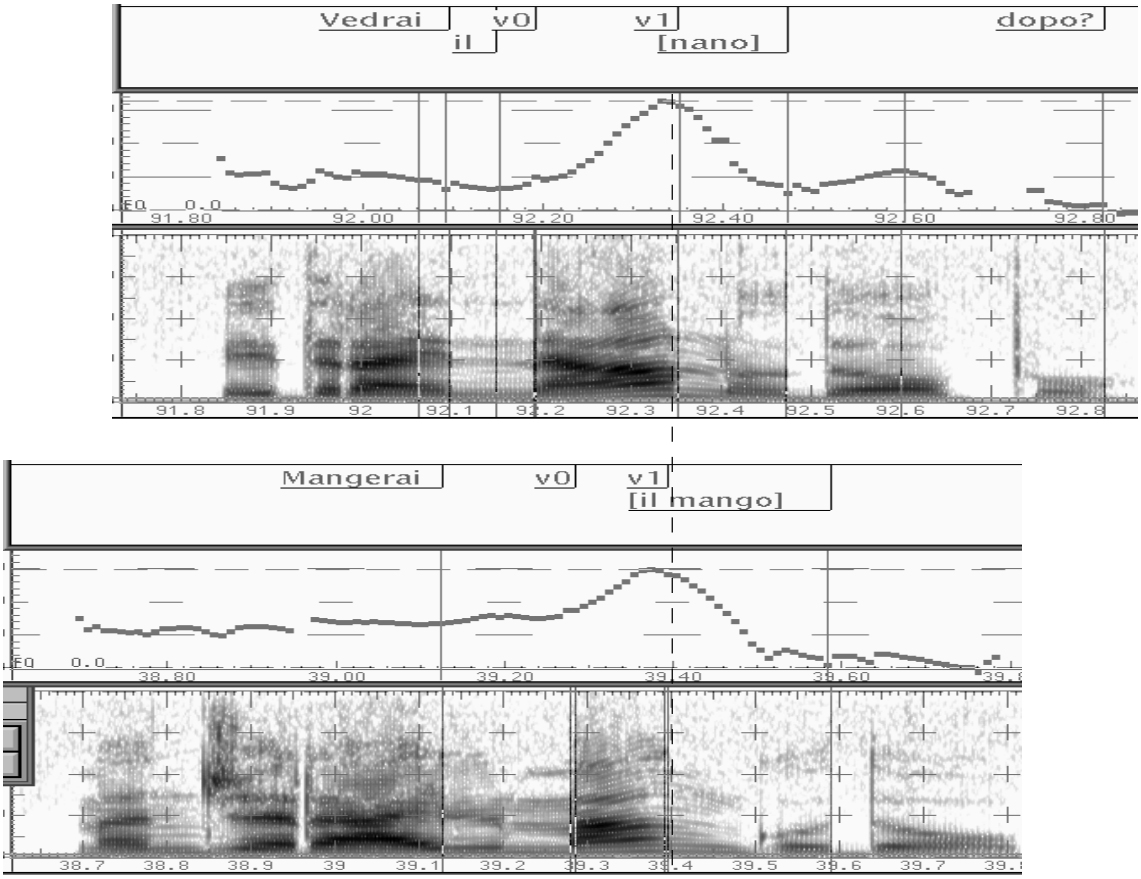

Figure 1.4: Tone labels, f_0 curve and spectrogram for the utterances *Vedrai il NANO dopo?* (upper) and *Vedrai il MANGO dopo* (lower). Stressed vowel offsets are lined up at the dashed line.

How do Neapolitan listeners deal with such a potential parsing ambiguity? The issue here is what counts as phonologically the same intonation pattern. If we take f_0 to be the main phonetic representation of tonal categories, then such pairs are awkwardly similar. The empirical question is whether such ambiguous pairs can be distinguished perceptually. Such a test would support an active role of perceptual cues in shaping the phonological representation of tone. Similar ambiguities seem to arise also as a result of segmental environment, though such cases will be discussed later. I therefore also explored the effects of syllable structure and segmental environment on both production and perception of tonal alignment. The goal was to determine whether relative or absolute target alignment is employed by the listener to differentiate between a question and and a statement pitch accent.

In the next section, the notions of phonological association and phonetic alignment are introduced. Then, in §1.4, I review a series of studies that investigate alignment in production and factors affecting it. §1.5 is devoted to alignment in perception, while in §1.5.1 I report the results of a pilot study which is at the basis of the perception experiments of this dissertation. Finally, in §1.6 I summarize some psychoacoustic notions of pitch perception in speech and non-speech stimuli.

1.2 Alignment in intonational phonology

Tonal alignment can be defined as the temporal synchronization of tones and segments, mediated by the prosodic structure. This notion has a great import in all the ramifications of current intonational phonology, since, in one way or the other, alignment of tones

with structural positions in the string can have dramatic repercussions for the phonological makeup of the melodic contour, as well as for its pragmatic meaning, in a way that has been found to be categorical (Pierrehumbert and Steele 1987; Kohler 1987; Pierrehumbert and Steele 1989; D'Imperio and House 1997).

Tune-text alignment is in fact responsible of different types of contrasts. For instance, in the classic example in (1a) (from Bolinger (1961); capital letters indicate accent placement), a high tonal target is placed on the syllable *d*e- of *d*eported for contrastive reasons, while in the non-contrastive, "default", rendition of the same word, as in (1b) we expect the same tone to be aligned with the stressed syllable *p*or-.

1a. This whiskey wasn't EXported from Ireland; it was DEported.

1b. Those prisoners were dePORted.

Moreover, a tone can either be aligned to cooccur with a stressed (therefore metrically prominent) syllable or it can be aligned with the edge of a phrase. Consider the LH rise of the American question utterance in Figure 1.1. The LH glissando, here, is the result of a L tone associated with the stressed syllable *no-* of *nominate*, plus a following H tone which is a property of the entire phrase (and is aligned with the last segments of the same). A similar LH rise can instead be the expression of a glissando in which both tones are associated to a stressed syllable. This is the case of the LH rise of questions and statements in Neapolitan Italian shown in Figure 1.2. Here, the contrast between the statement and the question rise is encoded through an alignment difference between tune and text (this issue will be further discussed in Chapter 6). Relative alignment can then be said to be contrastive in Neapolitan Italian.

All current theories of intonational phonology therefore incorporate some notion of alignment, though they differ in the specific mechanism for its implementation as well as in the status it is granted within the specific theory. Alignment is either a "consequence" of the notion of starredness and association (Pierrehumbert 1980; Pierrehumbert and Beckman 1988; Hirst 1988; Grice 1995a; Arvaniti, Ladd, and Mennen 2000), or is taken to be a binary (Bruce 1977; Ladd 1983) feature of level tones and, alternatively, of pitch configurations ('t Hart, Collier, and Cohen 1990; Gussenhoven 1984). Within the last few years considerable attention has focused on the temporal alignment of High and Low tones with the consonants and vowels in an utterance. But before discussing what tonal alignment is and the factors that affect it in production and perception, a review of the notion of association and how it relates to tonal alignment across various works will be presented.

1.3 Autosegmental-metrical theory of intonation and the notions of "association" and "starredness"

The present dissertation is couched within the autosegmental-metrical (AM) theory of intonational phonology, especially associated with the work of Pierrehumbert and colleagues (Pierrehumbert 1980; Beckman and Pierrehumbert 1986; Pierrehumbert and Beckman 1988), and, more recently, with the work of Ladd, Grice and colleagues (see Ladd (1996) for a review). Such work is also assumed by intonational transcription systems inspired by ToBI (Tone and Break Indices, (Beckman and Ayers Elam 1994)), for a variety of languages. One of the central tenets of such a theory is that tones are systematically associated with the text.

The idea of a "phonological association" between tune and text was first proposed within the framework of autosegmental phonology (Leben 1973; Goldsmith 1976) (whose essential contribution is recognized in the AM theory). Within classic autosegmental treatments, representations are regarded as arrays of elements arranged on different levels, or tiers. In principle, these tiers are autonomous, and group together linguistic objects, such as segments, tones, and so forth. Association is then the mechanism that autosegmental phonology employs in order to link elements on different tiers. In the first autosegmental works, association links tones to vowels, and reveals itself to be a useful notion to describe the phonology of tone in African languages. No real attempt at formalizing the notion of association relation can be found in Goldmith's work nor in other early autosegmental work. Goldsmith simply establishes a set of "well-formedness conditions" which, in their classical formulation (Goldsmith 1976), state that 1) each vowel must be associated with at least one tone; 2) each tone must be associated with at least one vowel; 3) no association lines may cross (more recent formulations of the association relation and of conditions constraining it are discussed below).

Within the AM theory of intonation, the notion of association usually refers to the "special link" between an autosegment on a tonal tier and a prosodic domain (such as the syllable). Nevertheless, while Goldmith's association rules were still couched in a strictly derivational fashion, in this dissertation a representation such as the one shown in Figure 1.5 (in which the vertical line stands for the association relation between a tone T and a syllable σ) will be simply taken to specify that the speaker will produce (and perceive) the physical realization of T (i.e., the specific pitch level manifesting T) "together with" the segments that compose σ (i.e., within the same temporal span).

Figure 1.5: Schematic representation of association between tone and syllable.

It is important to note that in autosegmental theory a tone is regarded as a “pitch level”, not a contour. In later AM intonation work, tone will often stand to mean “pitch target”, or, alternatively, “turning point” in the pitch curve (Bruce 1977; Bruce and Gårding 1978). This point will be taken up below.

The concept of “starredness” of a tone also goes back to classic autosegmental theory and is closely linked to the concept of association. Unlike tone languages, accentual languages are characterized by a mechanism of association, by which a specific tone and a specific vowel, each marked with a star (an asterisk), are selected first in the association procedure. A pitch accent is a pitch configuration that has the function of rendering one word more prominent than the others in an utterance, and in languages such as English and Italian it is a tone that is associated (that is, localized in a contrastive way) to a stressed syllable. This means that a pitch accent can only occur on syllables that are marked as stressed in our mental dictionary, as for instance the syllable “bil” in the English noun “billow”. We know in fact that the same syllable could not be accented if it occurred in the word “below”, where it is the second syllable that carries the stress. Therefore, whenever

we say that a syllable is accented, we mean that a melodic configuration, namely a pitch accent, is placed on it. Such a pitch configuration is analyzed as being made of target pitch levels, specifically a High or a Low level, or a combination of the two. We can roughly think of these levels as the notes on a musical score, the difference being that for intonation we deal with both "simple notes", which are the accents consisting of a single target level, as well as sequence of notes or "glissandi", which are accents composed of more than a target level (such as a sequence of a High and a Low, i.e., a rise). Also, differently from music, in speech one or more of such "notes" will bear a special relationship with its associated structural unit (vowel, syllable, mora, etc.), and will be marked by a star. In music, notes can only be associated to other notes.

In a strict derivational approach, starred elements, which are more prominent than unstarred ones, are assumed to have priority in the association process; only in a second stage will other vowels and tones be linked to each other according to "readjustment rules". A point to notice is that AM work on intonation allows for sparse tonal specification. There are in fact languages in which not every *tone bearing unit* (henceforth, TBU) needs to be linked to a tone and vice versa, an example of this type being English.

At roughly the same time as Goldsmith's dissertation appeared, another seminal work for current AM theory was being written, Bruce (1977). As a general remark, one must notice that the Swedish school of intonation has produced a great deal of work, since the seventies, on the notion of alignment and its impact on meaning (Bruce 1977; Bruce and Gårding 1978; Bruce 1983; Bruce 1987). Bruce's dissertation (Bruce 1977), specifically, lays the foundations of autosegmental work on intonation. Most notably, some of its central notions were adopted by a later seminal work, Pierrehumbert (1980). Though tonal

association is still loosely defined, in Bruce's work the syllable is used as the docking site for tone, which he labels "word accent". The phonetic correlate of a word accent is "a pitch obtrusion in connection with the stressed syllable" (Bruce, 1977:11). This concept is akin to the concept of "pitch accent" in English (Bolinger 1958), though a word accent is a lexical property of words and, unlike the English pitch accent, does not merely convey pragmatic meaning. Within the pitch obtrusion, it is not the configuration of the tone that interests Bruce, but the position of the "turning points", i.e., High and a Low targets within the contour itself. As he puts it : "This reflects the view that reaching a certain pitch level at a particular point in time is the important thing, not the movement (rise or fall) itself" (Bruce 1977, p. 132).

The relevance of this study, especially for later work on alignment, lies in capturing a very simple, yet still unnoticed, generalization about a property of Swedish word accents, generally known as "Accent I" and "Accent II". In Standard Stockholm Swedish, both accents have a falling configuration over the stressed syllable; different dialects of Swedish show different locations for the pitch peak in each of these accent types, though they keep one crucial distinction between them: the peak of Accent II is always realized later within the stressed syllable than Accent I. Bruce formalized this opposition by making use of a feature [+/-Accent II], whose correlate is "late vs. early timing of the pitch obtrusion" (Bruce 1977, p. 11). He will also employ relative peak timing in order to formally distinguish between accent types in Stockholm Swedish. In other words, in this account, association does not determine a specific pattern of alignment *per se*, since the whole accent shape is associated with a stressed syllable. Bruce in fact claims that one stressed syllable "in every lexical unit" is associated to a feature [+Word Accent]. The phonetic details of

peak alignment are instead determined by the accent feature specification. Tonal alignment is therefore taken to be contrastive in this theory. This idea shares some similarities with the work of Ladd (1983), which I will discuss below. The relevance of Bruce's contribution for alignment in production and perception will be discussed in sections §1.4 and §1.5.

Tune-text alignment is one of the three main components of a grammar of intonation listed by Pierrehumbert (1980), where the text is metrically organized into a grid (Liberman 1975; Liberman and Prince 1977). In Pierrehumbert's system, as in Bruce's, only two tones are employed, a high tone (H) and a low tone (L). These tones are the basic elements of a pitch accent. Specifically, pitch accents can either be monotonal (made of one component tone, e.g. H) or bitonal (made of two tones, connected together by a "+" sign, as in L+H). Pitch accents are characterized not only in terms of their component tones but also by "a feature controlling alignment with the text" (Pierrehumbert, 1980:9). Therefore, a LH sequence of tones can potentially be employed to describe two different pitch accents, depending on which one will be associated to a strong metrical position (i.e., a stressed syllable). Association is marked with an asterisk (the "star" of Goldsmith's work), which is used to reflect the association relation between the specific tone carrying it (also called "starred" tone) and the stressed syllable. So, in the case of the LH sequence, the grammar of intonation provides both a L*+H and a L+H* types. While, in the first, case L is associated to the stressed syllable, and H is simply realized at a fixed distance from L, in the second case it is the H tone that participates in the association relation.

Here, I would like to concentrate on the meaning of the starredness notion in the original Pierrehumbert's approach as well as in more recent interpretations of it. First, the star is essential to distinguish a pitch accent from edge related tonal events. The grammar of

English intonation includes, in fact, also unstarred tones, such as a phrase accent (e.g. L-), which is the tone that immediately follows the last pitch accent in an intermediate phrase and controls the pitch until the end of the phrase itself, and a boundary tone (e.g. L%), which is aligned at the periphery of the intonational phrase. Second, the star notation allows Pierrehumbert to distinguish between bitonal pitch accents which are characterized by the same glissando (say, a LH rise), but in which only one of the tones participates in the association relation with the stressed syllable, as for the above mentioned L+H*/L*+H pair.

Additionally, the star is a means to indicate "relative alignment". Namely, while L*+H aligns the L target with the stressed syllable and the H target "trails", in L+H* the L target "leads" while the H target is specified to cooccur temporally with the stressed syllable. Moreover, the star notation can also be taken to represent a "strength relationship" between the two tones, in a way that is akin to the stressed/unstressed relationship between syllables in a foot. This notion of metrical strength as a property of starred tones has been recently exploited by Grice (1995b). In other words, the T+T* (where T stands for "tone") notation can be replaced by a representation such as the one depicted in Figure 1.6 for the H+L* pitch accent (PA). Note that a PA is made of at least a tone (T). When it is made of two tones, one of the tones is weak (w) and the other is strong (s). The strong tone is the one that is marked with an asterisk in the classic autosegmental notation and in Pierrehumbert (1980). Finally, starred tones have been taken to be temporally more stable than unstarred tones (Frota 2000b) (but see §1.4 below for account of "tonal repulsion" acting on prenuclear H* in English), or more resistant to modifications in the f_0 domain (Arvaniti, Ladd, and Mennen 2000). That is, in terms of concretely measured targets, starred tones have

been proposed to be more resistant to local variability in alignment and scaling. Hence, in current AM theory starredness is related to a quite diverse array of phonological and phonetic characteristics.

Apart from the notions of starredness and association, Pierrehumbert's work is important also for introducing a view of tone that departs from the very concrete notion of "tone target" embodied by Bruce's work. For Bruce, tones are all phonetically implemented in tangible targets (which he will later refer to as "turning points"), each characterized by a specific f_0 value and temporal alignment. For Pierrehumbert's, instead, not all tones have a concrete phonetic manifestation. For instance, the L tone of H*+L (which has been later excluded from ToBI transcriptions of English) is not implemented phonetically but is interpreted as the trigger of "downstep" (a process by which successive high tones are realized at lower f_0 values than otherwise expected). This idea was common in the study of lexical tone in African languages. Also, for Pierrehumbert, not all visible f_0 valleys and peaks need to be mapped to tones, as in the case of the "sag" in the f_0 contour between two successive H* accents, which is interpreted as a result of non-linear interpolation between them.

To summarize, in Bruce (1977) both Accent I and Accent II are assigned the same phonological representation, that of HL falling accents in which both tones are associated to the stressed syllable. Only later (Bruce 1987) will Bruce employ the notion of starredness as it is represented by most current AM approaches and will differentiate between two structurally different pitch accents. That is, Accent I will be interpreted as having its low tone associated with the stressed syllable (H+L*), while in Accent II it is the H tone

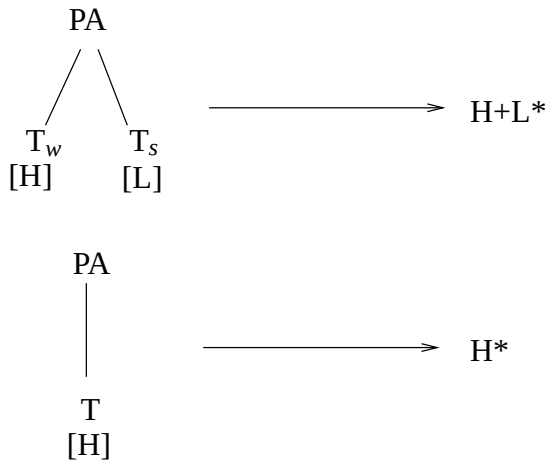

Figure 1.6: Schematic representation of a bitonal pitch accent (PA) and of a monotonal pitch accent.

that is associated with the same (H*+L). Therefore, in the original version of Bruce's theory, alignment does not depend on association, in that there is no explicit prediction about the phonetic alignment of tones and the stressed syllable. Also, tones are quite concrete events, in that they are identified with visible turning points within the f_0 contour. Pierrehumbert's and subsequent AM work will significantly depart from both notions, in that association (and therefore starredness) directly translates into alignment and the tones-to-f_0 target mapping is more complex.

Now, I will only briefly illustrate a theory where, similarly to Bruce's original theory, association and alignment do not imply temporal overlap of the elements that are in the association relation. This theory is exemplified by Ladd (1983), who proposes the use of the feature [+delayed peak] in order to represent the alignment of tones relative to the

syllable to which they are associated. Ladd's central concern was to capture similarity in meaning between pitch accents that appeared to differ only in form. These accents are the "plain" fall (H* L) and its "scooped" variant (L*+H L). In Ladd's theory, the scooped fall is characterized by the feature [+delayed peak], since its target H is reached later in the syllable than the target for H*. In fact, he proposes that both pitch accents can be represented as a HL fall, where only the scooped variant is [+delayed peak]. For Ladd, alignment is contrastive, as his analysis of the H* vs. L*+H opposition reveals, therefore it has a phonological status. But alignment is also phonetic, in that it can be employed to describe details of phonetic realization of the same pitch accent in different varieties of the same language (as in Swedish, for example). Therefore, the main difference from the treatment in Pierrehumbert and colleagues' theory is that alignment does not fall out from the association relation between starred tones and structural elements, but it is itself specified as an independent feature. Recent research has shown that Ladd's line of argument was not entirely justified, since the scooped and plain fall are not variants of the same pitch accent in English. In fact, their meaning can be contrasted in a categorical way, and peak alignment appears to be binarily and not gradually manipulated (Pierrehumbert and Steele 1987; Pierrehumbert and Steele 1989).

The viewpoint expressed in Ladd (1983) is rather similar to that offered by Gussenhoven (1984), though his "delay" feature applies to the contour (such as a "fall" or a "rise") as a whole and not to a specific target within the contour (such as a H or a L). For Hirst (1988), the alignment between tones and the phonemes associated to a specific syllable is mediated by a superordinate prosodic category (either the intonational unit, the tonal unit) to which both the syllabic and the tonal tier (which are on different planes) are linked, and which appears to dominate both of them.

In Pierrehumbert and Beckman (1988), work on Japanese lead to a very explicit formulation of association as a relation between a set of structural elements (such as syllables or moras) and one of the elements of a specific substantive tier (such as a the tone tier). First, the concept of association is defined in terms that are different from the dominance relation, which is simply a relation between two structural elements, such as a foot and a syllable. Formally, the association relation is defined as "an ordered pair $\langle s_i, N_j \rangle$ where s_i is an element of a substantive tier and N_j is a node of a tree T" (Pierrehumbert and Beckman, 1988:153). Moreover, "The association relation A obtaining between the entire substantive tier S and the tree T is then a set of such ordered pairs "sAn". From the point of view of the temporal interpretation of association, it represents (Pierrehumbert and Beckman 1988, p. 153) "temporal overlap between substantive elements and structural positions", or, in other words, if the structural element is a mora and the associated substantive element is a tone, what this means is that the tone will "occur simultaneously with any phoneme segments associated to that mora" (p. 119).

Though phonetic alignment falls out of the notion of association, it superficially appears that the two concepts might not be readily equated. One of the most evident reasons

why association does not directly and unequivocally translate in a certain pattern of phonetic alignment is that different languages tend to align tones towards a certain edge of the structural domain they are associated to. For instance, it appears that targets for starred tones, such as H*, in English tend to occur towards the right-edge (i.e., "right-peripheral" alignment in Pierrehumbert and Beckman (1988)) of the stressed syllable, while the opposite seems to be true for starred tones of Standard Swedish. Actually, in Standard Swedish there seems to be sociolinguistic pressure enforcing this kind of alignment, since non-standard dialects align starred tones later. Moreover, within the same language, a starred tone might be aligned differently depending on the specific pitch accent it belongs to. So, in Neapolitan Italian, the L* of L*+H tends to occur much earlier than the L* in H+L* of broad focus declaratives, as Figure 1.4 and Figure 1.3 show.

In sum, the notion of association in Pierrehumbert and Beckman (1988) entails a special relationship between a tone and a structural unit, which simply translates in "temporal overlap" between tones and segments. However, the specific details of the coordination between tones and the segments that are linked to the structural unit are, in such a model, not part of the phonological representation itself. As Pierrehumbert and Beckman (1988, p. 159) state:

> *The left-peripheral, central, and right-peripheral alignments cannot be distinguished by the association relation alone. If it is desirable to distinguish them–for instance to describe the relative timing of elements on different tiers that are attached to the same node–this must be done by additional stipulations. In cases with which we are familiar such stipulations take the form of language-specific rules of phonetic interpretation.*

The notion of "temporal overlap" embedded in the traditional notion of association between a tone and structural node is therefore not specific enough, since, as we saw above, target tones can privilege alignment with the left or the right edge of the syllable. Such seemingly small details of alignment are of paramount importance in perceiving linguistic contrast, though. Also, the alignment of visible acoustic targets alone can be, however, puzzling if we want to directly employ alignment as a window on the phonological representation of a specific pitch accent. For instance, in some cases, neither of the tones in a pitch accent appears to be aligned with the stressed syllable, as in the case of the Greek prenuclear rise (Arvaniti, Ladd, and Mennen 2000) and Glasgow English nuclear accents (Ladd 1996). Specifically, both cases are characterized by a LH rise in which the L is aligned before the onset of the stressed syllable while the H is aligned with postaccentual segments. Neapolitan Italian offers a mirror-like but similarly difficult situation. Namely, the L and H of both the question and the statement LH rises are aligned with the same stressed syllable (see Figure 1.7). How is then such a contrast represented in the mind of Neapolitan speakers? Which one is the "starred" tone? Can any of the properties that have been attributed to starred tones be found for either the L or the H in any of the two pitch accents?

Also, if acoustic alignment is the taken to be the "only" phonetic exponent of association, it becomes difficult to maintain that prenuclear H accents of English are associated with the stressed syllable. As mentioned above, we know that in English as well as in a number of other languages (such as Italian and Spanish), prenuclear accents often reach their target in the postonic syllable. On the other hand, Pierrehumbert and Beckman (1988) recognize the possibility that a tone can "spill over" onto the syllable following the TBU.

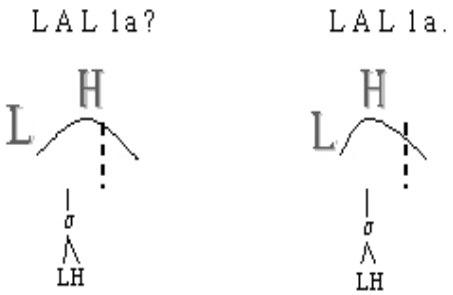

Figure 1.7: Schematic representation of tonal alignment in Neapolitan questions (left) and statements (right).

Also, from a perceptual point of view, it has been noticed that for a pitch movement to be "prominence-lending", it does not have to be necessarily completed within the boundaries of the TBU (Hermes, Beaugendre, and House 1997).

Therefore, we are presented with a puzzling situation. Either we have to say that the details of temporal alignment of starred tones (such as the prenuclear H* of English and other languages, or the low targets of L*+H and H+L* in Neapolitan) are due to (language-specific) phonetic implementation of different phonological entities (such as a "delayed" L* vs. an "early" L*), or that alignment itself is an essential part of the notion of phonological association. For instance, we might postulate the existence of alignment constraints requiring the second tone within a pitch accent to be as close as possible to the syllable

right-edge. The origin of such a constraint, in turn, might be related to, say, the goal of rendering the perceived target of the fall more salient by producing it within a spectrally stable region (for the concept of spectral stability, see §1.6 below), and/or to sociolinguistic factors (as in the case of Swedish starred tones) and/or to some general constraints of physiological nature. The position assumed in this dissertation is that the nature of phonological representations is quite complex, and, above all, it is phonetically grounded, as already assumed in Jakobson et al. (1952) and recently restated by Pierrehumbert et al. (in press) on the basis of a wealth of experimental findings. In other words, alignment is both phonetic and phonological, in that it describes the details of temporal coordination between tones and structural nodes, which, in turn, is part of the phonological representation for a specific tone. Also, I will assume that the notion of "starred" or "associated" tone might imply not only the visible acoustic details of temporal alignment with the segmental string, but also the perceived location of such targets. It is an empirical question how such target locations are mapped onto each other.

Finally, the notion of secondary association of edge tones, dating back to Pierrehumbert and Beckman (1988) and their work on Japanese, has been recently exploited to account for the peculiar behavior of phrase accents in a variety of languages. In Pierrehumbert and Beckman's original interpretation, some phrasal tones (i.e., phrase accents) can seek a secondary association with a TBU (see Pierrehumbert and Beckman (1988)), apart from being associated to the specific phrasal domain they belong to. A notion of secondary association is also employed by Gussenhoven (2000) in his analysis of boundary tones of Roermond Dutch (though he does not distinguish between phrase accents and boundary tones). In particular, Gussenhoven proposes that alignment and association are different

phonological mechanisms, though this viewpoint will be discussed later, in §6.2. Both alignment and association are here reinterpreted as constraints on output well-formedness. This is desirable within recent theories of constraint interaction (cf. Prince and Smolensky (1993)). The status of phrase accents and their secondary association with TBUs is also explored by Grice et al. (2000), especially regarding what they refer to as the "Eastern European Question Tune". Such notion of secondary association will be employed here to describe the falling section of the rise-fall configuration of Neapolitan questions and statements.

In sum, as the above examples show, the concept of association is embodied in rather different ways across the AM approach to intonation. Because of the terminological indeterminacy and the lack of explicitness of some of the accounts, alignment and association are either equated, freely interchanged with each other, or strictly segregated to different roles in the grammar. Nevertheless, the need to rely on some superficial manifestation of alignment in order to figure out pitch accent structure, especially when the phonological analysis is at preliminary stages (see the position expressed in Grice et al. (in press)) is recognized. It is therefore desirable that the phonological notion of "association" would relate to specific and well defined physical cues, in a way that echoes the notion of "specific property" of Jakobson et al. (1952) for segmental phonology. In other words, the "specific property" (or better, the "properties") of association should somehow be defined. The idea proposed in this thesis is that the notion of phonological association should rest as much as possible on empirically testable regularities, such as the mapping between acoustic and "perceptual" tonal targets. Note that the role of perception in shaping our idea of association does not exclude the role of production regularities in the representation. Most

obviously, there must be a basis in production for some property to be perceived; vice versa, if there is no robust perceptual grounding to a property, then the pattern will be marked in production.

The point of view that I think is most defendable is the one by which we should view association simply as "contrastive alignment". That is, while phonetic alignment is the measurable set of phenomena we are here concerned with, association is just the effect of contrastive alignment as a polarizing force acting on the observable phonetic variability (both in the production and the perception domain). I will then propose that constraints on alignment will determine the exact temporal location within the stressed syllable for both tones of the question and the statement LH rise in Neapolitan. Note that such "alignment constraints" should not be identified with the family of phonological constraints proposed by the Theory of Generalized Alignment (McCarthy and Prince 1993). In Chapter 3 I will show that such alignment is perceptually contrastive, while in Chapter 5 I will show that it is quite robust, allowing the parsing of prosodic structure even when syllabic structure and segmental environment render the contrast superficially less straightforward. Some of the factors affecting the acoustics of tonal alignment are reviewed next.

1.4 Acoustics of tonal alignment

A coarse classification of the major factors affecting tonal alignment in production is as follows:

- Prosodic and tonal factors

1. Tonal environment (upcoming or preceding tones, tonal crowding effects, vicinity to a prosodic boundary, stress clash).
2. Pitch range.
3. Focal structure.
4. Syllabic structure (onset duration, rhyme duration)

- Speech rate (but see (Ladd, Faulkner, Faulkner, and Schepman 1999))
- Segmental effects (intrinsic vowel duration, consonant voicing)

I will start by discussing some of the prosodic and tonal factors. In his treatment of alignment in Swedish word accents, Bruce (1977) stressed that the prototypical fundamental frequency (f_0) contours of Accent I and Accent II appear only when the words are 1) in focus and 2) in sentence-final position. In fact, when in initial position, the fall of Accent I or Accent II can be truncated and realized as a simple initial low, which is then immediately followed by the sentence accent (if the word is in focus). The shape of the same word accents in other sentential and focus contexts can therefore be different due to, among other things, "time-dependent" adjustments. For instance, the copresence of a word accent and a sentence accent on the same syllable (e.g. when the syllable is final) leads to a situation of tonal crowding that is resolved either through temporal readjustments of peak location or through melodic readjustments of the f_0 level reached by the tonal target. For instance, the fall in Accent II is anticipated when immediately followed by the sentence accent, suggesting that the sentence accent command has interfered with the word accent command. These and other effects are accounted for in terms of "interference" between accent commands, but the hypothesis is not mathematically formulated in a testable way.

Bruce also observed that the timing of the sentence accent is more heavily dependent than word accent on contextual factors. In fact, the sentence accent shows an influence of both preceding and following tonal context. Usually, the beginning of the sentence accent rise is delayed when it is preceded by a word carrying Accent II (in which the fall is aligned later than in Accent I). Conversely, an immediately following Accent I word has the effect of anticipating the end of the sentence accent rise. The fact that the sentence accent of Swedish undergoes more drastic timing readjustments than word accents lead Bruce to propose the generalization that "sentence prosody" is less stable than "lexical prosody". But in English, this is not generally true. In fact, the target of starred tones can actually be displaced in time because of the vicinity of another pitch accent or a boundary tone (Steele 1986; Silverman and Pierrehumbert 1990), while boundary tones tends to stay anchored to the edges of the prosodic domain.

Recently, in fact, some researchers (cf. Silverman and Pierrehumbert (1990, Prieto et al. (1995)) have shown that it is possible to quantitatively model the alignment of certain F_0 peaks and valleys with stressed syllables. These studies concentrated on the interaction of a number of diverse factors, such as duration of stressed syllable, distance to next stressed syllable, distance to the end of the word, etc., in determining systematic alignment differences. The most extensive study regarding the alignment of pitch accents conducted on English is Silverman and Pierrehumbert (1990). Systematic variation of tone target alignment is the object of this work. In a previous study (Steele 1986) it was found that nuclear accent peaks tend to be aligned much earlier relative to the onset of the stressed vowel when the stressed syllable is final in the word. Rate also interacted with peak delay, in the direction of later peaks for longer vowels. However, final syllables always presented

shorter peak delays than non-final syllables at equal speech rate, suggesting that rate and right-hand context affect alignment in rather different ways. The authors extend the study to prenuclear accents, with the goal of showing that prenuclear and nuclear accents are not inherently different (as Pierrehumbert's theory assumes), but are selected from the same set of pitch accent types. Similarly to the findings for nuclear accent timing, they found that the peak delay of prenuclear peaks is greater in slower syllables than in syllables where the lengthening is induced by vicinity to an upcoming stressed syllable (stress clash) or a word boundary. However, the authors also found that, all things being equal, nuclear peaks are aligned earlier in time than prenuclear ones. They propose that the reason for the discrepancy is that nuclear accents are immediately followed by a L tonal command for the end of the utterance, which might "push" the nuclear peak backwards.

Two of the hypotheses tested in Silverman and Pierrehumbert (1990) are those of "tonal undershoot" and "tonal repulsion". Briefly, both hypotheses stem from the idea, already very common in the segmental literature (Lindblom 1963; Lindblom and Studdert-Kennedy 1967), that two articulatory gestures in sequence, can interfere with each other when too close in time (*gestural overlap*). For instance, as a coarticulatory effect of vowel production, formant targets can often be "undershot", namely the actual formant value reached within utterance production is often different from the one obtained when the vowel is produced in isolation. Therefore, it is possible that a tonal target, if produced as a result of two competing tonal gestures, such as a rise and an immediately contiguous fall, might be undershot, i.e., the target f_0 value reached is lower than expected. Another possible outcome of gestural overlap is that, while the "underlying" target is attained in the melodic domain, it is instead anticipated in the temporal domain. That is, the peak or

valley might be displaced earlier in time ("tonal repulsion"). Both hypotheses of tonal undershoot and tonal repulsion, as possible outcomes of gestural overlap, were entertained by Silverman and Pierrehumbert (1990). However, neither of these hypotheses scored better than the other in Silverman and Pierrehumbert's results of data modeling.

But various points need to be made regarding this work. The first one concerns a general issue of choosing the right metrics for measuring tonal targets. The authors choose in fact to measure the shoulder in the f_0 contour near the accented syllable as the phonetic manifestation of the H target. This choice was in a sense dictated by their model assumptions. In fact, the theoretical framework employed by them is Pierrehumbert's model of intonation (Pierrehumbert 1980), which is what can be defined as a "target-interpolation" model. That is, tonal targets are defined in the time dimension, and rises and falls in the melody are obtained just by interpolating between a set of specified targets. Other possible choices, such as measuring rise slope or similar acoustic phenomena as indices of target location, were therefore excluded. The model adopted excluded also the possibility of measuring the location where the rise becomes steeper, as an indication of target location for the beginning of a rise, i.e., LH, gesture. This is because that would have implied that prenuclear H's are preceded by a L target. But, in the intonational analysis of English adopted by them, prenuclear accents are simple H tones, not LH combination.

Measuring peaks, though apparently a trivial task, turned out to be very difficult, though, especially because of cases where two competing yet plausible peak locations where found in the region for the H. For instance, it is very common for the f_0 to show dips caused by nasal segments, as it was the case in some target words of the study. Such dips make it difficult to select a unique peak location, since, in principle, both the preceding maximum and

the one following the dip are acceptable choices. In such cases, the authors measured the H target at the midpoint between the two observed peaks. Though consistency in measurements is to be appreciated, we do not know if such a measure is a good indicator of target location (see also criticism in Bruce (1990)). A possibility to entertain is, in my opinion, that by measuring both target H and the preceding valley one would obtain a better idea of the global accent configuration, which seems to be important in considerations of "invariance". As Bruce (1990) observes in his commentary, I believe that, though the pitch accent investigated is analyzed as a peak accent (H*), it is still manifested as a pitch rise in the actual phonetic representation (the pitch curve). It is an empirical question, then, whether the starting point of the rise would show as much variability as its target.

An outcome of the phenomena of tonal repulsion and undershoot proposed by Silverman and Pierrehumbert can be discerned, on one side, in the truncation of final low targets in rise-fall gestures. An example is found in Southern varieties of Italian such as Neapolitan. Compare the final rise-fall of the question *Vedrai il NANO dopo?* "Will you see the DWARF afterwards?" (capital letters indicate nuclear accent placement) with that of *Vedrai MA?* "Will you see MOM?", respectively in the upper and lower panel of Figure 1.8. In the first case, where there are three syllables after the accented one, the rise-fall is fully realized. When there is less material following the accent, two outcomes are possible. When the rise-fall is placed on a monosyllabic word (i.e., a syllable that is not followed by additional segmental material), as in the word *ma*, the fall section can be curtailed. This situation is shown in the lower panel of Figure 1.8. Such "truncation" is also found in Palermo Italian (Grice 1995a). On the other hand, it is possible to compress the entire rise-fall movement, and temporally anticipate the location of the peak even on

a word such as *NANO*, when this is final in the intonation phrase, as we can notice in the lower panel of Figure 1.9. In this Figure the utterance *Vedrai il NANO?* "Will you see the DWARF?" is compared to the utterance *Vedrai il NANO dopo?* (which is repeated for ease of comparison). The compression effect can be seen as an outcome of tonal repulsion.

Hence, it appears that, in Neapolitan, truncation can be found in intonation phrase final monosyllables (see Figure 1.8 above). Compression (and, consequently, tonal repulsion) is found in intonation phrase final words (see Figure 1.9 above). The degree of tonal repulsion does not seem to depend, however, on the number of syllables to the right of the accented syllable (D'Imperio 1995), though such a result is controversial (see D'Imperio (1998a) and D'Imperio (in press)).

Also, in Neapolitan Italian, when the constituent in focus is made of a single word[3],the fall of the rise-fall pattern of interrogatives occurs immediately after the pitch accent rise and seems to mark the end of the focus constituent. When the focus constituent is longer, the rise and fall appear to separate, with the rise staying anchored to the focal initial stressed syllable while the fall occurs later, reaching its target in the vicinity of the constituent right boundary. Figure 1.10 shows such a contrast by using different words. The utterances depicted are *Vedrai [mamma] domani?* "Will you see [mom] tomorrow?" with narrow focus on the word *mamma* (upper) and the utterance *Vedrai [la mano di Mammola] domani?* "Will you see [Mammola's hand] tomorrow?", with narrow focus on the longer NP constituent *mano di Mammola*. Notice in the lower panel of Figure 1.10 how the pitch stays high after the word *mano*.

[3]For details of the Neapolitan pitch accent inventory, see §2.1.

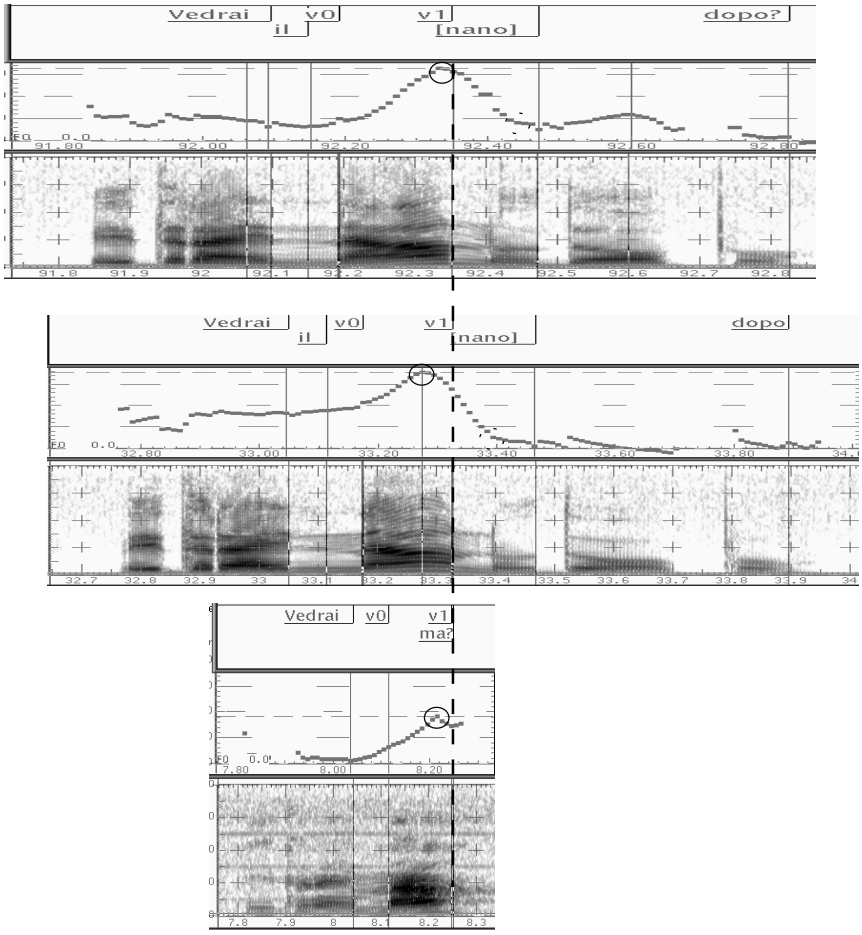

Figure 1.8: Tone labels, f_0 curve and spectrogram for the utterances *Vedrai il NANO dopo?* (upper), *Vedrail il NANO dopo* (middle) and *Vedrai MA?* (lower). Stressed vowel offsets are lined up at the dashed line.

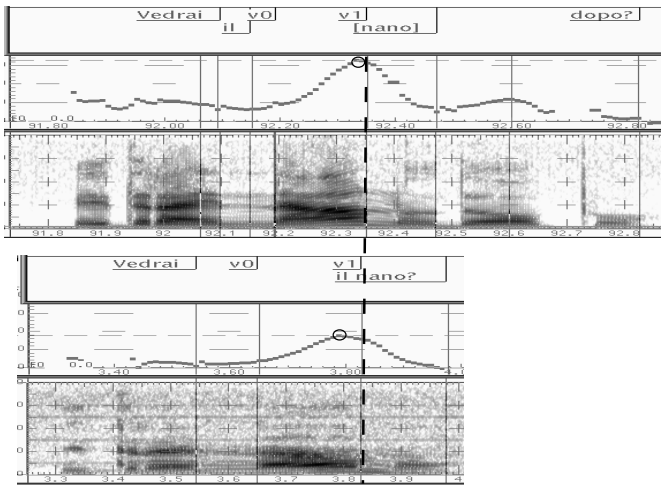

Figure 1.9: Tone labels, f_0 curve and spectrogram for the utterances *Vedrai il NANO dopo?* (upper) and *Vedrail il NANO?*. Stressed vowel offsets are lined up at the dashed line.

A production study (D'Imperio 1998a; D'Imperio in press) concentrated on the properties of the final constituent fall (as well as the initial rise) in early focus interrogatives with different focus constituent size. It was found that the final fall (which I analyze as a HL; see §2.1) is anchored to the last stressed syllable (when it is available) of multi-word focus constituents, thus resembling a regular pitch accent. However, when there is only one stressed syllable in the focus constituent, the existence of two separate H targets is obscured, since there are not two separate docking sites for them. In this case, only a high-fall is found. Specifically, it was also found that the target for the HL is reached later in single-word constituents, since it is as if it were "pushed forward" by the nuclear L*+H. It was also found that the initial rise of the focus constituent reaches its H target later in

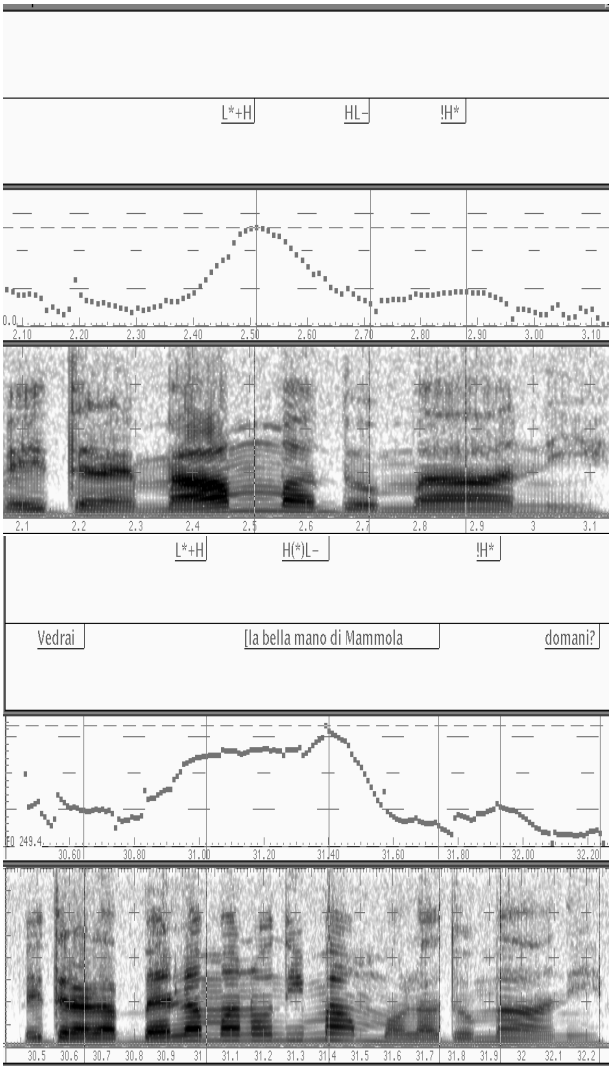

Figure 1.10: Tone labels, f_0 curve and spectrogram for the question *Vedrai MAMMA domani?* uttered with narrow focus on *mamma* (upper) and for the question *Vedrai LA MANO DI MAMMOLA domani?* uttered with narrow focus on the constituent *la mano di Mammola* (lower).

multi-word than in single-word constituents, which is taken as evidence for a "tonal repulsion" effect (Silverman and Pierrehumbert 1990) that produces temporal anticipation of this target when immediately followed by the HL tone (i.e., in single-word constituents). The tonal repulsion hypothesis, in fact, predicts that when two competing tones have to be realized on the same TBU (such as a stressed syllable, in our case), the target for the first tone will be moved earlier in time to make space for the second tone to be realized. An alternative strategy for the resolution of "tonal crowding" consists in not reaching the f_0 target level (tonal undershoot, (Silverman and Pierrehumbert 1990)). This strategy is, however, not reliably employed by the speakers analyzed so far.

The peculiarities of nuclear accent alignment had actually already been noticed by Silverman (1987). In his effort to apply Pierrehumbert's model of intonation to the implementation of a synthesizer, Silverman (1987) is faced with a conundrum: how to preserve the principle that prenuclear and nuclear pitch accents are drawn from the same pitch accent inventory and, yet, realize the difference between their peak alignment that is found in natural speech. What Silverman found was in fact that nuclear accents must be aligned much earlier in the syllable than prenuclear ones in order to sound natural.

Silverman (1987) also explores the contribution of segmental factors on the realization of the f_0 contour. Though Silverman did not specifically look at timing of peaks, his results are still relevant here in that they reveal something about the influence of segmental perturbation on the "shape" of pitch contours after stop release.

He finds in fact support for the hypothesis that the underlying intonation structure of an utterance actively interacts with segmental perturbations. His prediction is based on a layered model of f_0 production, where segmental and intonation perturbation are in a sort

of "trade off" relationship. A note on the contribution of Silverman to the naturalness of f_0 synthesis is in order. In his implementation of the target-interpolation model, Silverman (1987) criticizes text-to-speech implementations of tonal targets as very sharp peaks (a target might be realized, say, only in the space of a 10 ms-frame), as in Anderson et al. (1984). As Silverman notices, "Natural speech f_0 contours do not exhibit such sudden changes in direction, except when these are indicated by segmental perturbations" (p. 5.24).

The avoidance of sharp peak contours in synthesis has also been motivated from a perceptual point of view. 't Hart (1991) reports on the results of a perception experiment intended to evaluate discriminability of rise-fall patterns created by stylizing the pitch contour in a linear or a parabolic fashion. The linear versions, additionally, could either present a sharp or a "flat" peak. Subjects are reported as being unable to perceive differences between the linear and the parabolic versions. This was especially true when flat peak stimuli were compared to the parabolic ones. This leads the author to justify the common practice at IPO of straight-line stylization, with the addition of a short plateau (of 30-40 ms). Interestingly, the author comments that in order for the sharp peak stimuli to be perceived as identical to the parabolic ones he had to employ a "somewhat higher peak", the reason being that in those stimuli the peak is maintained during only one period. But details about how much higher the peak needed to be are not offered. Therefore, such a study supports the idea that linear transitions are perceptually undistinguishable from parabolic ones, with the proviso that they are stylized with a flat peak.

More recently, van Santen and Hirschberg (1994) showed that the timing of pitch contours can be modeled as the result of weighted effects of intrasyllabic segmental composition as well as syllable onset and rhyme duration. The key proposal of the work is a model

of time-warping that intends to reproduce an "accent curve" for a specific pitch accent in English. Within the accent curve, the authors model the timing of "anchor points" located at specific proportions of the entire accent gesture. The contour is warped in a non-linear way, so that earlier anchor points location will depend more on onset duration, and the location of later anchor points will depend largely on rhyme duration. The authors recognize one limitation of this study, which tested only very simple intonational contours, with one pitch accent (H*) and one boundary tone. An interesting feature of this work is the use of the phonological notion of "rhyme". In their measurements, the duration of the rhyme is made to start from (and includes) the last sonorant of the onset (if there is one). In other words, they explore the possibility that alignment is made relative to units that are not necessarily traditional phonological constituents. Analogously, it has been argued that alignment of prosodic elements (such as pitch accent alignment relative to specific prosodic domains) to each other need not be made only relative to constituents in the classic sense. Pitch accents, for instance, can be aligned relative to the edges of constituent heads (see Pierrehumbert (1993). It becomes then interesting to explore which are the units (both phonetically and phonologically) that are relevant in matters of tonal alignment, especially regarding the role of invariant "segmental landmarks".

Recently, Arvaniti et al. (1998) show that rising prenuclear accents in Greek begin at the onset of the stressed syllable and peak early in the following unstressed vowel, independent of syllable composition and duration. Analogously, Ladd et al. (1999) show that the beginning of the f_0 rise in English prenuclear accents is consistently aligned with stressed syllable onset, while the end of the rise is aligned with a location within the segmental stretch spanning from the end of the stressed syllable to the onset of the following vowel.

Such findings are taken by Ladd et al. (1999) to support the hypothesis that tonal alignment is the main feature of "accent identity" and that other features, such as "accent shape" or "accent duration" are not independently specified and controlled by the speaker. This issue will be explored in the experiments presented in this thesis. At the same time, other recent work has emphasized the importance of fine differences of alignment for conveying language-specific differences. Most of this work has been conducted through perception experimentation, therefore it will be reviewed in §1.5 and §1.5.1.

In sum, the temporal alignment of f_0 targets seems to be affected by a number of variables that can render intonation analysis quite difficult when the language system of contrasts is still unknown.

1.5 Perception of tonal alignment

Though general issues of pitch perception will be reviewed in §1.6, here I would like to discuss only some key linguistic studies that have shed light on the perception of tonal alignment. In particular, Pierrehumbert and Steele (1987) (see also Pierrehumbert and Steele (1989)) conducted a study that happened to be crucial for subsequent work on alignment. The authors investigated the timing properties of the "rise-fall-rise" pattern in English, represented by the two contours L*+H L-H% and L+H* L-H%.

The two pitch accents L*+H and L+H* can cue in English either an "incredulous exclamation" or a "forceful assertion", respectively. A formalization of this meaning distinction can be found in Ward and Hirschberg (1985). Indeed, Pierrehumbert and Steele (1987)

found that a categorical distinction exists between an utterance such as “Only a millionaire!” with a L*+H associated to the stressed syllable “mil-”, and the same sentence uttered this time with an associated L+H*.

It had been a matter of dispute whether the alignment of the peak would be gradiently or binarily manipulated within the rise-fall(-rise). The prediction was that, if the distinction is found to be binary this would suggest the existence of two pitch accent types, one with an early peak (L+H*) and the other with a late peak (L*+H). On the other hand, Ladd (1983) and Gussenhoven (1984) maintain that there is just one phonological category “rise-fall”, with a gradient variation in peak delay signaling a gradient variability in the pragmatic meaning (in an iconic way, analogous to the way that degree of pitch range expansion reflects “degree of involvement”). However, Pierrehumbert and Steele (1987) succeed in showing that the distinction is binary and not continuous, with L+H* cuing assertion and L*+H cuing incredulity.[4] The authors created stimuli through LPC resynthesis, in which the accent peak was delayed in steps of 20 ms (from 35 to 315 ms from the end of [m] in “millionaire”), while rise and fall duration were kept constant. Then they asked listeners to reproduce the stimuli heard. The subjects were recorded, and peak alignment was measured in their productions. The prediction was that if the contrast is categorical, the imitations would present a bimodal distribution, with ambiguous stimuli assimilated to the endpoints. If Gussenhoven (1984) were correct, the expected distribution of peak delays should be flat,

[4] It must be noticed, though, that Ward and Hirschberg (1985) hinted at another possible meaning for L*+H, that is “uncertainty”. The possibility of homophony for L*+H was then experimentally tested in Ward and Hirschberg (1988), which supports the polisemy hypothesis and finds that the incredulous reading has a more expanded pitch range than the uncertainty reading.

reflecting the continuous variation in peak delays in the mimicked stimuli. The hypothesis of bimodality was instead supported by the data.

They also noticed that the boundary between the two categories was rather late, i.e., between stimulus 10 and 11 in their series of 15 stimuli (where the stimulus with the higher number has later alignment). This might be because L*+H is in some sense a more marked tonal pattern; or, alternatively, because the subjects were biased towards L+H* responses because of some property of the original stimulus used for resynthesis. Some of these properties could be rise-fall shape, spectral tilt and relative amplitude. The authors suggest that these might be secondary cues to pitch accent category. I believe that the second hypothesis is very plausible, and that it would be interesting to explore the role of these secondary cues, and, above all, investigate under which conditions they might become "primary" in signaling pitch accent category. One of the perception experiments presented in the present study will involve base stimuli with inherent different properties. The possibility of an overall effect of "base type" is in fact investigated in Chapter 5 of this dissertation. Pierrehumbert and Steele also observe that the target for the L seems to be displaced later when H alignment is also later, but their results for the L target were not bimodal. This result leads the authors to admit that the L targets were probably not properly measured. Again, a problem in target measurement difficulty leads to uncertainty in the interpretation of the results. Finally, peak delay did not cause segmental durations to be longer.

Kohler (1987) can be considered as the first work showing that perception of tonal contrast can be categorical, in the sense used in segmental research. He employed the complete paradigm of categorical perception (both identification and discrimination tasks), by creating stimuli in which the pragmatic values "established", "new" and "emphatic" are manipulated.[5] The manipulations were obtained by shifting the f_0 peak in the German word "gelogen" (in the context of the sentence *Sie hat ja gelogen* "She's been lying"). The peak was shifted rightwards, in 30 ms steps, and could either occur before the stressed syllable "lo", within it or at the very end. However, it appears that the discrimination function was above 50% only when the stimuli were separated by two steps. This means that circa 60 ms are required in order to shift alignment perception. In a subsequent experiment Kohler (1990) manipulated f_0 peak alignment in order to shift perceived "word stress". However, it is not clear if the stimuli were totally neutral as to possible secondary cues. Though Kohler concludes that f_0 (peak) can directly cue stress "provided the duration of the stressed-syllable-to-be is not too short" (p. 137), there is a *caveat*. In fact, in the stimuli in which only f_0 was manipulated in order to shift stress, there was still a residual effect of duration (which the author accounts for in terms of preboundary lengthening) that might have helped the listener, together with spectral cues, to infer the metrical composition of the word.

[5] Note that in an autosegmental-metrical approach these three configurations might be represented as H+L* (fall from peak onto accented syllable) vs. (L)+H* L% (peak on accented syllable) vs. L*+H L% (late peak relative to the accented syllable).

In sum, no stress shift was obtained (even when an f_0 peak was placed on it) when the syllable of the base stimulus was short. This is revealing of the need for higher control in stimulus preparation for perception experiments, as well as the need to make clear the type of configurational contrast that one wishes to investigate.

Rietveld and Gussenhoven (1995) investigate the effect of segmental composition of the syllable on the alignment of peak accents (H*) in the Dutch "flat hat pattern". Two variants of this pattern, one with an early fall and the other with a late fall, have been described in the literature. While 't Hart et al. (1990) consider the two realizations of the hat pattern as contextual variants of the same contour, recently it has been found that they cue two different pragmatic meanings, with the early fall (whose initial high is also downstepped) conveying a sense of "definitive contribution to the discourse" on the part of the speaker. The question of the authors is then how to implement this difference in their synthesis system so that peak alignment is successfully conveyed.

Earlier, the authors had modeled peak alignment just as latency from stressed vowel onset. But the segmental composition of onset and coda had been found to shift perceived peak alignment for the same latency. In the 1995 paper, therefore, they create stimuli with varying onset and coda structure, so that there were from zero up to three consonants in the onset, and either a sonorant /m/ or an obstruent /p/ in the coda. The listeners' task was to identify either a "quiet, low-pitched" or a "more emphatic, high-pitched" contour. Such a task is prone to misinterpretation, though, since the listeners would concentrate on just one physical variable, i.e., pitch. It would have been more appropriate to have the listeners match a discourse context to the heard stimulus. In any case, proportion of "quiet, low-pitched" responses was employed to calculate "points of subjective equality" (PSE).

Then, the authors carried out a multiple regression in order to discover which factors affect the position of the PSE, including the following nominal and continuous measures: coda voicing, onset duration, duration of voiced section of the onset, P-center.[6] What they found was that the first three factors significantly contribute to explain the variance (precisely, 89% of the variance), but the contribution of P-center was not significant. It seems that the presence of voiced segments in the onset and in the coda determines the PSE location, with voiced codas pulling the PSE to the right, and voiced onsets pulling it to the left. It is interesting to notice that when both onset and coda are voiced, the PSE is right in the middle of the stressed vowel. The authors, however, did not control for total duration of the coda, whose contribution cannot therefore be evaluated. On the contrary, they claim that the total duration of the onset, independent of voicing, heavily affects PSE placement. This finding for perception seems to be in accord with the findings for production of van Santen and Hirschberg (1994), and it supports the notion of a "rhyme", governing alignment, which starts at the first sonorant segment of the syllable (including segments in the onset).

More recently, D'Imperio and House (1997) showed that in Neapolitan Italian, the intonation patterns for statements and questions are distinguished primarily by the alignment of the peak of a rise-fall pattern on the main stressed syllable. Since this experiment will constitute the basis for the perceptual manipulations presented later in this work, I will cover its details in more depth.

[6]The "P-center" is intended to reflect the "psychological moment of occurrence of a syllable". This measurement was obtained by applying an algorithm developed by Pompino-Marshall (1989).

1.5.1 Perception of questions and statements in Neapolitan Italian

The materials in D'Imperio and House (1997) consisted of a series of stimuli in which the f_0 peak of a rising-falling pitch accent (such as the one shown in Figure 1.11) was shifted forward or backwards within the stressed syllable ("peak-shift" manipulation). Since both a declarative base and an interrogative base stimulus were employed, the peak-shift series based on the declarative was called "decl-peak" and the interrogative based series was labeled "inter-peak" (see Figure 1.11). In another manipulation the fall of the declarative base was shifted without changing the rise, or the rise of the interrogative base was shifted without changing the fall. Here, the declarative based series was called "decl-fall" and the interrogative based series "inter-rise" (see Figure 1.12). Neapolitan Italian subjects listened to all stimuli and identified them as either a question or a statement. The results are shown in Figure 1.13), where percent of "statement" responses is plotted for the four stimulus series. Note that by shifting the peak at later timing continua in both decl-peak and inter-peak series (the solid and dashed line, respectively) caused an increase in question responses which appears to be categorical. In fact, stimulus 4 and 5 in both series shown in Figure 1.11 elicited around 100% of question responses. At a closer look, though, declarative base stimuli produced a higher number of statement responses at early peak timing locations than interrogative base "peak shift" stimuli, while more question responses were obtained for late peak timing in question base stimuli as opposed to declarative base ones. The effect of base stimulus on the perception of pitch accent category is investigated in Chapter 5 of this dissertation.

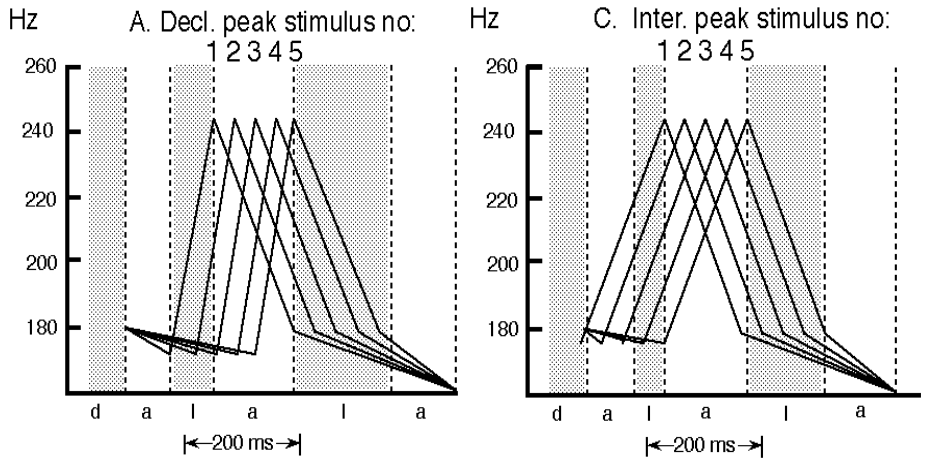

Figure 1.11: Stimuli in which the peak is shifted across the stressed vowel (from D'Imperio and House, 1997.)

The important point made in that study (which is also the most relevant for our purposes) was based on an implicit (and rather unjustified) assumption. D'Imperio and House in fact assumed that the target of a rising LH accent corresponds to the terminal point of the rise. While this assumption seems fairly unproblematical when both rise end and fall beginning coincide, as in the series shown in Figure 1.11, it becomes obviously questionable when the accent shape includes a "plateau", as in the series of Figure 1.12.

Equating rise end with target peak location lead to a paradoxical interpretation of the results of the inter-rise stimuli, shown in the right panel of Figure 1.12. Let us look closely at the details of these stimuli. In this series, the rise of a resynthesized question was shifted backwards within the stressed syllable in 35 ms steps. As we can see from Figure 1.12

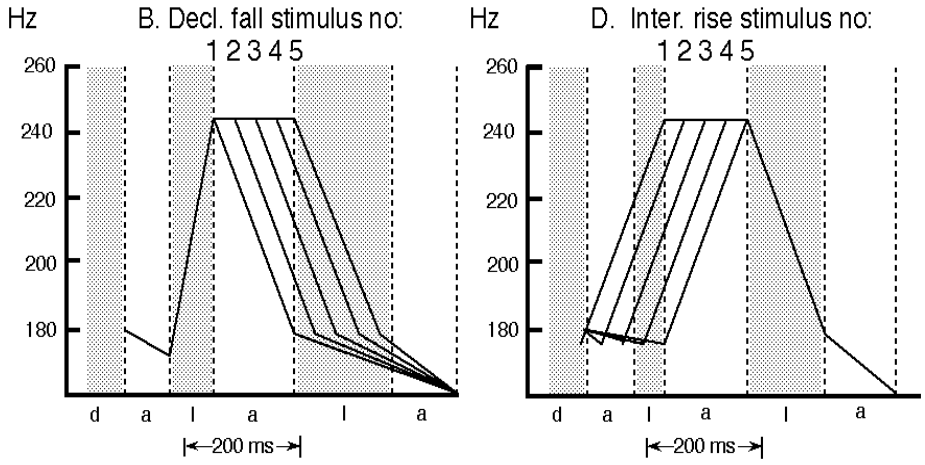

Figure 1.12: Stimuli in which the fall or rise is shifted across the stressed vowel (from D'Imperio and House, 1997).

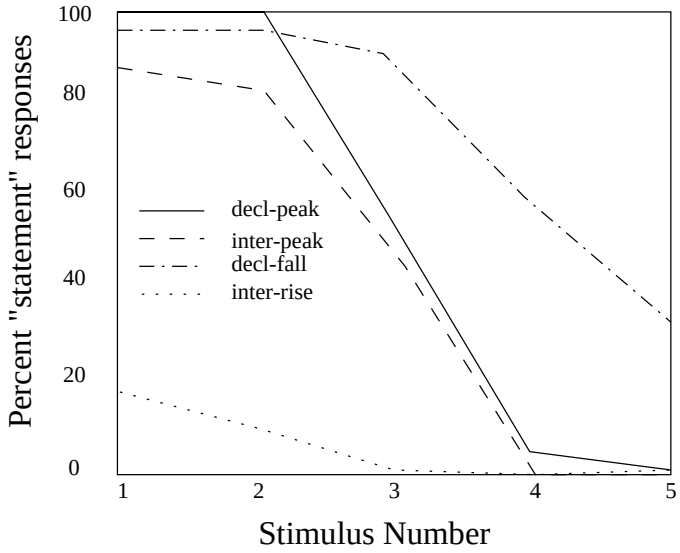

Figure 1.13: Results from D'Imperio and House (1997).

(right), the first stimulus has a rise terminating at the onset of the stressed syllable (unshaded central area). Therefore assuming that: 1. (as we saw above) Neapolitan questions are characterized by a late and not an early H target location and that 2. the end of the rise is the "target" of this accent configuration, we predict that we will obtain high percentages of statement responses for early continuum stimuli. However, as one can notice in Figure 1.13, a very high percentage of question responses was already obtained for stimuli at early continuum locations (around 80% for stimulus 1 and around 90% for stimulus 2). Thus, the location of the high target, by the implicit assumption assuming that it coincides with the end of the rise, cannot explain the pattern of results.

On the basis of this assumption, the authors suggested that peak timing cannot be the strongest cue for question/statement identification. Hence, D'Imperio and House (1997) suggested that rise perceptibility, and not (peak) target alignment, is the key. In other words, they proposed that it is a perceptible *rise* within the stressed syllable that cues questions more strongly that any H target timing location. Such a result was also interpreted as having implications for the old "level vs. configuration" view of intonational primitives (see Ladd (1996), section 1.2, for a review). The Dutch school of intonation, for instance, supports the primacy of "configurations" (rises or falls) as the primitives of intonation, and D'Imperio and House (1997) would, in a sense, offer perceptual support to such a view.

However, the results of that study were not conclusive for various reasons. In fact, different explanations can be invoked in order to account for the inter-rise results which are based on different hypotheses regarding the location of the H target. One hypothesis relates to the the "shape" of the accent. The rise-shift stimuli were in fact characterized by a high f_0 plateau, which resulted from splitting the rise and the fall in stimuli 1-4 (see Figure 1.12,

right). One obvious hypothesis would therefore be that in "plateau" stimuli the perceived target is not located at the end of the rise, but it is located elsewhere. Plausible locations would be the middle of the plateau, the end of the plateau, etc. Another hypothesis, based on informal findings in 't Hart (1991) and psychoacoustic data reported in Nábeľk et al. (1970) (which will be discussed in §1.6), would be that peak stimuli have a peak target that is perceived as being lower in pitch than those of plateau stimuli, even when the end of the rise reaches exactly the same f_0 peak value (but see §3.1 below). This hypothesis was tested in Chapter 3. Another hypothesis is that the location of the perceptual target is displaced in the time domain, and not (or at least not only) in the melodic domain. But the details of such a hypothesis will be clear after some peculiarities of tonal perception will be reviewed in the next section (§1.6).

A yet alternative explanation is based on rise slope gradient. Notice in Figure 1.11 (left) that the rise is much steeper for the decl-peak stimuli than for the rise-shift stimuli (Figure 1.12, right). Therefore, D'Imperio and House (1997) speculated that a shallower rise might be a cue to the question category. The import of slope in production and in perception are evaluated, respectively, in Chapter 2 and Chapter 3.

Hence, some of the hypotheses presented above were tested in the perception experiments presented in this thesis. For now, it will suffice to say that we need to distinguish between various explanations for the greater "questionhood" of rise-shift stimuli, one obvious possibility being that the perceived target of such stimuli might not correspond to any of the targets traditionally measured in the f_0 contour, such as the maximum value at rise end, or the peak just preceding the beginning of a melodic fall.

In sum, through f_0 event timing manipulations alone, D'Imperio and House (1997) were able to induce a clear question and a statement category, with a region of ambiguity at the center of the continuum. On the one hand, then, there is much about tonal alignment that is predictably influenced by the phonetic context. On the other, though, small differences of alignment can apparently also create clearly perceptible differences of meaning. In order to understand how both these things can be true, it is necessary to know more about how differences in tonal alignment are perceived. Hence, one of the purposes of this research will be to compare perceptual strategies in listeners speaking different languages, such as Neapolitan Italian and American English.

1.6 Psychoacoustics and "universals" of tonal perception

A question we need to ask is "what are the limits of our auditory system when it comes to perceiving pitch?". Indeed, I subscribe to the view that if a phenomenon cannot be heard, it cannot play a role in speech communication. However, caution must be exercised regarding the results of psychoacoustic studies, since they "stretch" the limits of our system in reflecting performance levels that are never attained in normal speech situations. Here I will present just some results of psychoacoustic studies on dynamic tones since they will bear upon some of the issues tackled in the perception section of this thesis.

Pitch perception has been considered for many years to be simply related to the perception of the fundamental frequency of a complex sound. However, the fundamental frequency component need not be present for pitch to be auditorily resolved. Indeed, Schouten

(1940) successfully provided evidence for a theory that had been already around for a century or so, that is the "temporal theory" of pitch perception. According to this theory, nerve fibers would fire according to a pattern established by the intrinsic periodicity of the signal (phase locking). Though energy at the fundamental might be missing (as in Schouten's experiment), one is still able to hear a "residue pitch". A competing theory, the "place theory" of pitch perception, supports instead the idea that the fundamental frequency of a complex sound stimulates a specific region of the basilar membrane. This theory accounts for cases of "missing fundamental" by invoking the effect of nonlinear distortions that would reintroduce energy in the fundamental region. Nevertheless, when energy in the lower harmonics is present (which is to say very often in normal speech communication, apart from when we are talking on the phone), it appears that the region that dominates in conveying a pitch percept is the one that spans the area between the third and the fifth harmonics (Ritsma 1967). In this region, in fact, individual harmonics can be separately resolved. Precisely, this dominance region is defined for signals with a fundamental frequency that is between 100 and 400 Hz and for levels up to 50 dB, which are appropriate values for speech.

At the present state of our knowledge, the controversy between the place and the temporal theory of pitch perception cannot be resolved yet. Moore (1989) states: "We may conclude, then, that the pitch of a complex tone can be derived from neural signals corresponding to individual partials in the complex, at least for stimuli with low harmonics and containing only a small number of harmonics. It is still not established whether the information about these partials is coded in terms of place or in terms of the temporal patterns of neural firing." (p. 180). Since Moore restricts his claim to stimuli with "a small number of harmonics", it is hard to see how this could apply to speech signals, which are instead characterized by a rich harmonic structure (at least for sonorant segments).

The Just Noticeable Difference (JND) for a frequency shift in pure tones has been the object of much research. This is the smallest detectable change in frequency in a stimulus. One of the main findings reported in the literature is that our ability to detect a frequency change is a function of tone frequency. In general, people are able to detect a change between a tone of 1000 Hz and one of 1005 Hz (Ritsma 1965; Nordmark 1968), but it seems that sensitivity decreases at frequencies both lower and higher than 1000 Hz. Also, precision in change detection appears to be lower for shorter stimuli.

Duration has also been recently found not to be significant as an independent factor in determining the "differential threshold" between two tonal glides. Madden and Fire (1997) used a two-alternative forced-choice task in which subjects had to say whether a glide decreased (or increased) in perceived pitch when compared to another glide. Stimuli had roving center frequency, between 0.5, 2, and 6 KHz (start and endpoint were equidistant from center frequency), and duration was either 40 or 400 ms. The authors notice that an effect of duration is present only for certain stimuli (i.e., not for those with roving center

frequency). Direction of the glide showed a general trend for downglides to have higher thresholds than upglides; but this trend was significant only for 50 ms stimuli, with transition span of the comparison signals being equal to 0 ERB (Equal Rectangular Bandwidth). Direction was not significant for any of the other conditions. The issue of asymmetries in direction of pitch movement will be discussed in greater detail below, for more speech-like stimuli.

Dynamic tones (tonal glides) have also been studied in relation to the question "to which tone level is a tonal glide matched when duration and other conditions are varied?". This issue is central to the experiments presented in Chapter 3 and Chapter 4 of this dissertation. Nábeläk et al. (1970) were among the ones that tackled this issue. The authors manipulated four variables, namely frequency change (calculated as the difference between the terminal frequencies), stimulus duration (either 12, 40 or 120 ms), transition duration (duration of the portion of the stimulus in which the frequency varies) and direction of the pitch movement (rising or falling). Specifically, they conducted a series of three experiments in which the subjects had to match a tonal glide to the pitch of a steady-frequency burst. In the first experiment, transition duration was equal to 100% of the stimulus, while in the other two experiments only 25% of the entire stimulus changed dynamically (while the remainder of the stimulus was either on a low or high level frequency). For the 100% transition condition, the authors found that glides are matched to the pitch of a tone with frequency near the glide mid frequency. They also noticed a tendency for glides with greater frequency excursion to be matched to frequencies corresponding to the final frequency of the glide and that this tendency is stronger for rising than for falling glides. On the other hand, the results for glides with only 25% transition duration were more complex.

If the glide was very short, the pitch was matched to a frequency that was in between the terminal frequencies; but when the glide was longer, its pitch was matched to a frequency corresponding to either one of the terminal frequencies. Precisely, long rising glides were matched according to the initial low plateau, steady-state, portion and, vice versa, long falling glides were matched with the initial high steady-state portion.

Similar results were found by Schouten and Peeters (1998), in which glides were matched for similarity to two steady-state tones. Matches in the main experiment were generally positioned at 2/3 of the glide.

Simplified speech stimuli have also been employed in order to find out how pitch perception works. Interestingly, some of this work suggests that complex pitch configurations (such as a rising or a falling pitch movement) are not perceived in their entirety, and that the perceived pitch target corresponds to the fundamental frequency at roughly two thirds of the length of the rise or fall (Rossi 1971; Rossi 1978), analogously to the psychoacoustic data presented above (Nábeläk, Nábeläk, and Hirsh 1970; Schouten and Peeters 1998). However, unlike earlier psychoacoustic studies, Rossi (1978) finds that the same generalizations can be made regarding the perception of falling and rising glissandi in speech. That is, for both rises and falls, the entire frequency change of the stimulus was matched to a frequency corresponding to 2/3 of the stimulus.

Figure 1.14 summarizes the psychoacoustic results regarding perceptual target location in the f_0 domain that can be extended at least to simplified speech stimuli. Basically, rising and falling glissandi in which no stationary transition is present the location of the perceived target will be at roughly 2/3 of the transition (indicated by the circle). On the other hand, when a plateau precedes or follows the glissando, the perceived target will be located within the plateau itself. Note that no remark is made relative to the temporal location of the target, but only to its pitch value.

Whether the type of generalization, though very simplistic, expressed in Figure 1.14 can be taken to be "universal" is a question for research. Work on the perception of intonation is still scarce, and the number of languages that have been studied are very few. Nevertheless, I will attempt to outline the results of some studies that might help us to start answering such a question. Such a problem will also be tackled in this thesis.

On one side, it appears superficially impossible to simply extrapolate from results on pure tones to complex tones and then to speech. This is because speech stimuli are characterized by many complicating factors, such as rapidly changing energy profile, varying spectral configurations and, most importantly here, a continuously varying fundamental frequency. This is a problem for auditory pitch modeling, as House (1990) points out. In fact, the only alternative would be a "step-by-step testing of the influence of all possible speech configurations upon the models" (House, 1990:29), which seems unfeasible. What are the possible solutions, then?

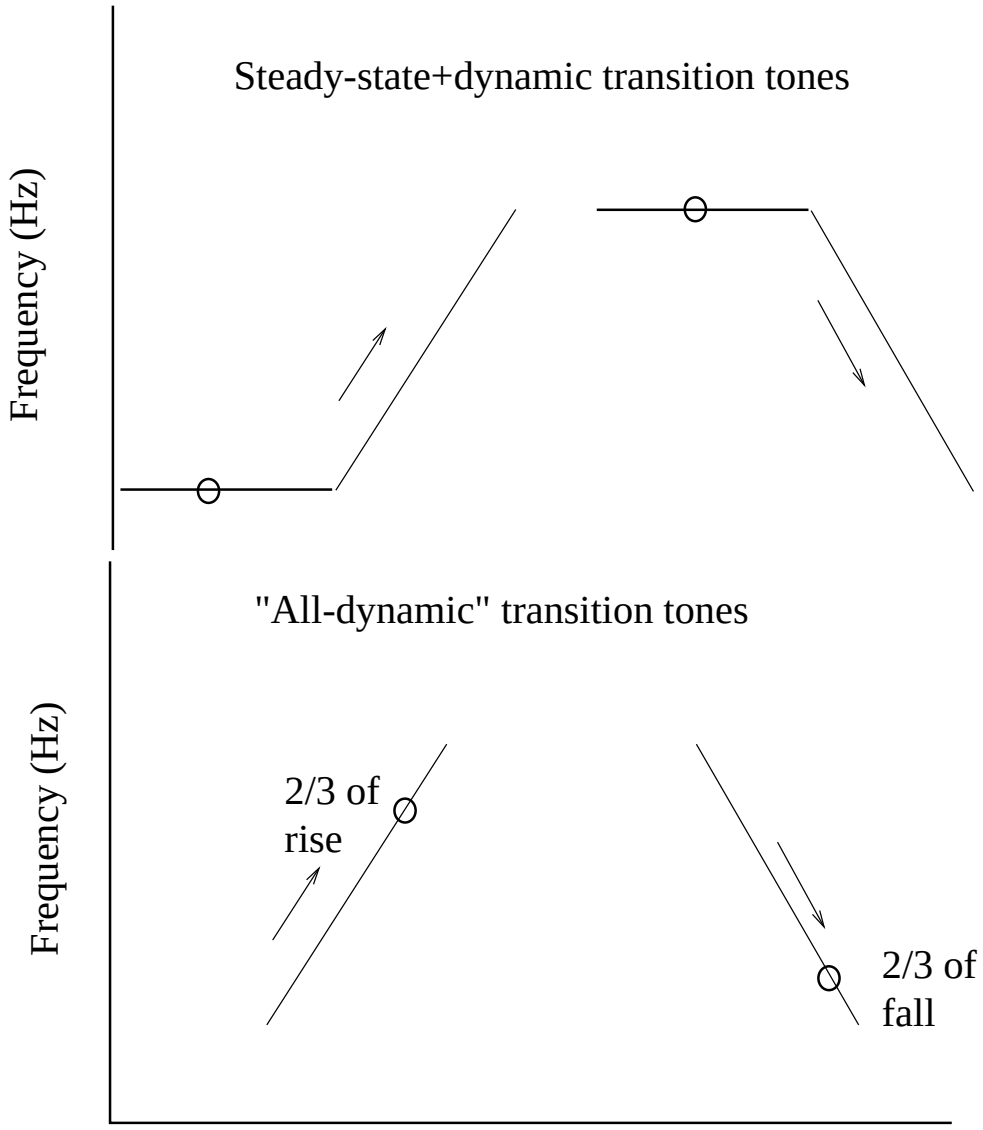

Figure 1.14: Schematic representation of perceptual target location for pitch of glissandi with and without a stationary (plateau) transition.

One proposal, which is certainly appealing, is the one offered by House (1990), which consists in using speech perception experiments to "cautiously infer back to pitch perception mechanisms in general and extrapolate forward, as it were, to perceptually based explanations of linguistically relevant tonal movement in spoken language" (p. 29).

But first, we need to define what we mean by "tonal perception". If we simply refer to the perception of pure tones and complex (non-speech) tones, psychoacoustic results cannot always be safely extended to the perception of speech-like stimuli. Some of the inherent properties of speech sounds that render them very different from psychoacoustic stimuli include:

- drastic energy modulation
- spectral discontinuities
- higher memory load
- influence of higher level linguistic structure

Some of these problems are fairly obvious, but have not been extensively explored. Dramatic energy shifts between vocal and consonantal portions of the spectrum can heavily affect our power for resolving pitch, for instance. 't Hart et al. (1990) quantify this masking effect as being equal to half an octave when the energy jump is in the order of 10–20 dB. Also, in continuous speech, the load on short-term memory is much greater than in the perception of isolated vowels, while the effect of higher level linguistic structure is still open to debate, and will be discussed below. Regarding pitch perception in

speech, the position of the Dutch school is that ('t Hart, Collier, and Cohen 1990, pp. 35–36): "in a communicative situation the listener does not perceive pitch in its proper sense: the variations of f_0 are translated into prosodic properties of the linguistic code, such as accentuation and the like".

As for spectral discontinuities, whether we perceive pitch in the same way in steady portions and unsteady portions of the spectrum is an empirical question that has been tackled by House (1990). Here I will review some results that seem to be in contrast with the hypothesis that intonation/pitch perception is universal.

The issue of spectral discontinuities and energy modulation has been recently investigated by House (1990) and House (1996). House (1990) formulates and tests two working hypotheses that relate pitch sensitivity to changing spectral information and spectral discontinuities such as: consonant release, intensity jump at vowel onset and rapid formant transitions at the beginning of the vowel. The two working hypotheses stem from a main hypothesis, i.e., the Spectral Constraint Hypothesis, which states that (House 1990, p. 34) "As the complexity of the signal increases, pitch sensitivity decreases". By this token, tonal perception would be favored in regions of spectral stability, such as during vowels. Following directly from this hypothesis, is the additional hypothesis that only during spectrally stable regions is tonal movement coded as pitch change. During spectrally unstable regions, instead, tonal movement would be coded as pitch levels which "can be then stored in short-term memory" (House 1990, p. 34). To test these hypotheses, the author carries out a series of experiments that appear to support both of them; this leads him to propose a model of "optimal perception of contour features", based on three constraints: (1) tonal movement must occur in a region of spectral stability; (2) tonal movement needs to be

synchronized with vowel onset (so that the initial part of the tonal movement would occur in the first section of the vowel) and (3) vowel duration must be greater than 100 ms. These conditions are viewed by House as "universally" constraining the human auditory system, and are taken to be responsible for some regularities observed in the languages of the world.

However, some years later, House was led to reconsider some of his assumptions in his investigation of movement perception through (C)VC syllables in Thai (House 1996), where the consonants in the stimuli are all sonorants. In this work, listeners had to match a tonal glide falling through a vowel with a glide falling through a VC sequence. The prediction, according to the Spectral Constraint Hypothesis of House (1990), is that the tonal movement falling through the VC boundary would be coded as a tone level and not as movement, because of the spectral discontinuity introduced by the consonant. However, it appears that in certain conditions, a falling tone is perceived as a configuration whose slope matches the slope of a fall spanning a single vowel. House also adds that, in Thai, a late fall in a CVC syllable is not categorized as a H tone but as a falling tone, which is independent evidence for the fact that the tonal movement through the coda is auditorily resolved as a (two-target) configuration and not as a single level target. The author than proposes to modify his model, so that optimal tonal perception would not be simply dependent on the spectral characteristics of the signal, but also on location relative to syllable position.

Therefore, a higher order linguistic construct, such as the syllable, is needed to account for the data. In such a model, tonal movement through codas would be better resolved than tonal movement through onsets. It remains to be seen if such a model can be applied to describe perception of tonal movement in other languages of the world. Therefore, the

universal validity of House's proposal must be subjected to empirical evaluation. Another observation is in order: the languages tested by House (Swedish and Thai) both have tones specified in the lexicon (though Thai has a much more complex tonal system than Swedish). I believe that another good testing ground for the theory would be represented by languages that do not exploit tone for lexical contrast, but that employ it to distinguish pragmatic meaning, such as English and Italian.

Moreover, tonal alignment continua appear to be exploited similarly in different languages in order to contrast pragmatic meaning. Remember that in German the early-middle-late alignment of an f_0 peak relative to the stressed syllable produced a three-way categorization (new, established and emphatic), while in English it might superficially seem that the same continuum induces only a binary category split in American English (Kohler 1990; Pierrehumbert and Steele 1987). Kohler did, however, include what would be H+L* in English, which Pierrehumbert and Steele (1987) did not explicitly test. It is likely therefore that a similar three-way categorization would obtain for English if those materials would be added.

Regarding the influence of higher order linguistic structure on the perception of tone, I will review another example of work suggesting that caution should be exercised when making claims about universality. Verhoeven (1994), for instance, questions the feasibility of extrapolating perceptual generalization to prosodic patterns that are different from the ones used in a specific experimental setting. Tonal perception (and in the specific case, tonal alignment) might not be so immune from a close interaction with language-specific intonational features. In contrast with the studies on alignment perception mentioned in §1.5, Verhoeven did not explicitly investigate target alignment, but alignment of "rises"

and “falls”. Specifically, the author is interested in the an asymmetry between the perception of rises and falls that had already been noticed in the perception literature on Dutch. According to earlier studies, f_0 excursion size is more easily discriminable in rises than in falls, which ’t Hart (1981) attributes to the greater familiarity with rising configurations on the part of Dutch speakers (but this hypothesis is not thoroughly discussed). Verhoeven then created a set of stimuli on the basis of the Dutch hat pattern (H* H*L) in the Dutch utterances *Renaat is ziek* “Renaat is ill” and *Renaat is in Parijs* “Renaat is in Paris”. In these utterances, 100 ms-long rises and falls, with a pitch excursion of five semitones, were progressively shifted forward in 10 ms steps. Both in a pilot study (where presentation order was not included in the design) and in a subsequent study, Verhoeven found an effect of direction of pitch movement (rising or falling). Basically, in an AX discrimination task listeners needed a smaller “gap size” (in ms) between the reference (standard) and the trial stimulus when discriminating two falls than when discriminating two rises. In other words, it seemed that fall alignment could be better discriminated than rise alignment.

But the part of the study that is especially relevant here is the one where Verhoeven challenges his own findings and the whole idea of a basic asymmetry between the perception of rises and falls. A clarification is in order: the Dutch school has always maintained the “Uniformity Assumption”, which basically claims that linguistic context is irrelevant when it comes to perceiving tonal features. But, as Verhoeven points out, this assumption is recognized as being not entirely justifiable even by the same proponents. While explaining the principles on which their “stylization” procedure of tonal contours is based, ’t Hart (1979, 377) states: “the perceptual tolerances which make stylization possible at all, do not seem to be uniform, that is, to our experience, one has to be very precise on some

places, whereas on other places it does not matter very much". Therefore, in order to test the assumption that rises and falls are perceived differently, Verhoeven performs another alignment experiment, but this time without the context of a multi-word utterance, and with no hat pattern superimposed.

In the second experiment, stimuli are made of isolated words, on which an isolated rise or fall is placed. Verhoeven's assumption is that "positional variation in isolated rises and falls is not exploited linguistically in the Dutch intonation system" (Verhoeven 1994, p. 82). The meaning of this statement is not entirely clear, however. It is possible that an isolated fall or an isolated rise cannot by themselves represent a minimal, grammatical intonational phrase in Dutch. However, it seems that an isolated rise can be the "root" (i.e., the only obligatory element of a contour in the Dutch grammar of intonation) of the contour, both in the form of an early rise (labeled "1") or as a late rise (labeled "2"). At any rate, the results of this experiment show no effect of pitch movement direction. For both isolated rises and falls the discrimination threshold is 82 ms. In my opinion, though these results are very difficult to evaluate, they are also very cogent for the point that the author is trying to make. We cannot make broad psychoacoustic generalizations until we test the stimuli in conditions where tonal variables are presented in linguistically comparable contexts. This is not necessarily an easy task.

A different kind of perceptual asymmetry between rises and falls is found by Hermes et al. (1997). Hermes and his colleagues are interested in determining the Accentuation Boundary for rises and falls, that is "the moment before which more than half of the subjects indicated the previous syllable as accented and after which more than half of the subjects indicated the next syllable" (Hermes, Beaugendre, and House 1997, p. 177). The

authors found that the Accentuation Boundary for rises is generally at vowel offset, while for falls it is later, i.e., between the vowel offset of test syllable 1 and the vowel onset of test syllable 2.

Some of the early literature on the perception of question versus statement contours also shows some potentially interesting data for the comparison of f_0 perception between speech and non-speech data. In Studdert-Kennedy and Hadding-Koch (1973), for instance, listeners had to judge stimuli both using the linguistic labels "questions or statements" or using the psychophysical terms of "rise" and "fall". In these stimuli, the f_0 was manipulated at specific points. Among these, both the peak and the "turning-point" (the point where the f_0 starts to rise for the terminal section of the contour) were determinant for the perception of a question in the linguistic task, but not in the psychophysical task. The same result was obtained with two groups of speakers, namely both Swedish and American listeners. Sine wave speech, manipulated in the same way as the speech stimuli, was also both linguistically and psychophysically judged. Both tasks revealed no effect of pitch peak, but a constant effect of the turning-point. The authors interpreted this finding by invoking a "direct linguistic function" of the peak, in the sense that peak perception would not be mediated by low-level auditory processes. The turning-point effect, on the other hand, cannot be attributed to a specialized speech-mode of perception, in that it is present in both the linguistic and non-linguistic judgments and in all stimulus types. In my view, the peak effect could also be interpreted as an indirect effect of pitch range in the linguistic judgments. We know in fact that peaks are instrumental to gauge some important linguistic functions, such as prominence, but it is not clear how it would have affected the stimuli in this study. Finally, since the linguistic judgments of Swedish and American listeners

were very similar (for both groups either a terminal rise or a high peak produced question judgments), the authors are led to conclude that the relationship between high pitch and the linguistic function of questions is universal, which is a view that still persists in certain literature. The findings presented in Chapter 4 will also challenge this universal view.

1.7 Summary

In this chapter, I have offered an overview of the work on association and alignment of tonal targets, with special attention to perception. In §1.2 and §1.3 I have reviewed the common understanding of the notions of phonological association and alignment, and, in particular, I have argued that the notion of association should be phonetically grounded, in particular it should be informed by discoveries at the perceptual level. Then, in §1.4 I have presented some of the relevant works showing the range of variability of phonetic alignment in production, while in §1.5 I have reviewed some of the work suggesting that alignment plays a major role in the perception of intonation contrast. With §1.5.1 I intended to give a concrete background to the experiments presented in Chapter 3, refsec:american and refsec:exp2 by summarizing the major points of a pilot perception study on Neapolitan Italian. Finally, in §1.6 I discussed the findings of some psychoacoustic work on the perception of pitch, with special attention to speech-like stimuli. The main idea to retain from such works is that it appears that tone transitions are not perceived in their entirety. Also, some of this work leaves open the issue of what are the "universals" of tonal perception. I also questioned whether target perception mechanisms would be employed linguistically or whether they are exclusive to psychoacoustic perception.

The remainder of this dissertation will be structured as follows: in Chapter 2, I present the results of a production study that aimed at defining some temporal, melodic and dynamic characteristics of question and (narrow focus) statement pitch accents in Neapolitan Italian. In Chapter 3, I provide details of a large scale perception experiment investigating tonal targets in Neapolitan Italian, and discuss its results. Then, in Chapter 4, two perception experiments performed with American English listeners are presented. The first one mirrors the linguistic task presented in Chapter 3, and shows that, despite phonetic and pragmatic differences within the alignment contrast for L+H* and L*+H in American English and Neapolitan Italian, the listeners were able to employ target alignment as a useful tool for categorizing the same resynthesized stimuli. The second experiment, more psychoacoustic in nature, shows instead that alignment is not consciously manipulated when judging the timing of accent configurations. In Chapter 5, I then discuss the results of an experiment in which the effects of syllable structure and base stimulus on target perception are investigated. I show that the results support the hypothesis of an invariant target alignment in perception, and that syllable structure effects are seemingly factored out. Finally, in Chapter 6, I discuss the implications of the experiments presented. I also offer a perceptual motivation for some observed phonetic alignment patterns and propose a phonological representation of the temporal alignment contrast in question and statement pitch accents of Neapolitan Italian.

CHAPTER 2

TONAL TARGETS IN NEAPOLITAN ITALIAN: TEMPORAL ALIGNMENT, SCALING AND SLOPE

2.1 Intonation structure in Standard and Neapolitan Italian

Before I describe the specific issues explored in the experiment presented in this Chapter, let me review a few relevant notions about Standard and Neapolitan Italian intonation structure. In particular, I will illustrate the main contrast we will be concerned with here, that is the one between (yes/no) questions and (narrow focus) statements. Note, though, that quantitative as well as qualitative data about the intonation of Standard (and regional varieties) of Italian are still scarce, and a consensus transcription system for Italian intonation has just been tentatively proposed (Grice, D'Imperio, Savino, and Avesani in press). Nevertheless, the facts that I will present below are mostly based on solid empiric inspection (D'Imperio 1995; D'Imperio 1997a; D'Imperio 1997b; D'Imperio 1999; D'Imperio in press).

In Italian, as in English, the location of pitch accents is not fixed in a prosodic phrase and the focus structure of an utterance can be entirely signaled by nuclear accent location. For instance, in a Subject-Verb-Object declarative, an accent can be placed on any of those elements, according to which of them is in focus. An alternative way of signaling focus is

to alter the basic word order in the sentence, since in an unmarked declarative sentence the rheme (new information) corresponds with the final position. Therefore, in an utterance such as *Giovanni mangia la mela* "John eats the apple", we can apply narrow focus on *la mela* in two different ways, that is either by placing the nuclear accent on it, as in *Giovanni mangia LA MELA* (where capital letters indicate nuclear accent position), or by displacing it to the left (topicalization), as in *La mela, mangia Giovanni*. This last case, however, will also present a narrow focal accent on *la mela*, which is in a sense redundant to word order (Sobrero 1993).

In Italian, therefore, as in English and Swedish, focus can be entirely signaled by intonational means. The relationship between focus structure and accent structure will be exemplified by drawing on English examples. For instance, the sentence "Ronnie loves Marie" can be the answer to the three possible questions in (1), depending on the position of the nuclear accent in the utterance.

1a. What's up?

1b. Who does Ronnie love?

1c. Who loves Marie?

Utterances that are felicitous in relation to question (1a) will have the focus structure in (2a), while utterances that are felicitous with question (1b) will have the focus structure shown in (2b). Both (2a) and (2b) will have a late nuclear accent placement. On the other hand, a felicitous answer to question (1c) will have the focus structure in (2c) and nuclear accent placement will be early, on "Ronnie" (capital letters indicate nuclear accent position).

2a. [Ronnie loves MARIE]$_{\mathrm{F}}$

2b. Ronnie loves [MARIE]$_{\mathrm{F}}$

2c. [RONNIE]$_{\mathrm{F}}$ loves Marie

In the case exemplified by (2c), no pitch accents will follow the nuclear one. In English, in fact, the nuclear accent is the last and most prominent pitch accent in an intermediate phrase. It is also the accent that is immediately preceding the phrase accent, a tonal event controlling the melodic contour in the region following the last pitch accent and until the boundary tone at the end of the intermediate phrase (Pierrehumbert 1980; Pierrehumbert and Beckman 1988).

The intended focal structure of an utterance can vary not only in the location of the nuclear accented word but also in focus scope. That is, with some specific accent placements, focus can in fact be broad or narrow. Broad focus is defined as "focus on whole constituents or whole sentences not just on individual words" (Ladd 1980), while narrow focus is a general label for cases in which smaller constituents, such as single words, are selected as the focus element of the utterance. An answer to question (1a) above would, for example, have broad focus, while felicitous answers to questions (1b) and (1c) would have narrow focus respectively on "Marie" and "Ronnie". Among some of the first attempts to provide an acoustic characterization of the prosodic mechanisms for signalling focus, Caldognetto and Fava (1972) and, later, Kori and Farnetani (1983), found that focus in Italian is expressed by an f_0 peak. This finding has been confirmed by later research, such as Avesani (1990) and Avesani (1995), in which the pitch accent expressing narrow focus is analyzed as a H*. The status of this pitch accent will be further discussed below.

The first attempt to employ Pierrehumbert's system for describing Italian intonation belongs to Avesani (1990). In this study, the author proposes an inventory of pitch accents and boundary tones for Standard Italian, based on her experience in implementing a text-to-speech system. Standard Italian interrogatives are generally cued by a combination of a L* nuclear accent followed by a rising H- phrase accent (plus a H% boundary tone) or, alternatively by a H+L* followed by a L-H% combination (Avesani 1990; Avesani 1995). Discussion about the necessity for postulating a phrase accent in Italian is offered in Grice et al. (in press). In both cases, the rising part of the contour, which is held to be responsible for the question meaning of the tune, is a property of the terminal part of the contour, as it is in English polar questions. In contrast to that, Southern varieties (at least those that have received attention in the literature so far) encapsulate the question meaning within a specific pitch accent type, that is generally also the nuclear accent of the phrase. This is true in Palermo and Catania Italian (Grice 1991; Grice 1995a), Bari Italian Grice and Savino (1995) and Neapolitan Italian (Caputo 1994; Caputo 1996; D'Imperio 1995; D'Imperio 1997b). This accent is a rising LH tone, where either L or H (or both) are phonetically aligned with the stressed syllable, depending on the specific variety of Italian considered.

In Bari Italian, the nuclear pitch-accent of polar questions is described as a L+H* "which involves a low pitch target just before a high accented syllable" (Grice and Savino 1997, p. 30). The star notation here must be just taken as a convenient way to mark that the H is aligned with the stressed syllable, since no contrast has been found with another LH rise accent in this variety. L+H* is the pitch accent typical of "queries"[1] and some

[1]Grice and her colleagues have extensively relied upon the HCRC Map Task procedure for data gathering, in which categories such as "query" and "check" are employed.

particular type of “checks” (namely those characterized by a very low degree of speaker confidence in his knowledge of the information shared with his interlocutor). Palermo Italian and Neapolitan Italian appear to employ L*+H for the same purposes.

It also appears that the final fall following the L+H* of Bari questions, which is attributed to the presence of a L- phrase accent plus an additional L% boundary tone (when in absolute intonational final position) can be curtailed if the accented syllable is the last in the word and no postaccentual material follows. This appears to be true also of Palermo Italian (Grice 1995a) and Neapolitan Italian (D’Imperio 1995) (as shown in §1.4). Therefore, if we had to make a typological classification similar to the one proposed by Grønnum (1991), who distinguishes languages as “truncating” or “compressing” (when it comes to choosing a strategy for temporal reorganization due to tonal crowding), some Southern varieties of Italian appear to be truncating, in that they prefer to curtail a tone in place of anticipating the preceding one. The situation appears to be more complex in Neapolitan Italian, as mentioned in §1.4 above. In fact, we can have both truncation and compression, depending on the specific situation.

A first proposal for adapting the ToBI framework to describe Neapolitan Italian intonation can be found in Caputo and D’Imperio (1995). In this work, it is proposed, among the other things, that H+L* should be used to describe the nuclear accent of declaratives, where the L tone is scaled at the speakers’ baseline (the bottom of the speaker’s range). A H+L* pitch accent which has similar phonetic characteristics has also been independently described for other romance languages, such as European Portuguese (Frota 1997; Frota 2000a), as well as in some varieties of English (see discussion in Grice (1995b)) and of Italian (Avesani 1990; Grice 1995a).

A proposal regarding the rise-fall of questions in Neapolitan, restated in Caputo (1996), is to transcribe the complex rise-fall movement of very emphatic Neapolitan Italian polar questions as a tritonal accent, since in some instances the entire tonal movement is completed within the accented syllable. Apart from theory-internal type of arguments against such a proposal (such as that tritonal configurations are cross-linguistically marked and that such an analysis has the result of overgenerating configurations that do not exist, by permutation; but see discussion in Grice (1995b)), remember from Chapter 1, §1.4, that long focus questions present a "split" of the rising and the falling sections of the rise-fall (see Figure 1.10). This contrasts with what happens in statements with long focus constituents, as we can see in Figure 2.1. Here the utterances are *Vedrai [mamma] domani* "You will see [mom] tomorrow" with narrow focus on the word *mamma* (upper) and the utterance *Vedrai [la mano di mamma] domani* "You will see [mom's hand] tomorrow", with narrow focus on the longer constituent *mano di mamma* (lower). Note the fall in pitch soon after the word *mamma* in the upper panel.

From the above observation, it was hypothesized that the constituent final fall of interrogatives is analogous to the sentence accent of Swedish (Bruce 1977), in that this tonal event marks the end of the focus constituent and contributes to the perceived prominence of the focal accent, without creating the perceptual impression of a phrasal break. I proposed to analyze such phrase accent as a HL- fall (D'Imperio 1997b; D'Imperio 1998a; D'Imperio in press), since the H can be docked to a stressed syllable in long focus constituents.

Statement long focus constituents are analyzed as having a more complex accent structure, similar to the structure described already for another Southern variety, i.e., Bari Italian

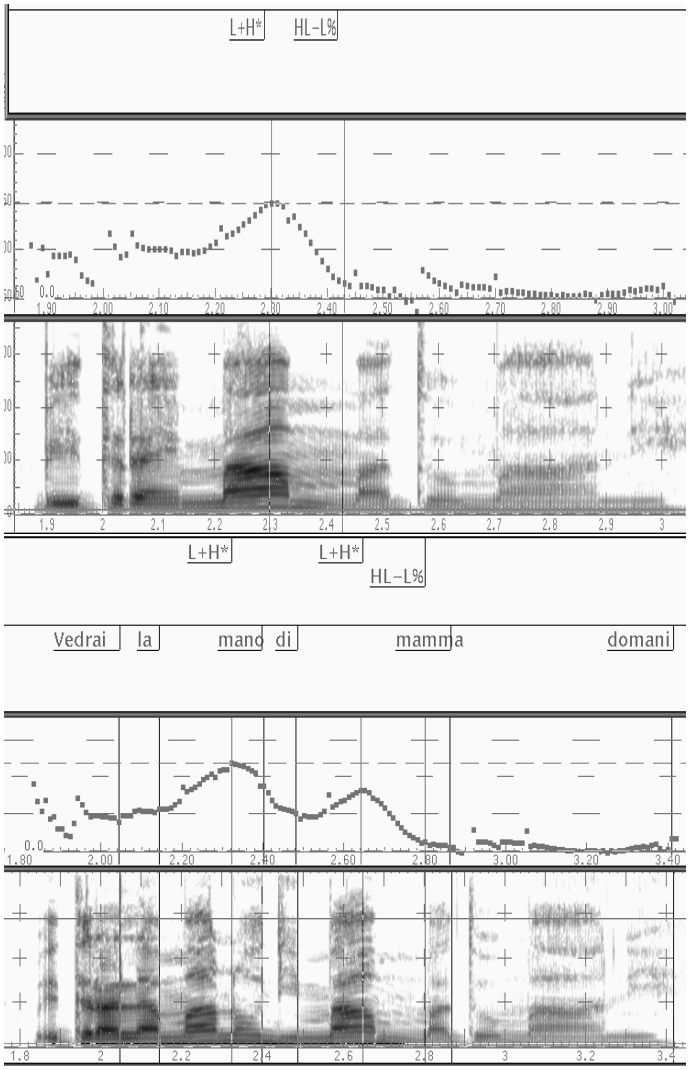

Figure 2.1: Tone labels, f_0 curve and spectrogram for the statement *Vedrai MAMMA domani* uttered with narrow focus on *mamma* (upper) and of the statement *Vedrai LA MANO DI MAMMA domani* uttered with narrow focus on the constituent *la mano di mamma* (lower).

Grice and Savino (1995, Grice and Savino (1997). Specifically, in multi-word narrow focus statements I postulate the presence of two pitch accents, a prenuclear one on the first stressed syllable and a nuclear one on the last stressed syllable, both represented as L+H*. In Bari Italian, it is narrow focus questions, and not statements, that seem to have a similar structure.[2]

When looking at declaratives with long focus constituents, as mentioned above, a fall remains to the immediate right of the constituent-initial rise, producing a medial L target that is not found in interrogative focus constituents. A second fall is later realized in the vicinity of the focus constituent end. At first, I analyzed the narrow focus pitch accent of declaratives as a H*+L (D'Imperio and House 1997), since the fall is completed on the stressed syllable and the peak is very early (close to the stressed syllable onset). A subsequent investigation of the focus constituent medial L target appears to provide evidence for a LH structure at the constituent-final accented syllable (D'Imperio 1999). In fact, it was found that the medial f_0 minimum has always the same value, independent of the length of the segmental material separating the first and second accent in the focus constituent. Regression line fitting showed that by increasing the number of syllables intervening between the two accents, the slope of the contour interpolated between the preceding peak and the medial low becomes shallower. This points to the existence of an actual L target,

[2] In this variety, Grice and Savino (1997) postulate the existence of a postfocal accent in early focus polar questions. Such postfocal accent is placed on a lexical item that is given in the discourse and belongs to the same pitch accent category as the focal accent. This phenomenon has been labeled "reaccentuation" by Cruttenden (1993), in that it has the result of "accenting again" a word previously mentioned in the discourse.

structurally belonging to the pitch event at the following nuclear accent in the narrow focus constituent. Therefore, I discarded the hypothesis of a simple "sagging" interpolation between two subsequent H peaks (Pierrehumbert 1980). Such an analysis of the narrow focal accent for declaratives is similar to the one proposed for the final HL- fall of interrogative focus constituents. In other words, the fall following the nuclear rise in narrow focus declaratives must be analyzed as the same tonal event in declaratives and interrogatives. An alternative analysis would be to consider the fall as part of the nuclear pitch accent, which would necessarily be tritonal (LHL). Such a choice is dispreferred for reasons outlined above. Evidence for a HL- (instead of a simple L-) structure of this phrasal tone comes from the fact that the nuclear pitch accent in the statement focus constituent is not necessarily downstepped, as we might expect for a sequence of bitonal accents. Also, I will invoke Occam's razor, and I will avoid postulating a redundant difference between statements and questions. That is, I will assume that the yes/no question tune differs from the narrow focus statement tune only as to focal pitch accent type, while proposing that the phrase accent type is shared between the two modalities.

Additionally, I postulate that the HL- phrase accent of questions and statements has a secondary association with the last stressed syllable within the focus constituent, and is transcribed as H(*)L- in long focus constituents, (where it does not compete with L*+H for its realization; see Figure 1.10) to underline its surface similarity with a regular pitch accent. This also predicts that the target H of rising pitch accents will compete temporally with the H target of the phrasal HL- in nuclear position. I will postulate that in Neapolitan Italian the nuclear accent is the last full fledged accent in an intermediate phrase. For instance, H(*)L- of question long focus constituents does not constitute a full-fledged pitch

accent, while the second L*+H of statement long focus consituents (see Figure 2.1) does. The analysis of the HL- as a unitary tonal event in both modalities predicts also that the duration of the HL- fall will be constant. This prediction will be tested in the production experiment presented below.

It is also relevant to say that the new analysis of the declarative narrow focus pitch accent as L+H* was instrumental in deciding the starredness status of the L in L*+H of interrogatives. In fact, from alignment facts alone, it is not clear whether the L or the H are associated to the stressed syllable in interrogative focal accents, since both the L and the H target are realized within the boundaries of the stressed syllable. However, since the existence of another LH rise must be recognized for declarative narrow focal accent, I employ a different notation based on surface alignment facts. A phonological analysis of this contrast is proposed in Chapter 6.

A preliminary system for some basic patterns of Neapolitan Italian is proposed here (from D'Imperio (in press) with some modifications).

The pitch accent types identified so far are:

- H*/(!H*): prenuclear (H*) and postnuclear in questions (!H*).
- H+L*: nuclear, broad focus declarative.
- L+H*: nuclear, narrow focus declarative.
- L*+H: nuclear, interrogative (polar questions, broad and narrow focus).

Several earlier investigations (D'Imperio 1995; D'Imperio 1997a; D'Imperio 1997b; D'Imperio and House 1997) have revealed an interesting interaction between focus scope and modality for utterances with late nuclear accent. The relevant tunes are listed in Table 2.1, where the prenuclear accent is in parentheses (since it is not obligatory) and nuclear as well as edge tones follow. The sentences upon which focus scope and modality are manipulated are *Mamma ballava da Lalla* "Mom used to dance at Lalla's" and *Mamma andava a ballare da Lalla* "Mom used to go dancing at Lalla's", in both declarative and interrogative modality.

First, we observe that the pitch accent type for the nuclear accent position in broad focus declaratives (H+L*) is not the same as the one employed in narrow focus declaratives (L+H*). The broad focus accent appears to be "downstepped" and acoustically less prominent than the accent preceding it. This is not the case in narrow focus declaratives, which present a very salient nuclear accent. This is a LH rise, that is followed by a fall to the speaker's baseline, as we can see in Figure 2.1. Here the utterance *Vedrai MAMMA domani* "You will see mom tomorrow" was produced as a statement with narrow focus on the word *mamma* (upper).

Returning to narrow and broad focus declaratives, it appears that the difference in accent structure helps the listener to infer the intended focus structure of the utterance. No perceptual ambiguity was in fact found between broad and late narrow focus declaratives in a perceptual experiment described in D'Imperio (1997a). This contrasts with what has been argued for English, i.e., that accent placement and accent type are ambiguous to the purpose of signaling focus structure (Ladd 1980) in utterances such as (2a) and (2b) above.

Declaratives	**Interrogatives**
Broad focus	Broad focus
(H*) H+L* L-L%	(H*) L*+H HL-L%
Mamma ballava da Lalla.	*Mamma ballava da Lalla?*
“Mom used to dance at Lalla’s”	“Did mom use to go dancing at Lalla’s?”
Late narrow focus	Late narrow focus
(H*) L+H* HL-L%	(H*) L*+H HL- L%
Mamma ballava [da Lalla].	*Mamma ballava [da Lalla]?*
Early narrow focus (short constituent)	Early narrow focus (short constituent)
L+H* HL-L%	L*+H HL- !H* L%
[Mamma] ballava da Lalla.	*[Mamma] ballava da Lalla?*
Early narrow focus (long constituent)	Early narrow focus (long constituent)
(H*) L+H* L+H* HL-L%	L*+H H(*)L- !H* L%
Mamma [andava a ballare] da Lalla.	*Mamma [andava a ballare] da Lalla?*
“Mom used to go dancing at Lalla’s”	“Did mom use to go dancing at Lalla’s?”

Table 2.1: Declarative and Interrogative accent structures in Neapolitan Italian.

Unlike what has been found for declaratives, interrogatives show the ambiguity that is generally found in late accent placement utterances of English. It is no surprise, therefore, that the question tunes of broad focus and late narrow focus utterances possess an identical contour, as we can see in Figure 2.2. Here, a broad focus rendition (upper) and a late narrow focus rendition (lower) of the question *Mamma andava a ballare da Lalla?* "Did mom use to go dancing at Lalla's?" are shown. Notice that the narrow focus rendition has a slightly expanded pitch range within the L*+H accent and the prenuclear H*.

Now, compare the LH rise in the medial panel of Figure 2.2 and the one in the lower panel (L*+H and L+H* respectively), which shows the declarative *Mamma andava a ballare da LALLA*, with narrow focus on *Lalla*. Such rises are quite similar, in that a rise-fall pattern, with a very salient and discernible peak, characterizes both of them. However, the timing characteristics of the peak and the rise are different and perceptually relevant to the purpose of signaling question vs. statement modality, as discussed in §1.5.1 (D'Imperio and House 1997). Specifically, the peak is reached later within the stressed vowel of narrow focus interrogatives, and the perceptibility of the rising movement was argued to be helping the identification of a question (see §1.5.1). Specifically, in Neapolitan Italian this accent is similar to the "scooped" (L*+H) accent of English, but with timing differences for the L and the H targets (see §4.1 for details), and, consequently, of the intervening slope. Note that the timing difference is very subtle in Neapolitan, with the rise to the peak being all contained within the boundaries of the stressed syllable (in English, the H target is generally reached much later, such as in the postonic syllable).

An additional issue involves the description of the tunes listed in the bottom half of Table 2.1. It is here that we discover an important difference in the postfocal contour

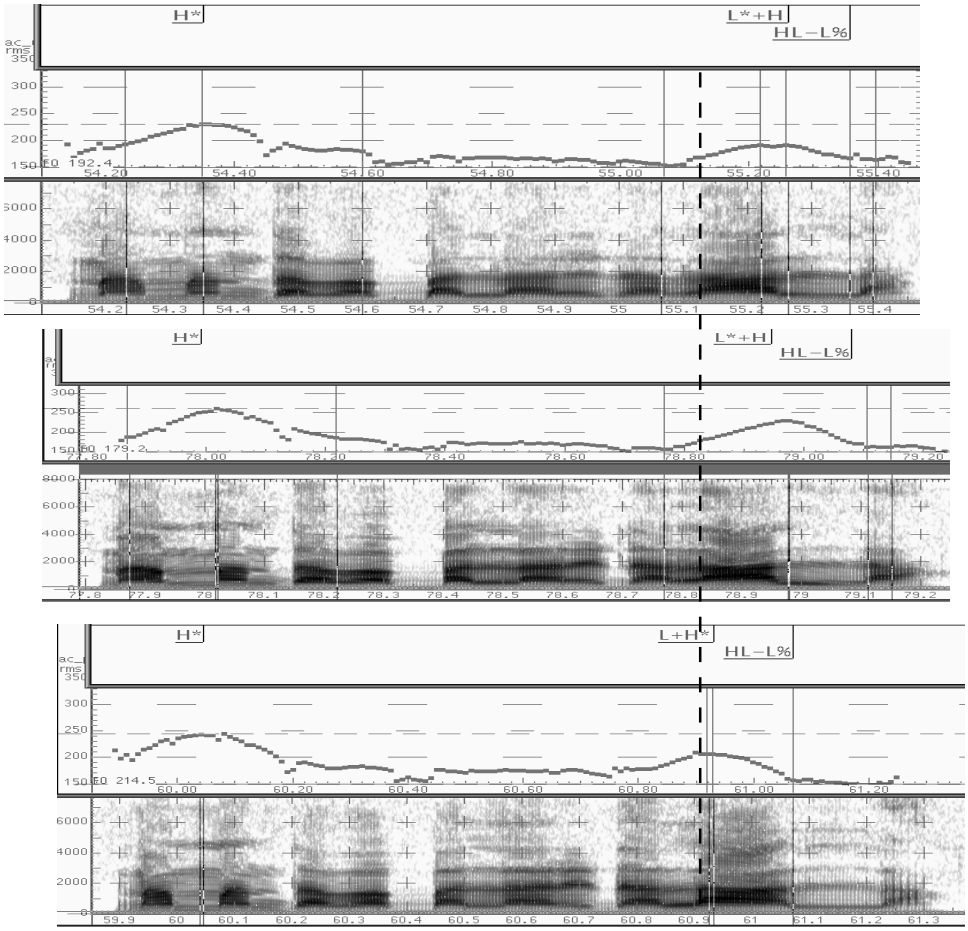

Figure 2.2: Tone labels, f_0 curve and spectrogram for the question *Mamma andava a ballare da Lalla?* uttered with intended broad focus (upper) and narrow focus on *Lalla* (medial) and for the statement *Mamma andava a ballare da LALLA* uttered with narrow focus on *Lalla* (lower). The onset of the stressed vowel is indicated by the dashed line.

of early focus interrogatives as opposed to that of declaratives. Similarly to English and Standard Italian, early focus declaratives do not present any accent after the focal L+H*. Interrogatives, on the other hand, present a small pitch obtrusion, associated to the last stressed syllable of the intonational phrase (!H*), which follows the salient initial rise-fall accent. This tonal excursion cannot be analyzed as a mere boundary tone, since its location is strictly determined by the location of the stressed syllable to which it is associated. It is possible that the small tonal excursion is not the nuclear accent of the phrase, despite being final, and that it is a postfocal accent which is downstepped as it happens in the case of Swedish postfocal accents (Bruce 1977; Bruce 1982). A perception experiment and a production experiment reported in D'Imperio (1997b) tested this hypothesis. The results support the hypothesis that the focal accent is also the nuclear accent, i.e., the most prominent accent of the utterance, while the postfocal accent is merely postnuclear.

Such findings reveal therefore a feature of nuclear accents of Neapolitan Italian early focus questions that makes the language similar to Swedish. In Swedish, in fact, the focal accent is not positionally defined as the last accent in the phrase (unlike English), but it is marked by a separate tonal event in a sequence of accents, i.e., the "sentence accent" (Bruce 1977). Also, while all accents following the nuclear one are generally suppressed in English, they are not suppressed after the focal accent of Swedish. In Neapolitan, I propose that a downstep process applies to all postfocal accents.[3]

In the production experiment presented next the contrast between the nuclear accent of yes/no questions (L*+H) and of narrow focus statements (L+H*) will be explored in terms

[3]Though this is a predictable process, I will transcribe downstepped accents with a preceding "!" for labeling purposes.

of measured targets in the f_0 contour. The section of the contour taken into consideration will actually include the pitch accent plus the fall for the HL- phrase accent. These tones are acoustically manifested as a series of a low, a high and a final low pitch target, which I will conveniently refer to as LHL configuration. That is, though the tonal analysis postulates the existence of two underlying H targets (one for rising pitch accent and the other for the falling phrase accent), just one H target is acoustically realized when the focus constituent is made of a single-word.

2.2 Rationale

In §1 I mentioned that tonal targets can be defined according to two dimensions, i.e., *alignment* (the exact temporal location) and *scaling* (the exact f_0 value). Both dimensions will be investigated in the production study presented here in order to acoustically characterize tonal targets for the two rising accents of Neapolitan Italian.

A number of factors in a variety of languages have been shown to affect the temporal location and the f_0 height of tonal targets, as mentioned in §1.4. As for tonal alignment, it appears that tones are timed to cooccur quite systematically with specific segments in the string. Such regularities appear to be language-specific. Arvaniti et al. (1998) show, for instance, that Greek rising prenuclear accents begin at the onset of the stressed syllable and peak early in the following unstressed vowel, independent of syllable composition and duration.

Analogously, Ladd et al. (1999) show that the beginning of the f_0 rise in English prenuclear accents is consistently aligned with the stressed syllable onset (independent

of speech rate manipulations), while the end of the rise is aligned with a less specifiable location (within the segmental stretch from the end of the stressed syllable to the onset of the following vowel). As mentioned already in §1.4 above, such findings suggest that tonal target alignment is one of the defining properties of "accent identity", while other more holistic features, such as accent shape, duration and slope are not controlled by the speaker, hence are not specified.

Apart from determining a contrast in terms of lexical accent location, alignment can determine pitch accent type (which in turn cue a specific pragmatic contrast; (cf. Ward and Hirschberg (1985, Pierrehumbert and Steele (1989). In Neapolitan Italian, peak alignment appears to be influential in determining whether an utterance is a yes/no question or a narrow focus statement (D'Imperio and House 1997). Question peaks are in fact later than statement peaks, all else being equal. Therefore, the question/statement contrast will be one of the main factors of the study presented here.

Given the analysis of the HL- phrase accent presented above, we also expect that the L2 target for the fall will be realized later in questions than in statements. This is because the HL- starts at the H target location for the preceding rise, given its secondary association with the last stressed syllable of the focus constituent. Hence, since the question H is later, L2 should be later as well.

Moreover, other work shows that segmental phonological contrasts, such as contrastive length, can affect peak alignment. For instance, Ladd et al. (2000) suggests that Dutch prenuclear rising accents are affected by the phonological length of the accented vowel. Basically, the peak is located at the end of the accented vowel when the vowel is phonologically long, while it is located at the end of the following consonant when it is short.

Such a regularity appears to hold despite the fact that phonologically long vowels are not always longer (when duration is measured) than phonologically short vowels. Italian is an example of a slightly different case. In this language, in fact, stressed vowels show a large durational difference between open and closed syllables, especially when in penultimate position within the word (D'Imperio and Rosenthall 1999), though vowel length is not contrastive *per se*. Namely, stressed vowels are long when in an open syllable, while they are short in a closed syllable.

One of the hypotheses tested here is that tonal targets are timed to occur at a specific, absolute location in the utterance (e.g., at a certain temporal distance after vowel or syllable onset). This predicts that peaks will be invariantly produced at such locations, independent of structurally-dependent vowel duration differences. Alternatively, if target alignment is proportional to the entire vowel and/or syllable duration, target location is expected to be different in open vs. closed syllables (analogously to the Dutch situation). Therefore, one of the factors manipulated in the acoustic study presented here is whether the accented vowel is in an open or a closed syllable.

In addition, the segmental composition of the postaccentual syllable will be manipulated. Arvaniti et al. (1998) found, for instance, that prenuclear peaks in Greek are earlier relative to the onset of the stressed syllable when the postaccentual syllable has a nasal onset than when the onset is a stop and, even more so, when it is a fricative. However, in the case of Greek, the peak is timed to occur with the postaccentual vowel nucleus, therefore, given that fricatives are longer than nasals, the later peak is a consequence of a longer segmental stretch preceding the peak (from stressed syllable onset). In the case of Neapolitan nuclear accents, instead, there is no *a priori* reason to expect that the segmental

composition of the postaccentual onset would affect peak alignment. We expect, though, that it might affect the alignment of the target of the falling phrase accent, that is second low target in the rise-fall tonal contour (L2 henceforth), since from previous observations it was noticed that it is generally realized at some location within the postaccentual syllable. In such a case, accentual syllable duration might affect the alignment of such a target if measured relative to the left edge of the accented syllable.

Scaling of f_0 values has been reported to be affected by the question/statement distinction in a variety of languages. This distinction seems also to be relevant for the perceptual identification of questions and statements. For instance, the higher peaks of Hungarian questions have been reported to help listeners to perceptually identify questions versus statements (Gósy and Terken 1994). Therefore, I will also test the hypothesis that Neapolitan questions are characterized, analogously to Hungarian, by higher fundamental frequency peaks. Additionally, from impressionistic observations, it appeared that, apart from the H peak, the two low targets (the L for the LH accent, L1 henceforth, and L2) considered here show different values in questions and statements. Hence, in addition to peak f_0 level, L1 and L2 f_0 values will be analyzed here.

Finally, according to a strict autosegmental view, slope gradient within a rise or a falling f_0 contour is hypothesized to be irrelevant and entirely dependent on the timing and scaling characteristics of the tonal targets involved. Such a hypothesis will be tested here by measuring both LH rise (from L1 to H) and and HL fall (from H to L2) slope.

2.3 Methods

2.3.1 Corpus

The corpus consisted of a group of sentences in which modality (question or statement; QS henceforth), structure of the stressed syllable and segmental environment within the target word were varied. The stressed syllable within the target words could either be closed or open (Open/Closed, henceforth). Closed syllables within the target word were always closed by a nasal. In order to manipulate possible segmental effects induced by the onset of the postaccentual syllable, the nasal could, in turn, be followed by either the same nasal (thus, resulting in a geminate consonant) or by a voiced stop (Nasal/Stop factor, henceforth). Target open syllables could also be followed by either a nasal or a stop as the onset of the following syllable. Target stressed syllables were always penultimate in the target word and the word was embedded as the direct object in a fixed carrier sentence such that focus scope was always narrow over the object.

The above combinations resulted in a corpus of 8 sentences, shown in Table 2.2. In all cases, the test word was a noun, followed by an adverb. The expected stress pattern was a nuclear LH accent on the target word, followed by a fall. Specifically, a L*+H was produced as the nuclear accent of yes/no questions and a L+H* as the nuclear accent of narrow focus statements, as expected.

2.3.2 Experimental Procedure

Two speakers of Neapolitan Italian, one male (LD) and one female (MD, the author) recorded the material. The speakers were both brought up in Naples and spoke Standard

	Open syllable/nasal	**Open syllable/stop**
Q	*Vedrai [il nano] dopo?*	*Vedrai [il mago] dopo?*
	Will you see [the ninth] afterwards?	Will you see [the magician] afterwards?
S	*Vedrai [il nano] dopo.*	*Vedrai [il mago] dopo.*
	You will see [the ninth] afterwards.	You will see [the magician] afterwards.
	Closed syllable/nasal	**Closed syllable/stop**
Q	*Vedrai [il mammo] dopo?*	*Vedrai [il mango] dopo?*
	Will you see [the *mammo*] afterwards?	Will you see [the mango] afterwards?
S	*Vedrai [il mammo] dopo.*	*Vedrai [il mango] dopo.*
	You will see [the *mammo*] afterwards.	You will see [the mango] afterwards.

Table 2.2: Corpus sentences. Square brackets indicate scope of focus. (Note that *mammo* can be defined as a man that takes on the tasks of a mother, i.e., *mamma*, feminine).

Italian with a Neapolitan accent. In order to ensure the statistical validity of the results, given the relatively low number of participants, each of the 8 target utterances were repeated 10 times, for a total of 160 utterances. [4] The recordings were made, for MD, in a double-walled sound-proof booth in the Phonetics Lab of The Ohio State University. LD was recorded in a studio at the University "Federico II" of Naples. Both subjects used a head-mounted microphone recording into into a Marantz tape recorder. The sentences were written on cards, whose order was randomized. The cards were read one time each. The subjects spoke at a normal pace and as naturally as possible. Misread or disfluent utterances were repeated. The recordings were then digitized at 16 bit, 16 kHz using ESPS Waves$^+$ on a SUN Sparc 10 station. Individual sentences were stored in separate files, and f_0 was extracted every ms using the getf0 pitch tracker provided with the ESPS software package. This utility uses an auto-correlation method of f_0 estimation and dynamic programming to adjust for local intensity variations and f_0 trends.

2.3.3 Acoustic measurements

The f_0 tracks were inspected in combination with waveform and spectrogram for each utterance. Both duration and f_0 measurements were performed. Specifically, within the rise-fall configuration, the f_0 value in Hz was measured for points equated with the first L target (L1), with the H target and with the second L target (L2). In most cases, a peak value corresponding to the H was easy to locate, that is the H target was measured at the single highest f_0 point within the accented syllable. This measurement was labeled

[4] Note, though, that the main goal of this study was mainly to obtain a good range of values within a speaker, therefore the number of participants was not a concern.

"F0max" (see Figure 2.3, for instance). In some cases, instead of a well-defined peak, the accented syllable showed a short f_0 plateau (see Figure 2.4). In such cases, the H target was measured to be the point halfway through the plateau, whose boundaries were marked as "pl1" and "pl2". Such points were determined as the first and the last high f_0 points within a series of points with very similar value.[5] We do not know whether this location corresponds to a production or perceptual target (though the perceptual issue will be explored in the following chapter), but it ensured consistency of measurement. Plateaus were common both in question and statement utterances of this corpus, in that 29 plateaus in questions and 35 in statements were globally measured.

[5]Similarity was determined in base of a criterion by which the f_0 value for such points could not differ more than +/- 2 Hz relative the previous point.

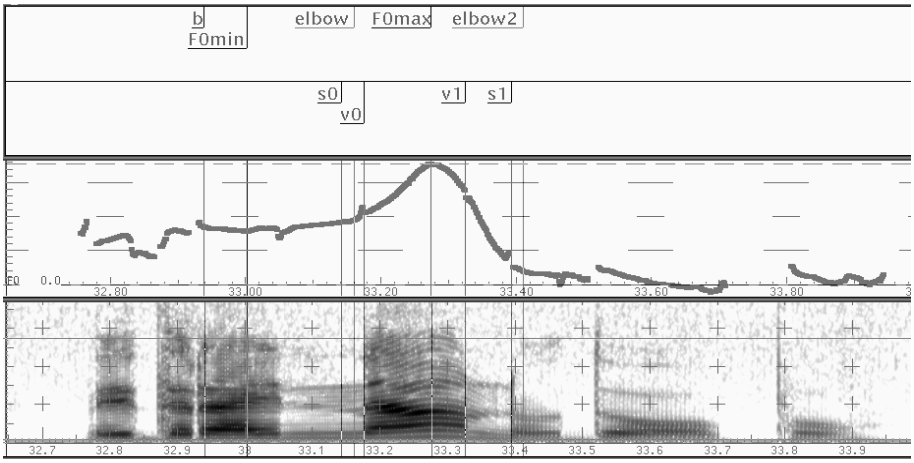

Figure 2.3: Tone labels, f_0 trace and spectrogram for one of the statement utterances of the corpus, showing the duration measurements (s0, v0, etc.) as well as f_0 measurements (F0min, F0max, etc.). The target word here is *nano*.

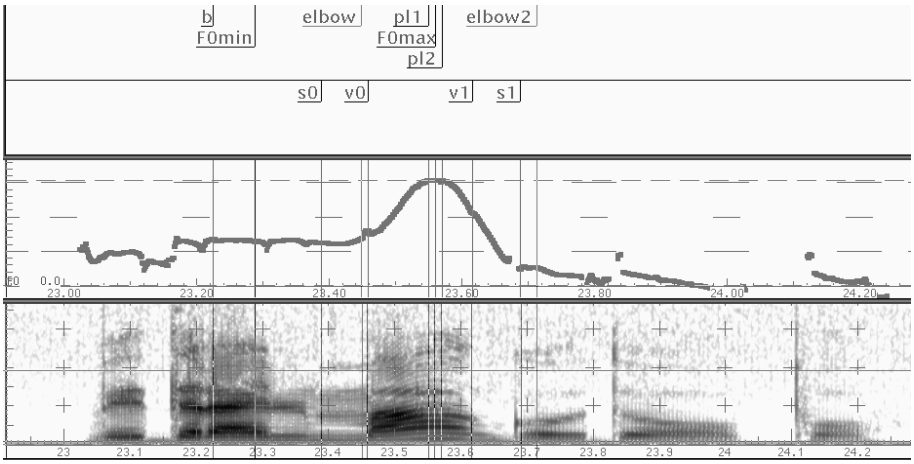

Figure 2.4: Tone labels, f_0 trace and spectrogram for one of the statement utterances of the corpus, showing the duration measurements (s0, v0, etc.) as well as f_0 measurements (F0min, F0max, etc.). The target word here is *nano*.

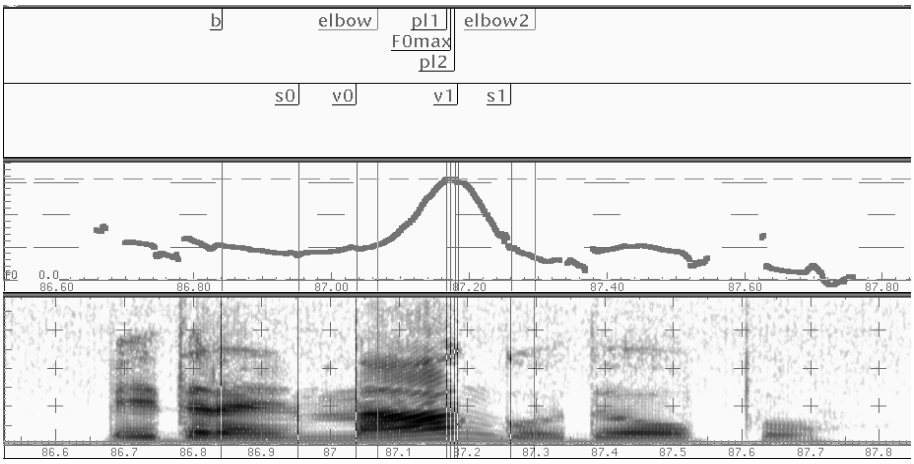

Figure 2.5: Tone labels, f_0 trace and spectrogram for one of the question utterances of the corpus, showing the duration measurements (s0, v0, etc.) as well as f_0 measurements (F0min, F0max, etc.). The target word here is *nano*.

The measurement of L1 and L2 proved to be more challenging. L1 was particularly difficult to discern in statements because of the frequent lack of a visible trough at either accented syllable onset or vowel onset. In other cases, more than one possible trough could be considered as a likely candidate for the L1 location. An example is given in Figure 2.3. As it can be seen, two visible troughs are discernible. The first one precedes a shallower rise, right after a location labeled F0min, which is the f_0 minimum value found automatically within the region between the beginning of the utterance ("b" henceforth)[6] and a point 100 ms before syllable onset). The second one precedes the actual accent rise (a little earlier than the vowel onset). The assumption here is that the actual low target (L1)

[6] "b" was always taken at the onset of the first stressed vowel within the utterance.

is the point at which the f_0 starts to rise rapidly. By this token, L2 should be positioned at the location closer to the stressed vowel, where the f_0 curve bends upward, which will be referred to as the "elbow". It is particularly challenging, though, to perform such a measurement in an objective and consistent way. Therefore, in order to overcome this problem, an automatic procedure[7] was employed by which two straight lines were fitted to the f_0 segment going from "b" to F0max (see Figure 2.6). The parameters of the two linear models were estimated by means of conventional linear least-squares methods. To estimate the elbow position, i.e., the intersection of the two fitted lines, two linear regressions were computed for each possible elbow location (from 1 to *n*, where *n* is the number of samples within the f_0 segment). The location eventually selected as the "elbow" was the one leading to the smallest total modeling error. The elbow was then automatically inserted in the "label file" of a specific utterance, at the location corresponding to the x-intersection of the fitted lines.

In other cases, a less ambiguous low target before the rise was discernible, as in Figure 2.4. However, even in those cases, the location of the elbow was calculated automatically, for consistency sake. The same procedure was also employed in order to locate the end of the fall from the measured peak (L2), as depicted in Figure 2.7. This location was labeled "elbow2" (see Figures 2.3 and 2.4).

Duration measurements were made from waveforms in combination with spectrograms. Standard criteria of segmentation were followed for such measurements (Peterson and

[7] A program originally written by Mary Beckman and slightly modified by the author was employed. A similar procedure was employed already in Pierrehumbert and Beckman (1988), D'Imperio (1995) and D'Imperio (1996).

Lehiste 1960). The locations marked were accented syllable onset (s0), vowel onset (v0), the postaccentual vowel onset (s1) and stressed vowel offset (v1). Such marks are shown in Figure 2.3, 2.4 and 2.5. Boundaries between laterals and nasal (e.g., i*l n*ano) were marked at locations where sudden changes in amplitude and formant structure occurred. Vowel onset was marked at the start of high amplitude periodicity. Since one of the target words (*mammo*) contained a geminate, the problem arose of where to place the syllable boundary. Because of the lack of an uncontroversial criterion, apart from arbitrarily cutting at the geminate halfpoint, it was decided that s1 would be placed at the onset of voicing of the following vowel (within the postaccentual syllable), after the nasal burst. For consistency sake, s1 was always placed at this location also in the other target words. In other words, the stressed syllable could be made, in the simplest case, of a CVC sequence (i.e., *nan*o and *mag*o), a CVCC sequence where the postvocalic CC sequence was a geminate (e.g., *mamm*o) or a CVCC sequence where the postvocalic CC was a cluster (i.e., *mang*o). Figures 2.3, 2.4 and 2.5 illustrate how the various labels were placed, for both statements and questions.

Duration and latency measurements included (i) stressed vowel and syllable duration; (ii) the distance of elbow1 (L1) from both v0 and s0; (iii) the distance of F0max (H) from v0 and s0, as well as from v1 and (iv) the distance of the elbow2 (L2) from v1, s1 and F0max. The measurements are schematized in Figure 2.8.

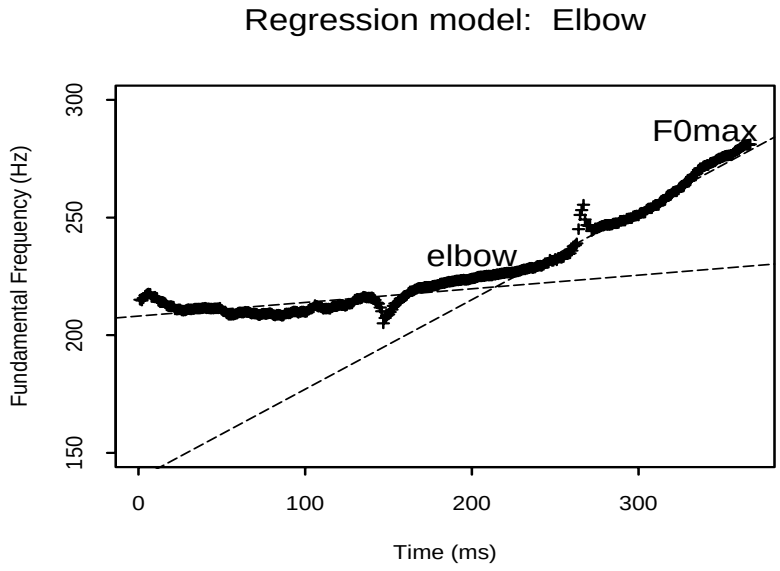

Figure 2.6: f_0 trace with fitted lines that intersect at the elbow time location.

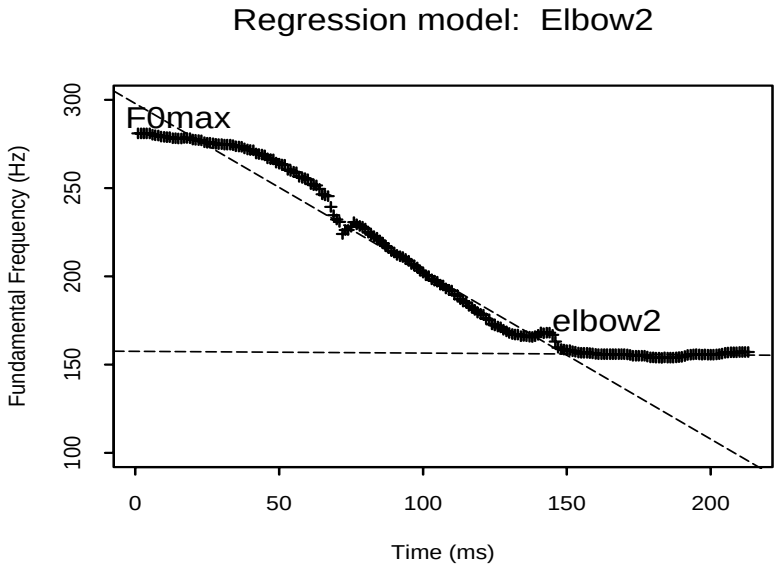

Figure 2.7: f_0 trace with fitted lines that intersect at the elbow2 time location.

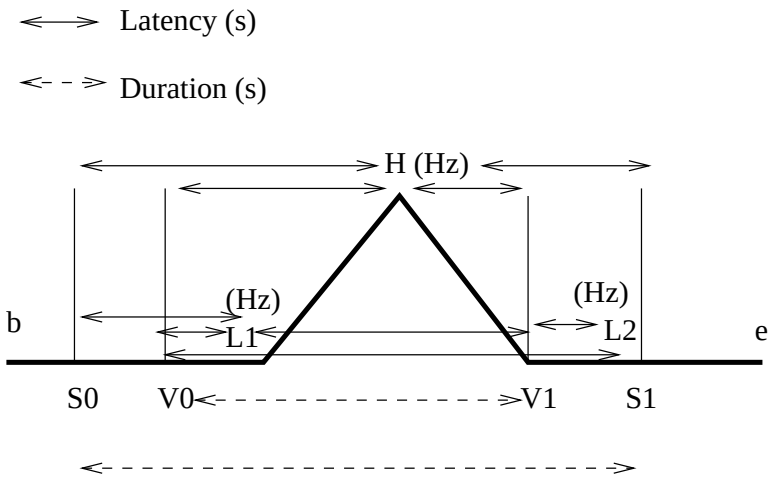

Figure 2.8: Simplified measurement scheme.

2.4 Results

2.4.1 Vowel and syllable duration

The duration results showed that the vowels associated with a L+H* accent (i.e., the accented vowels in the statements) were longer than those associated with a L*+H accent (i.e., the accented vowels in the questions). This was particularly true for MD. Also, as expected, vowels were longer in open than in closed syllables, as we can see in Figure 2.9. Among the vowels in closed syllables, there was a trend for those followed by a postaccentual stop onset to be longer than those followed by a geminate nasal, but this was significant only for LD. Finally, the intervocalic interval ("Sdur" henceforth) was obviously longer in closed syllables. All data were tested in terms of the various main effects and interactions. A series of overall four-way Analyses of Variance (ANOVA), with QS, Open/Closed, Nasal/Stop and Speaker as between-items factors, were run on the results. Additionally, three-way ANOVAs were run on the data for each individual speaker, with QS, Nasal/Stop and Open/Closed as main factors. The cutoff point for significance was 0.01.

The mean duration of the accented vowel for LD was 158 ms for questions and 164 ms for statements, while for MD it was 127 ms for questions and 138 ms for statements. Mean stressed syllable duration for LD was 358 ms for questions and 372 ms for statements; for MD it was 302 ms for statements and 263.5 ms for questions, confirming an overall difference between the accented syllable of questions and statements. The overall analysis on vowel duration revealed in fact a main effect of QS [$F(1,144) = 33.3; p = 0$] as well as an expected effect of Open/Closed [$F(1,144) = 363.7; p = 0$]. Speaker was also significant

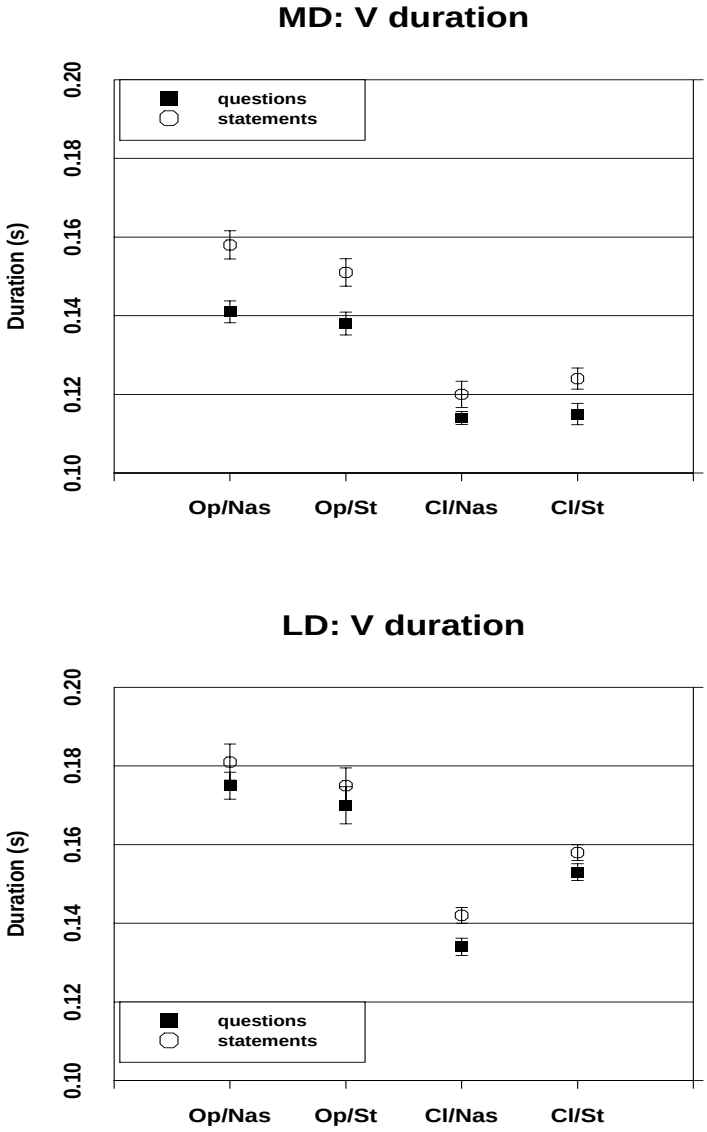

Figure 2.9: Mean vowel durations in questions and statements for MD and LD (Op/Nas = open syll., nasal; Op/St = open syll., stop; Cl/Nas = closed syll., nasal; Cl/St = closed syll., stop). Standard error is indicated by vertical bars.

[$F(1,144) = 356; p = 0$]. Note in the lower panel of Figure 2.9 that, especially for LD, closed vowels tend to be longer when followed by a stop. The interaction between the segmental factor (Nasal/Stop) and Speaker was however significant [$F(1,144) = 8; p < .01$]. Three-way ANOVAs on each speaker revealed in fact that Nasal/Stop was significant only for LD [$F(1,72) = 7.29; p < .01$] while it was not for MD [$F(1,72) = 1.27; p = .3$].

2.4.2 Temporal Alignment

In this section, I shall report on data regarding the alignment of the three major tonal targets, that is L1 (elbow), H (F0max) and L2 (elbow2), relative to various segmental landmarks: accented vowel onset (v0), vowel offset (v1), accented syllable onset (s0) and postaccentual vowel onset (s1). Means for each speaker for L1, H and L2 latency relative to v0, for both questions and statements, are reported in Figure 2.10. Apart from the latency measures, mean f_0 values at each target are shown in the Figure. These will however be discussed in §2.4.3. At a first approximation, note that all three targets are realized earlier in statements than in questions. A detailed analysis of the latency results will follow.

For MD, whose rise-fall configuration was globally later than for LD, L1 was located just before the stressed vowel onset in statements and at or soon after the same location for questions. For LD, L1 was located quite before the stressed vowel onset for statements, while it was located immediately before it in questions. The H peak tended to be timed with the offset of the stressed vowel for MD questions, while it was timed 36 ms before vowel offset in LD. In statements, H was timed to occur around the 2/3 of the vowel for MD, and closer to vowel onset for LD. Basically, question peaks were aligned 37 ms later than statements for MD and 81 ms later for LD. Finally, for MD, L2 tended to occur at

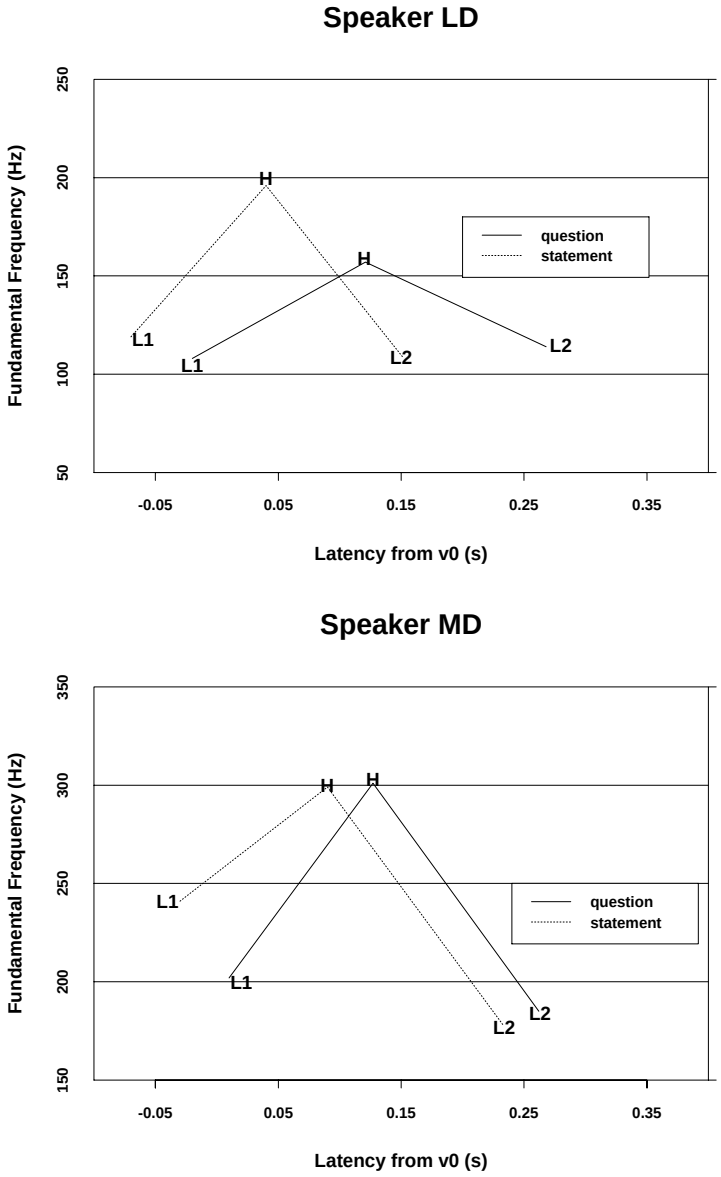

Figure 2.10: f_0 values and latency from v0 (vowel onset) for L1, H and L2 for speaker LD (upper) and MD (lower).

the postaccentual vowel onset for statements, and around the middle of the same vowel for questions. LD differentiated L2 target alignment more drastically between the two modalities. Namely, in statements, L2 was aligned with stressed vowel offset, while in questions it was timed with the onset of the postaccentual vowel.

A more detailed analysis of the results, as well as of the statistics, follows. First, let us look at the results for the elbow latency relative to vowel onset. As shown in Figure 2.10, L1, measured in term of the elbow location, was earlier in statements (-0.07 s for LD and -0.03 s for MD) than in questions (-0.02 s for LD and 0.01 s for MD). In other words, the elbow was circa 50 ms earlier in statements for LD, and circa 40 ms earlier for MD. The results for L1 alignment showed in fact that that the QS factor affected significantly the means when the latency was measured relative to vowel onset (EltoVons in the graph) [$F(1,144) = 68.8; p = 0$]. Also, the Nasal/Stop factor reached significance [$F(1,144) = 7.45; p < 0.01$] as well as Speaker [$F(1,144) = 55.6; p = 0$] while Open/Closed [$F(1,144) = 0.4; p = 0.84$] did not. None of the interactions was significant.

Figure 2.11 shows means and standard error for the EltoVons latency measure for both modalities, separately for each speaker. The EltoVons measure was selected among all latency measures for the elbow since it appears to be less variant (see discussion on peak latency measure below). From the Figure, we can see that the elbow was indeed later for questions than for statements, for both speakers (though LD alignment was earlier overall). Also, elbow latency was even later when the target syllable was followed by a stop, but this was true only for MD statements. However, when measured relative to syllable onset (EltoSons), Nasal/Stop was again significant [$F(1,144) = 14.6; p < 0.01$]. The significance

was confirmed for both speakers in individual one-factor ANOVAs. The interaction between QS, Nasal/Stop and Speaker was not significant, though very close to the rejection threshold [$F(1,144) = 6; p = 0.02$]. In fact, when looking at LD results in the lower panel of Figure 2.11, one can notice that statements are less affected by Nasal/Stop than questions, as far as such a measure is concerned. But the opposite was true for MD. In any case, the result of a long-distance effect of syllable structure on tonal alignment, though partial and seemingly inconsistent among speakers, was novel.

When measured relative to vowel offset (EltoVoff), elbow alignment values were expectedly smaller (meaning that the target is *closer* to vowel offset) in closed syllables than in open syllables. Vowels are in fact shorter in closed syllables than in open syllables in Italian, which was confirmed by the results presented in §2.4.1 (see Figure 2.9). Hence, the Open/Closed factor was significant for this measure [$F(1,144) = 24.6; p < 0.01$], while the interaction Open/Closed by Speaker was not [$F(1,144) = 1.2; p = 0.3$]. The Nasal/Stop effect was not significant, though quite close to the rejection threshold [$F(1,144) = 4.7; p = 0.03$]. In fact, three-way ANOVAs on each speaker's data revealed that the factor was only significant for MD [$F(1,72) = 7.7; p < 0.01$] but not for LD [$F(1,72) = 0.16; p = 0.7$]. QS was significant as well, in that question elbows were closer to vowel offset than statement elbows [$F(1,144) = 105; p = 0$].

Analogous results were found regarding the latency of H (F0max) relative to vowel onset (HtoVons). These results are shown in Figure 2.12. Specifically, the latency was smaller for statements (0.04 s for LD and 0.09 s for MD) than for questions (0.121 s for LD and 0.127 s for MD). Here, QS was again significant [$F(1,144) = 490; p = 0$], while

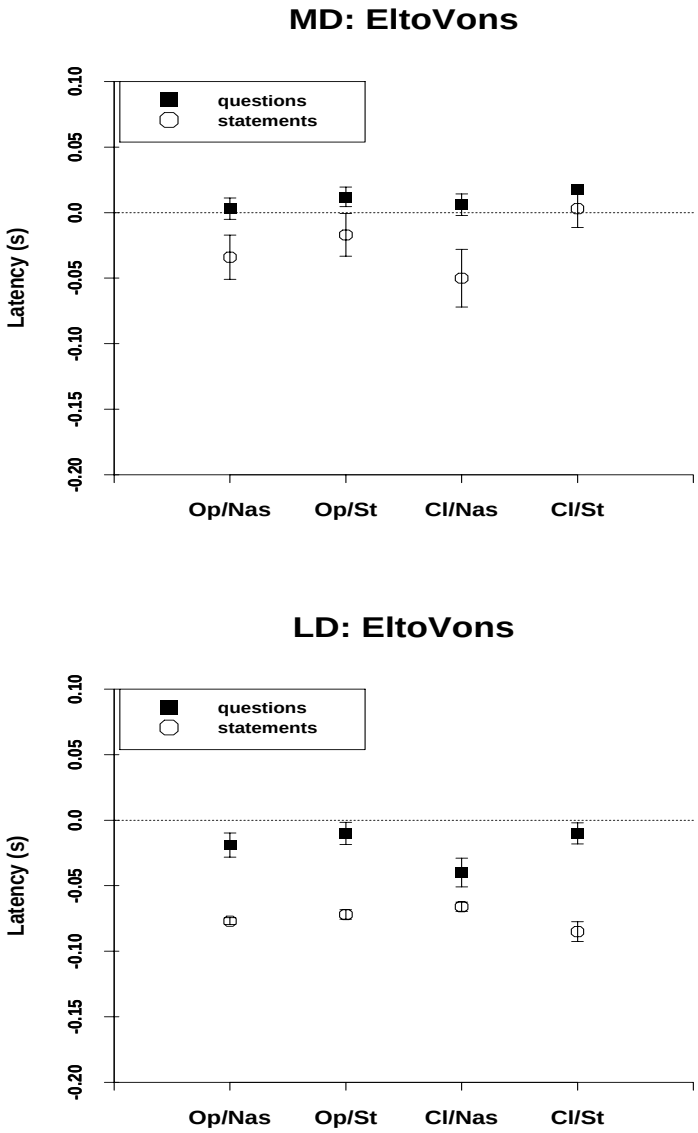

Figure 2.11: Mean L1 (elbow) latency from vowel onset (EltoVons) for MD (upper) and LD (lower) (Op/Nas = open syll., nasal; Op/St = open syll., stop; Cl/Nas = closed syll., nasal; Cl/St = closed syll., stop). The dotted line indicates vowel onset. Standard error is indicated by vertical bars.

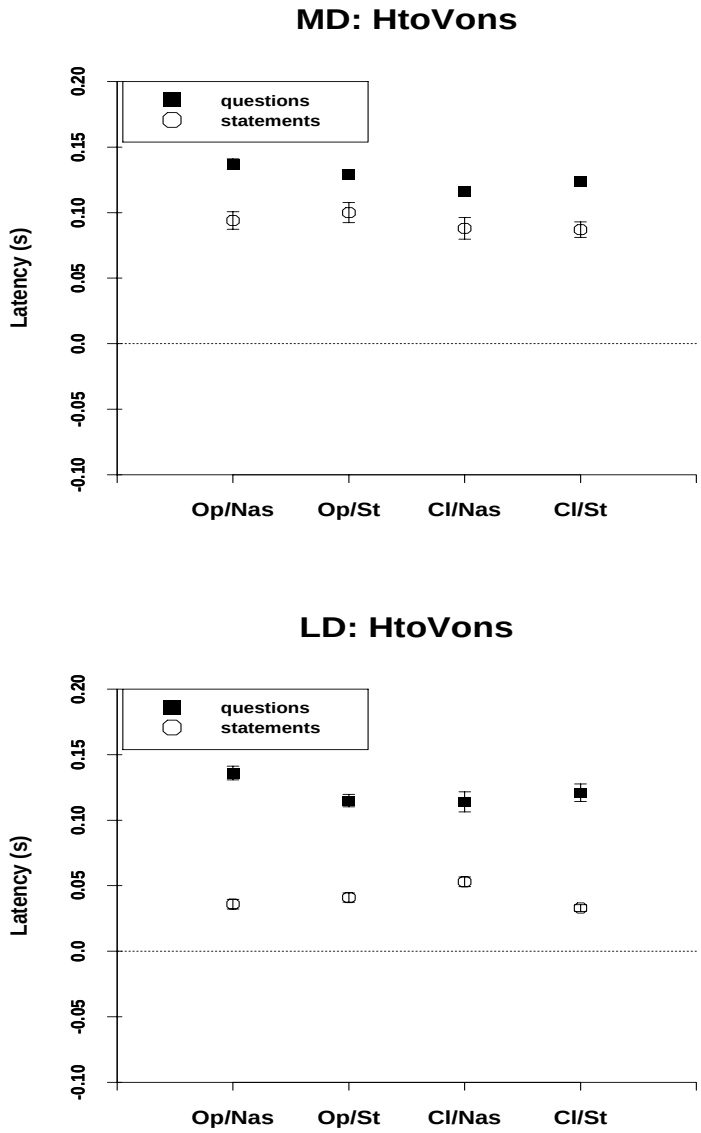

Figure 2.12: Mean H (F0max) latency from vowel onset (HtoVons) for MD (upper) and LD (lower) (Op/Nas = open syll., nasal; Op/St = open syll., stop; Cl/Nas = closed syll., nasal; Cl/St = closed syll., stop). The dotted line indicates vowel onset. Standard error is indicated by vertical bars.

Open/Closed was not [$F(1,144) = 5.9; p = 0.02$], though very close to the rejection threshold. This means that, despite the duration difference between stressed vowels in open and closed syllables, peak alignment relative to vowel onset is stable and therefore not proportional to overall vowel duration. Nasal/Stop was also not significant [$F(1,144) = .8; p = 0.4$]. The Speaker effect was again significant [$F(1,144) = 122; p = 0$].

When measured relative to syllable onset (HtoSons), only QS [$F(1,144)$ = 295; $p = 0$] and Speaker [$F(1,144) = 33.5; p = 0$] were significant, while neither Open/Closed [$F(1,144) = 0.74; p = 0.4$] nor Nasal/Stop [$F(1,144) = 2.6; p = 0.1$] were. The only significant interaction was QS by Open/Closed by Nasal/Stop [$F(1,144) = 7.5; p < 0.01$].

Therefore, it appears that the only strong effect for H alignment relative to vowel onset is the QS factor. In fact, if we look at Figure 2.13 (in which the various peak latency measures are compared for MD) we notice that the HtoVons (as well as the HtoSons) measure is pretty constant throughout the conditions for either questions or statements, while the other alignment measures do vary in a more dramatic way. If the peak is timed to occur relative to vowel or syllable onset, we do not expect any effect of syllable structure, which is what appears to emerge from the results. On the other hand, we expect those effects to reveal themselves in measures that are relative to the right edge of the stressed syllable, such as the alignment relative to vowel offset (HtoVoff) and to syllable offset (HtoSoff). In other words, in closed syllables the peak would have to be located closer to vowel offset, since vowels are short. Alignment should then result to be constant only relative to the left edge of the target syllable.

Such a prediction was supported by the data. When H latency was calculated relative to vowel offset (HtoVoff), apart from the expected effect of QS [$F(1,144) = 754.2; p = 0$],

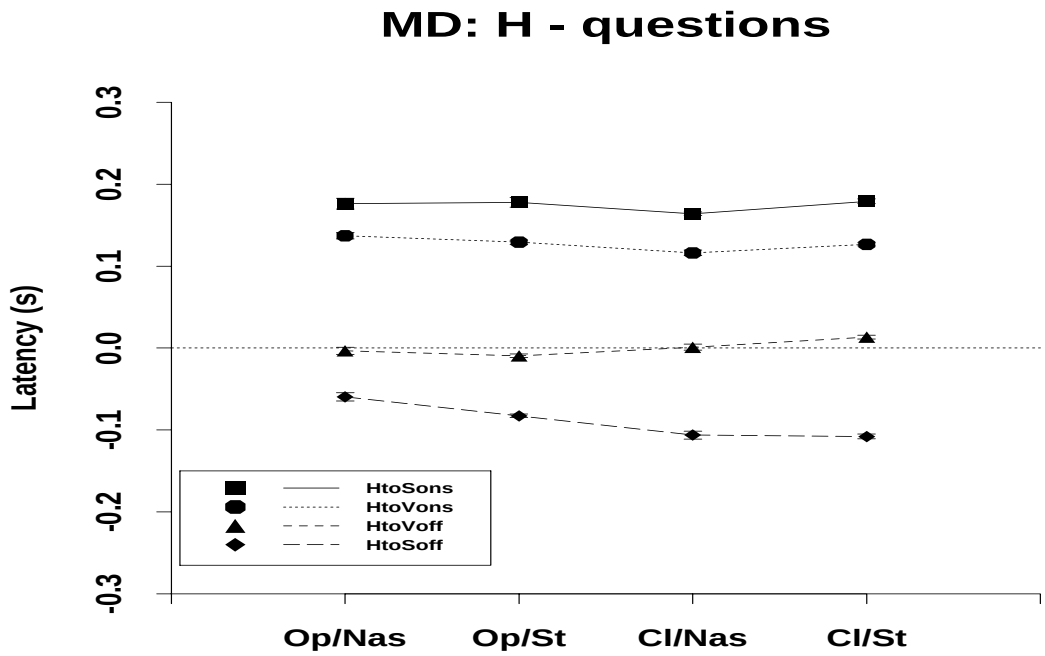

Figure 2.13: Mean H (F0max) latency from vowel onset (HtoVons), vowel offset (HtoVoff) and syllable offset (HtoSoff) for MD (Op/Nas = open syll., nasal; Op/St = open syll., stop; Cl/Nas = closed syll., nasal; Cl/St = closed syll., stop). The dotted line is the reference point for each latency measurement. Standard error is indicated by vertical bars.

Open/Closed was highly significant [$F(1,144) = 84.03; p = 0$]. Moreover, there was no Open/Closed by Speaker interaction [$F(1,144) = 2.5; p = 0.12$] (analogously to what was found for HtoVons).

As one can notice in Figure 2.13, peak values were later (closer to vowel offset) in questions than in statements, as expected. H peaks appear to be even later when the stressed syllable is closed. For instance, peaks are found above or slightly beyond the vowel offset mark (the zero reference value) for closed/nasal and closed/stop syllable of MD questions. The effect of syllable structure on HtoVoff is exemplified in Figure 2.14. The utterance depicted here is a repetition of the question *Vedrai il mango domani?* produced by MD, where the syllable is closed. In this case, as observed in Figure 2.13, the peak (at the L*+H label) visibly occurs within the nasal coda, right after the v1 mark. In sum, in closed syllables, the peak tended to be within the coda for speaker MD, and towards vowel offset for speaker LD (who generally presented earlier latency values for the entire LHL configuration).

For both speakers, the latency of L2 (elbow2) from vowel onset was greater for questions (0.268 s for LD and 0.262 for MD) than for statements (0.151 s for LD and 0.233 s for MD). Indeed, the latency measure was always greater in questions, independent of the reference location chosen (vowel offset, syllable offset or syllable onset), therefore only results relative to vowel onset will be shown (see Figure 2.15). It is interesting to note that mean latency values for L2 were very similar for questions in both speakers, but not for statements. This point will be discussed in §2.5 below. The results of the four-way ANOVA revealed a significant effect of QS [$F(1,144) = 607; p = 0$], Nasal/Stop [$F(1,144) = 17.04; p < 0.01$] and Speaker [$F(1,144) = 160.7; p = 0$], but no Open/Closed

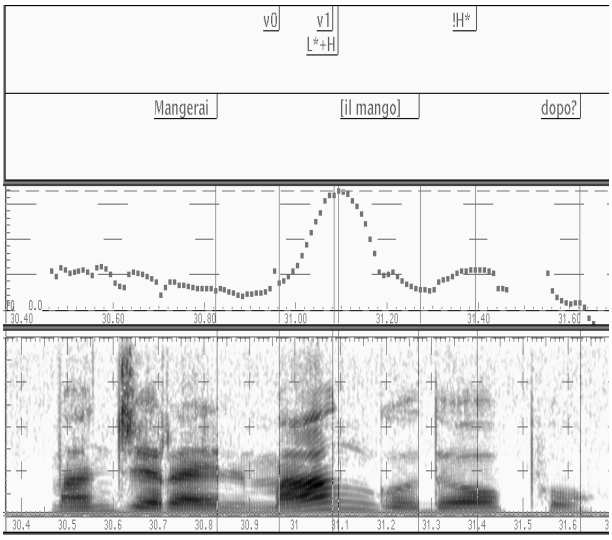

Figure 2.14: f_0 trace for the question *Vedrai il mango dopo?*, showing stressed vowel boundaries (v0 and v1).

effect [$F(1,144) = 0.84; p = 0.36$]. However, Nasal/Stop by Speaker [$F(1,144) = 16; p < 0.01$] was significant, This interaction is due to the fact that Nasal/Stop was significant only for MD [$F(1,72) = 26.07; p < .0001$] and not for LD [$F(1,72) = .01; p < .9$].

When measured relative to vowel offset, Open/Closed had an effect [$F(1,144) = 106; p = 0$] (as well as Speaker [$F(1,144) = 474; p = 0$], while Nasal/Stop was not significant [$F(1,144) = 10.9; p < 0.01$]).

As mentioned in §2.1, the HL- fall has been recently analyzed as a phrasal event (D'Imperio 1997b) for questions and, more tentatively, for statements (D'Imperio 1999). An index of the unitary nature of such HL event might be a constant latency of L2 relative to the preceding peak (El2toF0max). If this same relationship holds for both statements and

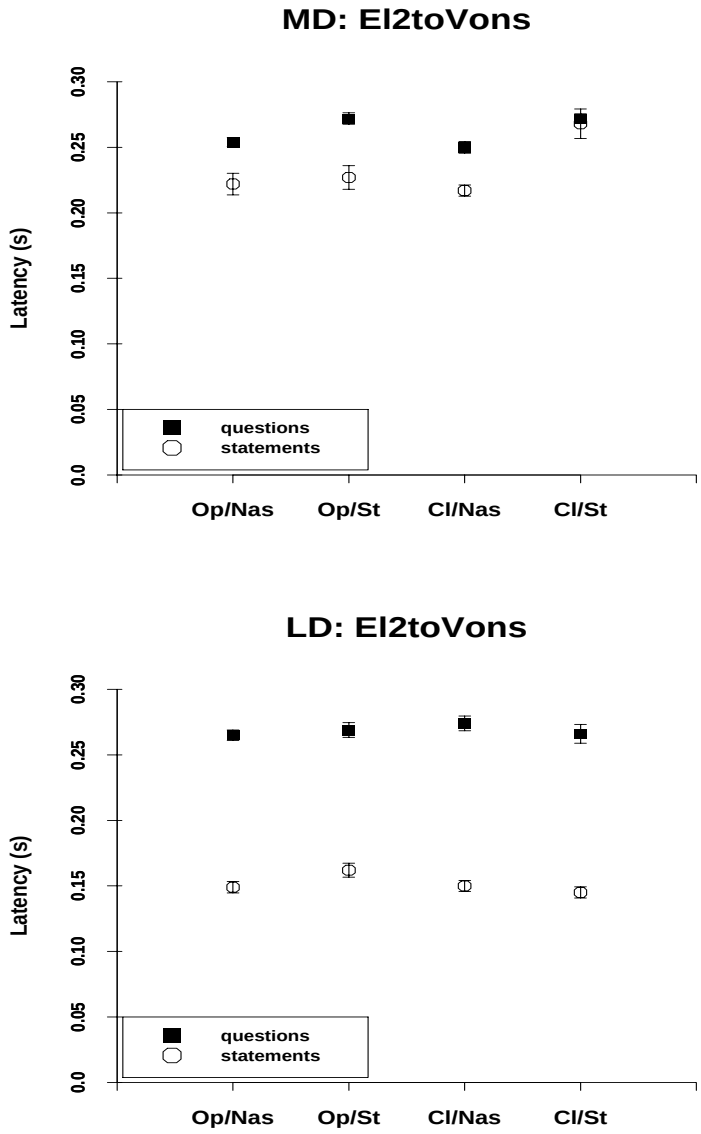

Figure 2.15: Mean L2 (elbow2) latency from vowel offset (El2toVons) for MD (upper) and LD (lower) (Op/Nas = open syll., nasal; Op/St = open syll., stop; Cl/Nas = closed syll., nasal; Cl/St = closed syll., stop). Standard error is indicated by vertical bars.

MD	**Questions (s)**	**Statements (s)**
Open/Nasal	0.116	0.126
Closed/Nasal	0.134	0.129
Open/Stop	0.143	0.125
Closed/Stop	0.145	0.181
LD	**Questions (s)**	**Statements (s)**
Open/Nasal	0.129	0.113
Closed/Nasal	0.160	0.098
Open/Stop	0.154	0.122
Closed/Stop	0.145	0.109

Table 2.3: Mean El2toF0max latency values in seconds for speaker MD and LD.

questions, this would be additional evidence that both modalities share the same phrasal event. The results for the El2toF0max latency are shown in Table 2.3. As one can notice, the hypothesis that the HL fall would show the same duration in statements and questions was verified only for speaker MD, but not for speaker LD. QS was in fact significant only for LD [$F(1,72) = 86.1$; $p = 0$], but not for MD [$F(1,72) = 1.31$; $p = 0.26$].

Therefore, while it appears that the distance between H and L2 is constant independent of modality for speaker MD, this is not the case for LD. Hence, the nature of L2 in questions and statements is not clear from the latency measures alone. The f_0 measurements presented in the next section will help in shedding more light over this issue.

2.4.3 Fundamental frequency target values

Mean values for both speakers and for all major tonal targets in the test utterances are presented in Figures 2.16 and 2.17. As it was already shown in Figure 2.10, elbow f_0 values were lower for questions (108 Hz for LD and 202 Hz for MD) than for statements (119 Hz for LD and 241 Hz for MD). The ANOVA run on L1 f_0 values for both speakers uncovered in fact a highly significant effect of QS [$F(1,144) = 154$; $p = 0$], and Speaker [$F(1,144) = 2893$; $p = 0$], while Open/Closed did not reach significance [$F(1,144) = 0.43$; $p = .51$]. The Nasal/Stop effect did not reach significance [$F(1,144) = 5.315$; $p = 0.02$], though the value was close to the cutoff point. However, when separate 3-way ANOVAs were run on each speaker, it was found that this effect was significant only for LD [$p < 0.01$], but not for MD [$p = 0.27$].

As mentioned in §2.2, given the fact that questions tend to have higher global and local (especially within the peak) f_0 values in various languages of the world, I tested the hypothesis that H values would be higher for questions than for statements. The results showed that the f_0 peak (H) was apparently not affected by QS for speaker MD, but it was for LD. This speaker actually showed the reverse of the expected effect, that is his statements presented higher peaks than questions. However, when listening to LD utterances, it is clear that he put more emphasis on the statements than on the questions, though he was not instructed to do so. This might be due to the fact that narrow focus statements are harder to produce than narrow focus questions: there is in fact no difference between the pitch accent of narrow focus and broad focus questions, while the opposite is true for narrow focus and broad focus statements. Narrow focus statements are, then, "marked" and

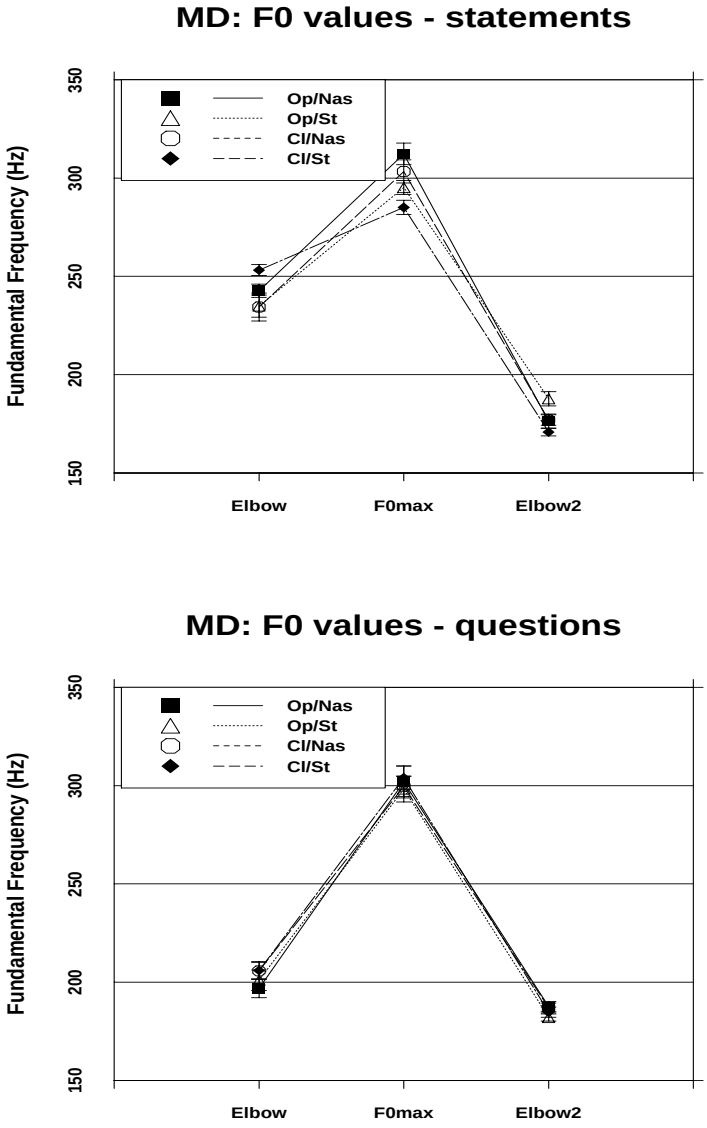

Figure 2.16: Mean fundamental frequency values for elbow1, F0max and elbow2 for MD (Op/Nas = open syll., nasal; Op/St = open syll., stop; Cl/Nas = closed syll., nasal; Cl/St = closed syll., stop). Standard error is indicated by vertical bars.

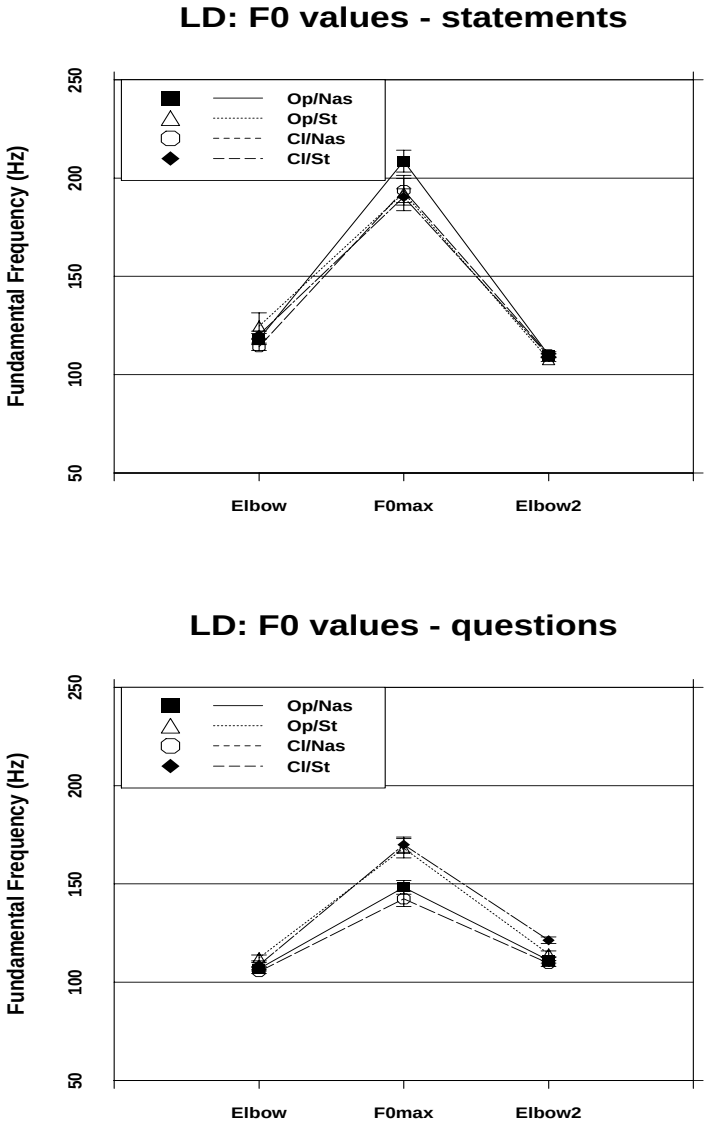

Figure 2.17: Mean fundamental frequency values for elbow1, F0max and elbow2 for LD (Op/Nas = open syll., nasal; Op/St = open syll., stop; Cl/Nas = closed syll., nasal; Cl/St = closed syll., stop). Standard error is indicated by vertical bars.

intrinsically emphatic in this variety. It is possible, therefore, that LD tried hard to signal the narrow focus contrast on statements, putting more effort than in the question production. The heightened effort might have translated in greater emphasis, and, consequently, heightened f_0 values. I believe that the gradient raising of the peak f_0 value is therefore optional, showing strong inter- and intra-speaker variability, in a way that is similar to the use of the "emphatic stress" discussed by Ladd (1996, pp. 199-201).

Hence, it is plausible that the difference, which was only significant for LD [$F(1,72) = 105.28$; $p = 0$], can be merely attributed to individual differences in signaling narrow focus. In other words, there might be just no systematic difference in peak height between two modalities, as speaker MD shows [$F(1,72) = 0.24$; $p < 0.63$]. In the ANOVA conducted on H f_0 values for both speakers, Open/Closed did not reach significance [$F(1,144) = 2.7$; $p < 0.1$], nor did Nasal/Stop [$F(1,144) = 0.09$; $p < 0.77$], while Speaker was obviously significant [$F(1,144) = 1966$; $p = 0$].

Finally, it is interesting to note from Figure 2.10 and both Figure 2.16 and 2.17 that the L2 target (elbow2) presented higher values in questions (114 Hz for LD and 185 Hz for MD) than in statements (109 Hz for LD and 178 Hz for MD). This was true for both speakers. In fact, QS had a significant effect [$F(1,144) = 31.64$; $p < 0.01$], while neither Nasal/Stop [$F(1,144) = 1.33$; $p = 0.25$] nor Open/Closed [$F(1,144) = 0.61$; $p = 0.44$] were. Speaker was also significant. [$F(1,144) = 4185$; $p = 0$]. This result is particularly interesting in view of the contradictory latency data for L2. This point will be discussed in §2.5.

2.4.4 Slope values

The gradient of the LH rise and the HL fall from H within the rise-fall of questions and statements was object of a careful acoustic analysis. At a first approximation, it appeared as if questions were characterized by a shallower rise than statements, since their peak is reached later and spans the entire stressed vowel. Such an impressionistic remark was already made in D'Imperio and House (1997) and was invoked speculatively to account for the intrinsic difference in the results for their declarative-base and interrogative-base stimuli (see §1.5.1). In that experiment, all things being equal, the interrogative-base stimuli produced more question responses than the declarative-base stimuli. The reason for such a discrepancy was then attributed to the difference in rise slope, which had not been controlled for, and was steeper for declarative base stimuli. Therefore, rise and fall slope within the LHL configuration were calculated here through various means, in order to find out if systematic differences between questions and statements could be discerned.

First, the peak velocity within the LH rise was calculated through an automatic procedure that produces a smooth velocity curve of a given f_0 segment and then selects the peak value of that curve. The smoothing is done by convolving with a 35 ms triangular window centered around the target frame.[8] In principle, such a procedure should be somewhat insensitive to noise in the data, since each point on the velocity curve is the weighted mean of itself and 34 surrounding points (17 points before and 17 after, since f_0 sampling was effected at every ms). However, close inspection of a few cases revealed that the procedure

[8] The program was written by Mary Beckman and the choice of 35 ms was modeled after a previous program written by Mary Beckman and Kevin Cohen to calculate jaw velocity.

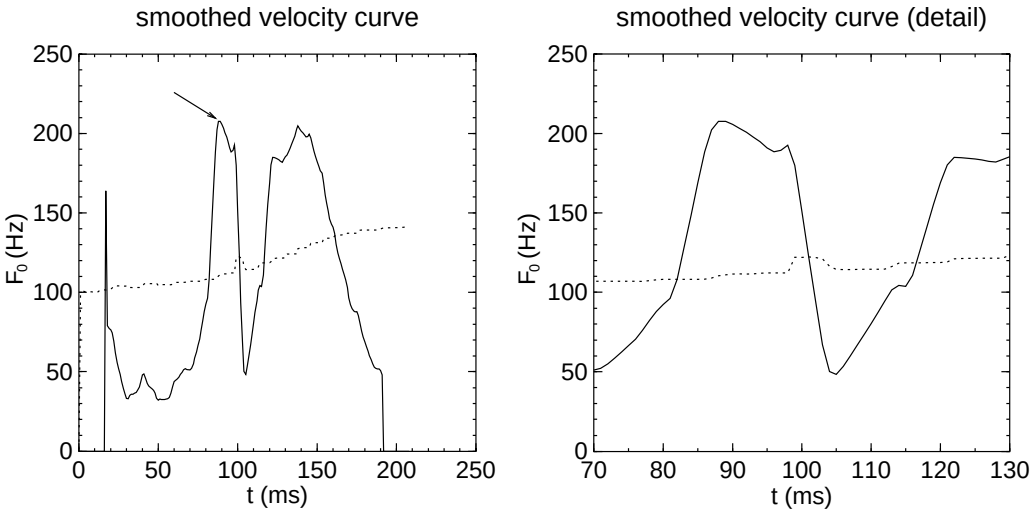

Figure 2.18: Fundamental frequency curve (dotted line) and smoothed velocity curve (solid line) for a statement produced by speaker LD (left) and detail of the accident (right).

is indeed still quite sensitive to segmental artifacts in the data. An example is shown in Figure 2.18 (left). Here the actual f_0 data of a segment of curve spanning from the syllable onset to F0max (dotted line) and the relative velocity curve (solid line) produced by the automatic routine are superimposed. Note that the sudden f_0 jump, though small in intensity, is responsible for the change in direction of the velocity curve (first peak in the solid line). Even more dangerous is the fact that the sudden change in direction of the velocity curve produces a peak that is going to be picked by the automatic routine as the actual peak velocity (indicated by the arrow in Figure 2.18, left). This event is shown in greater detail in Figure 2.18 (right).

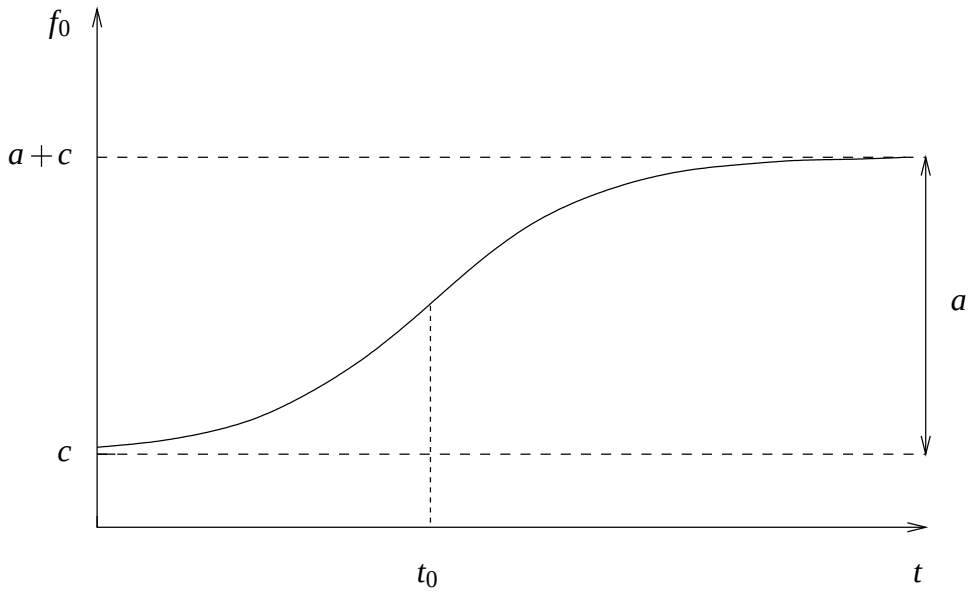

Figure 2.19: Graph of a logistic curve (2.1).

Since the f_0 segments (both rises and fall) present a shape that can easily be modeled in terms of a logistic curve, and because the analysis window could not be enlarged further[9], I decided to employ a logistic model fitting in order to obtain slope values. Note that a simple straight line fitting would not yield satisfactory results, given the particular shape of the f_0 segments under investigation. Consequently, a linear fitting would crucially obtain higher modeling error values than a logistic fitting.

[9]Window size was already equal to 35 ms, which represents between 23% and 35% of the average f_0 segment duration (a LH rise was between 120 and 150 ms, in general). Hence, it would have been undesirable to enlarge it further. In fact, if the window size becomes too wide, the dynamic characteristics of the curve can be lost and the emerging pattern is closer to an average of the f_0 excursion over time.

Specifically, to estimate the peak velocity of an f_0 segment (either the LH rise or the fall within the LHL configuration), a logistic model was fitted to it:

$$f_0(t) = c + \frac{a}{1 + e^{-b(t-t_0)}} \tag{2.1}$$

where t stands for time, t_0 for the temporal coordinate of the curve inflection point, and a, b and c are the parameters of the model (see. Fig. 2.19 for an example). The peak velocity was defined by the slope of the model at its inflection point, i.e., the first derivative of the model measured in t_0. The first derivative of the model is:

$$\frac{d\,f_0(t)}{dt} = \frac{abe^{-b(t-t_0)}}{\left[1 + e^{-b(t-t_0)}\right]^2} \tag{2.2}$$

Here, the peak velocity was therefore estimated by $\frac{ab}{4}$, again expressed in Hz/ms. This model resulted in a very good fit to the data, as can be see in Figure 2.20. As already noticed above, in the comparison with the aforementioned smoothing method, this model is much less sensitive to noise in the data.[10] In order to check the modeling results, visual inspection of each fitting was carried out.

Mean slope values and standard error relative to the f_0 rise region from syllable onset to F0max, obtained through the logistic curve modeling described above, are shown in Figure 2.21. The effect of the QS manipulation on slope values was not significant [$F(1,144)$ = 2.75; $p = 0.1$]. Analogously, neither the Open/Closed [$F(1,144)$ = 0.02; $p = 0.9$] nor the Nasal/Stop manipulation [$F(1,144)$ = 0.06; $p < 0.8$] reached significance. However Speaker was significant [$F(1,144)$ = 33.41; $p = 0$] as well as two of the two-way interactions, i.e., QS by Nasal/Stop [$F(1,144)$ = 15.62; $p < 0.01$] and QS by Speaker [$F(1,144)$

[10] Also, such a procedure does not take into account possible zeroes in the data, in that they are simply ignored by the modeling routine.

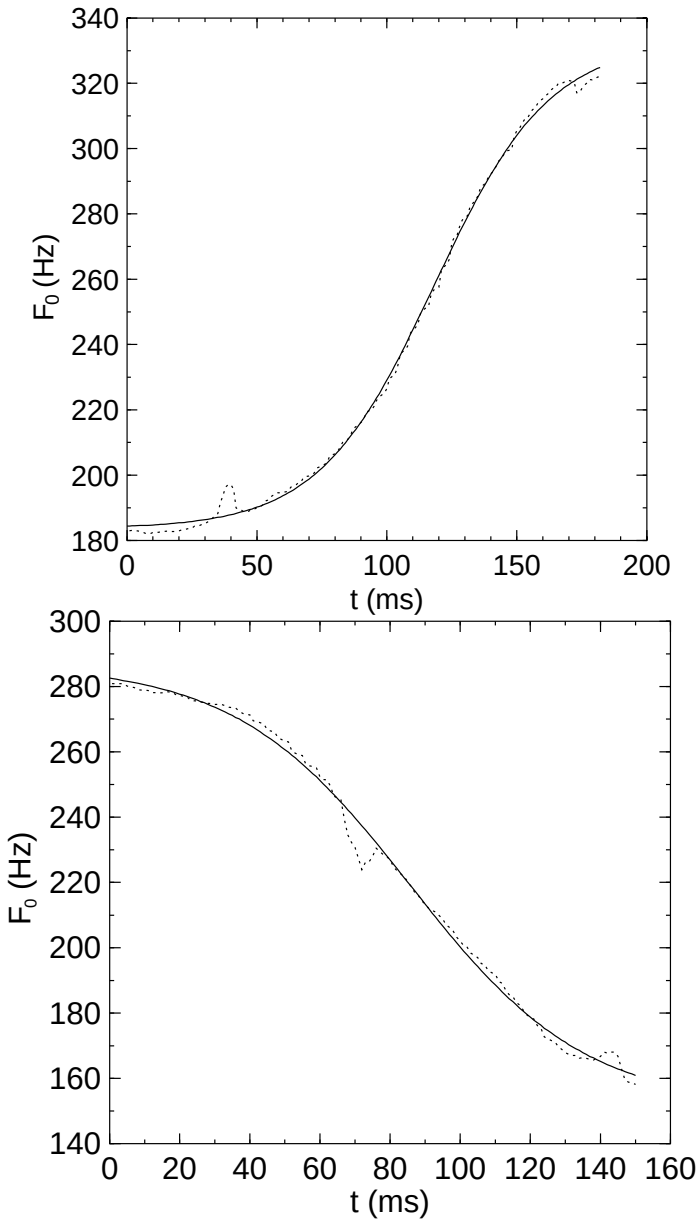

Figure 2.20: Fundamental frequency curve (dotted line) and logistic curve (solid line) fitted to the LH rise within a question produced by MD (upper) and the (H)L fall within a statement produced by LD (lower). The two perturbations correspond to the nasal-vowel edge in the rise and to the vowel-nasal edge in the fall.

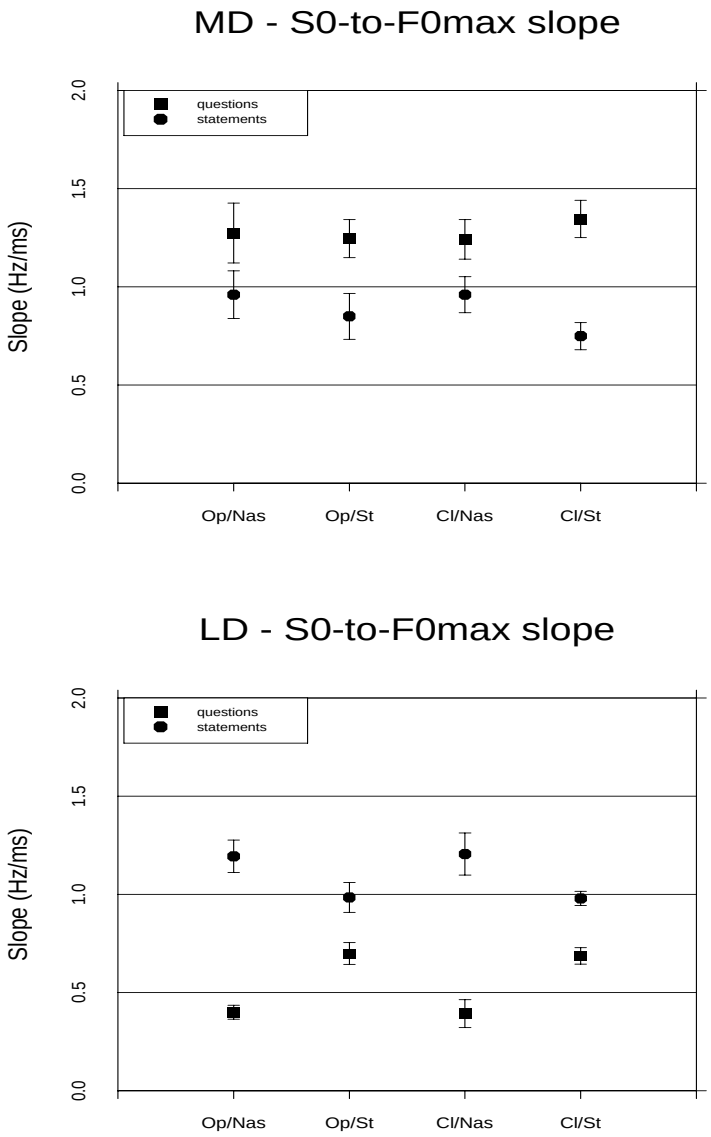

Figure 2.21: Mean slope values in Hz/ms as obtained through logistic modeling of the LH rise in questions and statements for both speakers (Op/Nas = open syllable, nasal; Op/St = open syllable, stop; Cl/Nas = closed syllable, nasal; Cl/St = closed syllable, stop). Standard error is indicated by vertical bars.

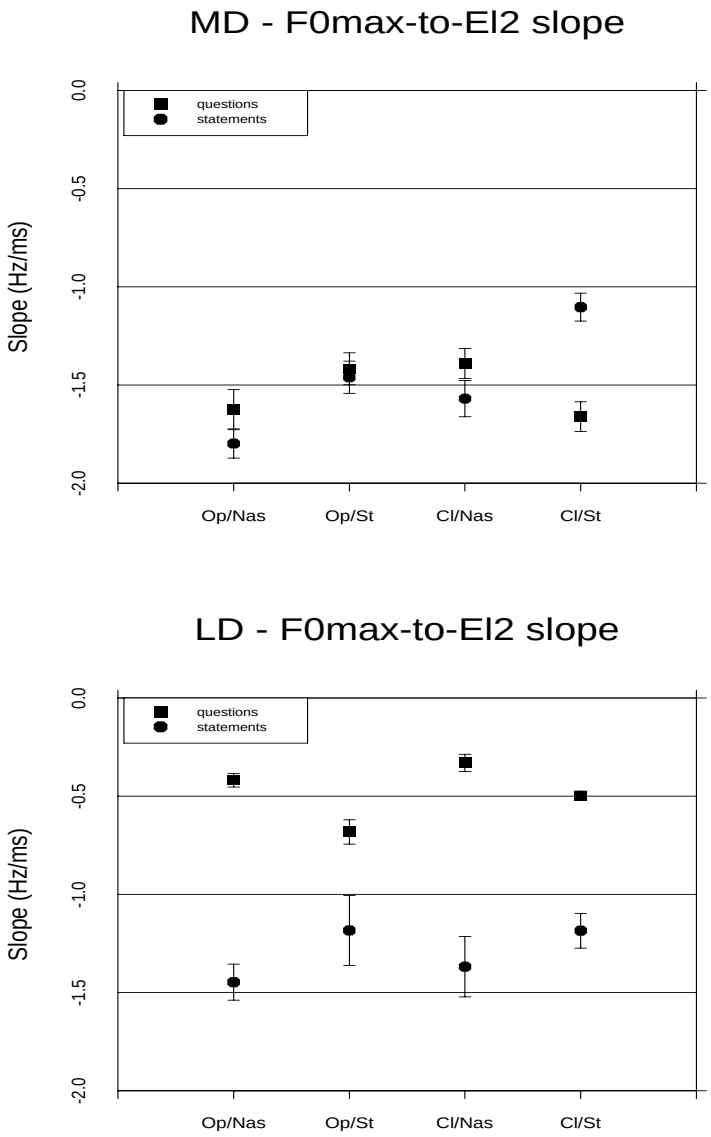

Figure 2.22: Mean slope values in Hz/ms as obtained through logistic modeling of the H to L2 fall in questions and statements for both MD and LD (Op/Nas = open syllable, nasal; Op/St = open syllable, stop; Cl/Nas = closed syllable, nasal; Cl/St = closed syllable, stop). Standard error is indicated by vertical bars.

= 109.64; $p = 0$]. The last interaction is particularly interesting. In fact, as expected from the higher f_0 mean value for statement peaks for LD, slope values for such speaker were greater for statements than for questions (see Figure 2.21, lower panel), and the difference was significant in a three-way ANOVA [$F(1,72)$ = 130.23; $p = 0$]. However, the opposite was true for speaker MD (see Figure 2.21, upper panel). For this speaker, statement rise slopes were shallower than for questions, and the result reached significance in a three-way ANOVA. [$F(1,72)$ = 27.06; $p < 0.01$]. Remember that no significant difference for f_0 peaks was reported for MD, though a minor trend for higher question peaks was noticeable.

A logistic curve was also fitted to the f_0 segment in the region between the peak and the following L2 (elbow2). At a first approximation, it appeared in fact that the H to L2 fall is shallower in questions than in statements. However, this was only true for LD, who had higher peaks for statements than for questions (see Figure 2.22, lower panel). Therefore, the difference in slope was entirely predictable from the f_0 results.

2.5 Discussion

In this production experiment I first tested two hypothesis regarding the temporal alignment of the main targets (L1, H and L2) within the rise-fall configuration of Neapolitan yes/no questions and narrow focus statements. Some previous studies (D'Imperio 1995; D'Imperio 1997a; D'Imperio 1997b; D'Imperio and House 1997) had revealed acoustic and perceptual differences regarding the timing of both L1 and H according to whether the utterance was a question or a statement. Hence, the hypothesis that the pragmatics of the utterance would affect the timing of the peak and the preceding low was expected to

be supported by the data, and indeed it was. Consequently, it was also expected that the timing of L2 would be affected by the coupled "displacement" of L1 and H in questions and statements, since the HL- fall is here assumed to start at the location for the H target of the LH rise, for both questions and statements. The hypothesis was confirmed by the results, which are quite novel (though suggested in D'Imperio (1998b)). The latency effect is predicted on the basis of the postulated secondary association of HL- to the last stressed syllable of the focus constituent. Thus, the results show that the pragmatic contrast affects the alignment of all three targets in a very consistent way. Specifically, L1, H and L2 were earlier in statements than in questions. This result was confirmed for both speakers.

The study also examined the stability of target alignment under variability in both syllable structure and segmental composition of the postaccentual onset. Specifically, it was hypothesized that if targets are aligned relative to syllable or vowel onset, i.e., the left edge of a phonological or phonetic domain/unit (such as the stressed syllable or the stressed vowel), duration variation induced by either structural or segmental factors would not affect the results. The results appear to support this hypothesis, but just as far as the syllable structure factor is concerned. Namely, the Open/Closed factor affected the alignment of L1 and L2 only when alignment was measured relative to the right edge of the syllable (both accented vowel offset and following vowel onset). This was not the case when alignment was measured from the left edge of the syllable. Generally, the targets were timed closer to both vowel offset and postaccentual vowel onset when the target syllable was closed. As mentioned above, closed syllables are characterized by shorter vowels in Italian. Therefore, if targets are to occur at a certain distance from the left edge of a specific unit, the position

relative to the right edge would vary in order to meet the alignment condition. The findings parallel those of Dutch prenuclear L (Ladd, Mennen, and Schepman 2000), in that no relative displacement of L1 alignment was found as a result of differences in phonological length of the stressed vowel when measured relative to syllable (or vowel) onset.

Additionally, one cannot exclude the H peak from such a regularity. In fact, the Open/Closed effect, though almost significant in affecting overall H latency from vowel onset, was not significant in a three-way ANOVA on LD data. Moreover, when H latency was measured relative to syllable onset, the syllable structure effect globally disappeared, offering additional support to the central hypothesis that the tonal targets might be timed to occur at a specific distance from the left edge of the stressed syllable. On the other hand, a syllable structure effect was found for H latency only when it was measured relative to vowel offset, i.e., relative to the right edge of a likely domain unit. The findings are reminiscent of those of Ladd et al. (2000), where a phonological length effect was found on H alignment, which was only measured relative to stressed vowel offset. If peaks were timed to occur relative to vowel offset, however, a very complex situation would arise in Neapolitan Italian. Namely, the alignment of H peaks in statement pitch accents that are associated to a closed syllable (see Figure 2.24) can have alignment specifications that render them almost undistinguishable from question pitch accents associated to an open syllable (see Figure 2.23). Notice that in both cases the accent peak is located very close to the offset of the stressed vowel. This effect was perceptually tested in Chapter 5.

A surprising segmental effect was found, on the other hand, affecting target alignment even when measured relative to vowel and syllable onsets. Such a result was rather speaker-dependent, though. Specifically, the alignment of both L1 and L2, relative to the left edge

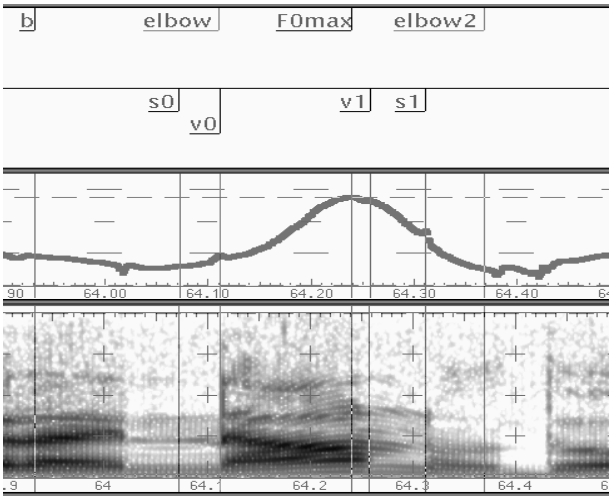

Figure 2.23: Labels, f_0 trace and spectrogram for the question *Vedrai il nano dopo?* produced by MD, showing stressed syllable boundaries (s0 and s1) and stressed vowel boundaries (v0 and v1).

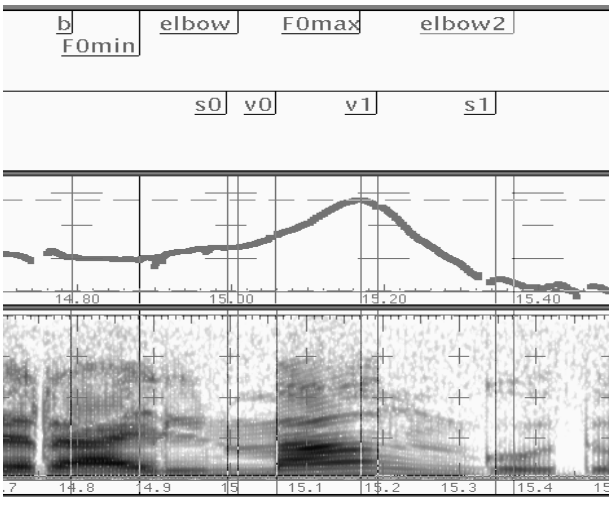

Figure 2.24: Labels, f_0 trace and spectrogram for the statement *Vedrai il mango dopo* produced by MD, showing stressed syllable boundaries (s0 and s1) and stressed vowel boundaries (v0 and v1).

of the stressed syllable, was later if the postaccentual syllable had a stop consonant in onset position. However, this was consistently true (for both L1 and L2) only for MD. LD showed such an effect only for L1 alignment (and quite markedly for questions). A speaker-dependent segmental effect, though on the alignment of H, was also found for Greek by Arvaniti et al. (1998).

Here, remember that for MD, L2 was generally aligned around the onset of the postaccentual vowel onset in statement utterances. Hence, when syllables are closed and followed by a stop, L2 alignment for MD is disrupted by a localized f_0 rise, which is a consequence of a microprosodic effect in the immediate surroundings of the voiced stop. An example is shown in Figure 2.24. In such conditions, the automatic procedure tended to place L2 later than in other utterances, at a point where f_0 starts to fall again after the small rise. But this was the case only for MD closed syllables followed by a stop. As indicated above, L2 generally cooccurred with a much earlier location, i.e., vowel offset, for LD, as it can be seen from Figure 2.25. This might be why the Nasal/Stop effect was consistently found only for MD.

Another interesting result was the syllable structure effect when elbow2 latency was measured relative to the right edge of the accented vowel, i.e., vowel offset. In this case, the target was later in closed than in open syllables. How can we account for such an effect? At this point, one can speculate that the HL- falling gesture has a rather fixed duration. Hence, by the time the f_0 fall is completed, the speaker is closer to the syllable offset and might already be into the post-accentual vowel (as for both MD and LD questions) when the target syllable is open, hence shorter.

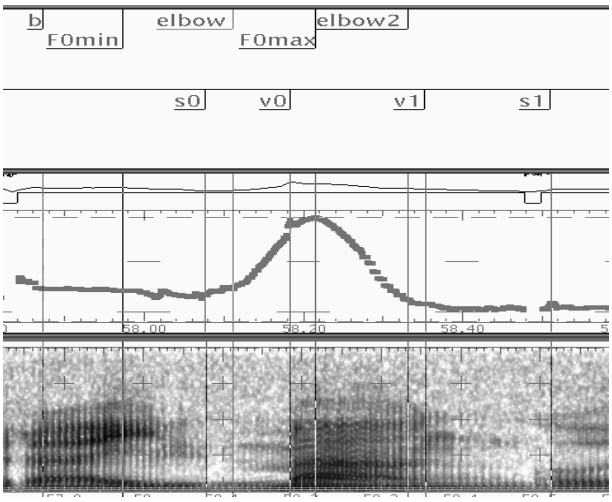

Figure 2.25: Labels, f_0 trace and spectrogram for the statement *Vedrai il mango dopo*, produced by LD.

On the other hand, data regarding the latency of L2 relative to the H peak was quite idiosyncratic, therefore the results do not allow one to draw clear conclusions relative to the duration of the fall from the H peak. I will get back to this point below.

Though the segmental effect on L2 alignment can somehow be accounted for, it is harder to account for such an effect on L1. When latency was measured relative to vowel and syllable onset, both speakers tended to place L1 at later positions when the syllable was followed by a postaccentual stop onset. Such a long-distance segmental effect was entirely unexpected. Unlike the results reported in Arvaniti et al. (1998), no segmental effect on H alignment was found in the present study, independent of the reference point for the alignment measure. The most consistently timed target appears therefore to be the H peak.

When measured relative to syllable onset, only the question/statement contrast appears to affect its alignment, in that question peaks are aligned relatively later than statement peaks.

The latency of L2 relative to the preceding H was also measured. This was done in order to decide whether the HL- fall has a fixed duration in both questions and statements, which would support a homogeneous analysis of this tonal event for both modalities. But while the L2 latency from the preceding H target was the same between question and statements for MD, this was not the case for LD. Here, the rather idiosyncratic behavior of L2 alignment can be accounted for by the overall timing difference between LD and MD target alignment. Remember that MD aligns tonal targets later than LD. L2 was in fact aligned much earlier in LD statements (around accented vowel offset) than in MD statements. The same target was instead aligned very similarly for questions of both LD and MD.

At first, this appears to be problematic, since it has the effect of producing a discrepancy between the HL- fall duration of questions and that of statements only for LD. Nevertheless, the reason for the LD discrepancy might be found in an opportunistic tendency to avoid the realization of the L2 target within a consonantal segment. First, since the H peak is displaced to the right in questions, the subsequent fall results to be displaced too. This is because of secondary association of HL- with the last stressed syllable of the focus constituent. Therefore, L2 results to be displaced as well. However, while such a displacement is proportional to H displacement in MD (so that there is no difference in the latency of L2 from the H peak for the two modalities), LD displaces L2 a bit further than expected. If LD had moved L2 forward so that the HL- fall would have the same duration in both questions and statements, the tone would have been realized either at the onset of the postaccentual

syllable or within the coda of the accented syllable. Apparently, such a strategy is dispreferred. Specifically, LD displaces the fall gesture as far as possible in order to allow L2 to be realized within the postaccentual vowel. The reason for such a strategy is not entirely clear at this point, but it can have its roots in a tendency for tonal events to be realized in regions of relative spectral stability (House 1990).

Therefore, the analysis proposed here is that the H target for the rising accent and the H target for the HL phrase accent become "merged" in a unitary measurable target. No "rearticulation" of the H target is hypothesized here (which would be in any case very hard to test empirically). That is, only one measurable H target should be found in both statements and questions. Plateau configurations, which are quite frequently found for both questions and statements, as mentioned in §2.3.3, are the only cases in which the presence of two acoustically measurable H targets might be discerned. However, such plateau might be the result of "keeping" a unitary H target for a longer time (as in playing a longer musical note). Another outcome of tonal crowding is that H targets are earlier when the nuclear rise is on a single-word focus constituent than when it occurs on a multi-word focus constituent. Quantitative data backing up this claim are already available for questions (D'Imperio 1998b; D'Imperio 1999; D'Imperio in press), while only qualitative data are presently available for statements (though quantitative results are being analyzed).

The above findings are taken as supporting the hypothesis that the HL- phrasal fall has a secondary association with the nuclear syllable. Hence, while the H target of HL- is merged with the H of LH pitch accents, its L target (i.e., L2) follows the H at a fixed distance (though this result needs to be verified on more speakers), unless a constraint on tonal alignment with the segmental string forces the speaker to "move" L2 in regions of relative spectral stability.

The f_0 evidence regarding L2 is quite puzzling, though. If we assume that the HL- fall is the result of the same phrasal gesture in statements as well as in questions (as already proposed for the HL fall of questions in D'Imperio (1997b)) then we would expect L2 to attain the same f_0 level in both modalities. But this was not the outcome of the f_0 analysis. Instead, L2 was significantly different in f_0 terms between questions and statements, for both speakers. Namely, L2 was higher in questions than in statements, though the difference was, on average, only equal to 5 Hz for LD and to 7 Hz for MD. Though small, such a difference appeared to be accompanied by a significant difference in slope gradient between the two falls. Nevertheless, individual analyses revealed that the slope difference was consistently maintained only by LD. This is quite suspicious, given two points. First, as noticed in §2.4.3, LD produced always higher H peaks for statements than for questions. Hence, the steeper slope gradient for his statements can be plausibly analyzed as an epiphenomenon of the different emphasis degree attributed to utterances in the two modalities. Second, as discussed above, it appeared as if LD employed a strategy to displace L2 as far as possible towards the onset of the postaccentual vowel. This could only render the HL fall of questions shallower than the fall of statements. A different control of the fall in questions and statements is therefore a question open for future research.

Also, a consistent difference in the F_0 level of L1 (elbow) was found. For both speakers, the elbow was higher in statements than in questions, and the difference was especially conspicuous for MD (39 Hz, on average).

Unlike L1 and L2, H peaks did not show a consistent f_0 difference between questions and statements. Though the effect was significant when results of both speakers were pooled, it was found that only LD produced a significant difference for the individual results. Moreover, the difference went in the opposite direction to that expected from other languages, for which questions are generally characterized by higher peaks and expanded f_0 range (see Herman (1997) for Balinese, Jun and Oh (1996) for Korean, Maekawa (1991) for Japanese, *inter alia*).

Finally, the hypothesis that the LH rise would be shallower in questions relative to statements was not supported by the overall data. Moreover, the hypothesis that the slope of the HL- fall would not differ according to modality was verified only for MD data, while support was found for LH rises. Again, the speakers showed an opposite behavior. Such a behavior was expected, to a certain extent, given the difference in the f_0 height results. In fact, LD showed steeper LH rises for statements, while MD showed steeper LH rises for questions.

Slope will be therefore manipulated for both rise and fall in the perception experiment that will be presented next. Given the contradictory slope results, we do not necessarily expect such a factor to influence perception in a significant way. Additionally, if the perception results show indeed an effect of slope, we expect the effect to be in the direction of LD data. In other words, we expect steeper slopes to signal statements and shallower

slopes to signal questions. Nevertheless, one is compelled to remark that slope is a parameter that shows too much variability, depending dramatically on f_0 excursion, which, in turn, depends on the degree of emphasis applied to the utterance.

Also, though the reverse was expected, f_0 peaks were higher for statements than for questions. However, this result was inconsistent among speakers. Unlike H peak results, L1 and L2 were consistently affected by the question/statement contrast. Here, though, neither structural nor segmental factors seemed to affect the results. Hence, as for rise-fall alignment, also rise-fall f_0 height will be manipulated in a perception experiment in order to test its impact on the question/statement identification.

To summarize, the hypothesis that the alignment of L1, H and L2 (expressed in terms of the elbow, F0max and elbow2 measure) would be affected by the question vs. statement contrast was confirmed. This replicates the findings of D'Imperio (1995), D'Imperio (1996) and D'Imperio (1997b). Additionally, the hypothesis that segmental environment would affect the alignment of the H peak when measured relative to syllable onset was not confirmed. The hypothesis was instead supported for L1 (elbow) and partially for L2 (elbow2) alignment. Finally, the hypothesis that syllable structure would affect target alignment relative to the left edge of the syllable was not confirmed, while a difference was found when alignment was measured relative to the right edge of the syllable. Assuming that the strong hypothesis of invariant alignment of such targets relative to syllable or vowel onset holds, it will be interesting to see how such factors interact in the perception of the question/statement contrast.

CHAPTER 3

PERCEPTION EXPERIMENT I: PERCEPTUAL "TARGETS" AND PITCH ACCENT IDENTIFICATION IN NEAPOLITAN ITALIAN

3.1 Rationale

The scope of this chapter is to first verify some hypotheses about the role of target alignment in the perception of pitch accent contrasts in Neapolitan, and, then, to propose a theory of tonal target perception in an abstract two-dimensional space. The main question I would like to answer here is "what is the perceived target of a complex, dynamic pitch transition, specifically of a pitch accent?". By perceived target I mean the auditory percept of a pitch transition (such as a rise, or a fall), defined by a set of coordinates including both the frequency domain (scaling) and the time domain (alignment). Within the psychoacoustic literature, we saw that a tone glissando can be matched to a single perceived frequency value (see §1.6). That is, a rising accent, analyzed as a sequence of a L and a H tone in standard autosegmental terms, might be mapped by the listener to a single perceived target value in a bidimensional (time by f_0) space.

What I would like to propose is that speakers have knowledge of some abstract coordinates for the perception value of a pitch accent, and that they employ such a stored value to either identify linguistically significant contrasts or to produce them. One of the

possible linguistic contrasts that Neapolitan speakers can produce is that between the LH rise of yes/no questions and the LH rise of narrow focus statements, whose shapes are very similar (as we saw in Chapter 2), though the timing of their tonal targets appears to be systematically manipulated. Of course, this is a simplified view of the mapping of shape onto target localization, and I shall not propose a model here. I just propose that the results presented in this Chapter appear to show that timing is very important for the definition of perceptual tone targets and that the computation takes into account local f_0 values. I will also postulate that the whole rise-fall contour might be abstractly translated into a target, "gestalt" configuration, that is the result of a weighted averaging process of the melodic configuration within the accent. The experiments presented here are also intended to test the main hypothesis that a perceived target value can be manipulated not just through obvious peak alignment manipulations, but by more subtle manipulations of peak shape and low target timing.

The notion of target in the production literature has been around for quite some time. For instance, in the vowel production literature (Lindblom 1963; Moon and Lindblom 1994), formant target values are those that the speaker attempts to reach when producing a specific vowel in a specific context. The concept of a target in tonal production is more controversial, as discussed in §1.4 in Chapter 1. Additionally, within the vowel perception literature, we find the notion of perceptual target for specific formant configurations (Johnson, Flemming, and Wright 1993; van Son 1993). On the other hand, within the tonal literature, the notion of target can be solely extrapolated from psychoacoustic results (see §1.6 above). It appears, for instance, that tonal transitions can be mapped onto a (pitch) target value that corresponds to the pitch of a stationary tone. I argued in Chapter 1 that it

is not easy to use psychoacoustic results in order to explain linguistic data, but I also advocate the view that we should employ some of those results in order to formulate working hypotheses that can be tested with linguistic data. This is a process of continuous testing and then inferring back from the data, in absence of any widely accepted theory of tonal perception in speech.

In Chapter 2, it was found that all three acoustic targets (L1, H and L2) of the rise-fall characterizing questions and statements in Neapolitan are timed differently according to modality. Namely, all three targets are timed later in questions than in statements. Hence, in §3.3.1, I first test the hypothesis that by shifting the alignment of those targets within the stressed syllable, while keeping everything else constant, I can indeed determine a perceptual category shift from question to statement. Again, my assumption here is that a perceived *gestalt* is mapped onto some value, which is associated to a linguistic category (question vs. statement). Then, in §3.3.2 I test the effect of peak *shape* on perceived target location. What I will propose is that pitch accent shape indirectly affects tonal target perception by modifying the temporal alignment of the perceived tone value. By tone value I mean the target value within a tone transition that is attended to perceptually, and that is salient for the listener. Alternative locations could either be, say, the 2/3 of the transition proposed in the psychoacoustic literature (Nábeľek et al. 1970), or values determined in alternative ways, which I shall mention later.

As discussed in Chapter 1, recent work appears to show that tonal alignment is systematically manipulated both in production and perception. Nevertheless, the alignment of acoustic targets proves to be a very unstable index. It is indeed quite puzzling why certain

tones tend to be aligned "beyond" the boundaries of the associated TBU, e.g., the prenuclear peak accents of a variety of languages (such as English, Spanish, Italian and Greek), as mentioned in §1.1. A concept that is not foreign to the vowel perception literature is that target production might be guided by the communicative situation, and that therefore production targets could be a "consequence" of perception targets. Hence, we can employ such line of reasoning for tonal perception purposes. For instance, one plausible account of the superficially aberrant alignment of tonal targets in production might have its basis in the psychoacoustic of the perception of pitch glissandi. In other words, the lack of synchronization between accent peak and associated TBU might be accounted for in terms of perceived target. That is, the target measured in the fundamental frequency curve, and that generally corresponds to a H or a L tone, might be mapped in specific ways to the target that is perceived as such by the listener.

Within the psychoacoustic literature, it has been suggested that the perceived pitch of a rising glissando is lower than the measured frequency at the very end of the same glissando (see §1.6). Speculatively, this might be the basis for the preference of H accents to be aligned late, even beyond the boundaries of the TBU. This is because the perceived pitch value might not only be lower than the measured f_0 peak, but might also be perceived as being earlier in time. Thus, the temporal relationship between the tone and the TBU would be in some sense kept, despite the "acoustic" misalignment. In other words, it is plausible that such perceived value, which appears to be melodically different from the value measured within the acoustic "target", might also be perceived as displaced in the time dimension. To make an analogy with the production literature, remember the hypothesis proposed by Silverman and Pierrehumbert (1990) about "tonal repulsion" of tonal targets

(see §1.4), which accounts for the displacement of acoustic target peaks. What is proposed here is that the mapping between alignment of tonal targets in perception and alignment of tonal targets in production is a complex one, but that we can discover more about it by using linguistic as well as psychoacoustic tasks.

Furthermore, it has been found that if the glissando is followed or preceded by a plateau, the perceived pitch value will correspond to the pitch of the plateau (Nábeľk et al. 1970) (see §1.6). As mentioned in Chapter 2, instances of plateaus were found for both statement and question utterances, therefore I do not assume that such a phenomenon can cue a specific modality *per se*. I also assume here that there is just one peak target location in such configurations, and not two. That is, I assume that the target H is maintained for a longer time than in peak accents, which results in a plateau. The other crucial assumption is that such plateaus are also perceived in terms of a single H target, and that parsing two separate targets is not possible (when plateaus are as short as in the present experiments).[1] In such cases, though, it might be difficult to decide where the acoustic target is to be measured in the first place. Either plateau onset, offset or even its midpoint might be equally good candidates in absence of perception evidence. One of the goal of this thesis was in fact to gauge temporal localization of perceived targets in such tough cases. This was done indirectly through question/statement identification. 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

[1] It must be noticed that it is common practice in some speech synthesis techniques (as within the Dutch IPO-style synthesis) to employ short plateaus of 30-40 ms also to model simple peak accents, in that they result in higher stimulus naturalness.

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.

Remember from §1.5.1 that the results of D'Imperio and House (1997) could be interpreted in a variety of ways regarding the rise-shift manipulation (Figure 1.12), which scored a majority of question responses already at early locations within the continuum. Namely, it was reported that (though peak alignment could indeed shift the perception from a question to a statement and vice versa) the perception of a "clearly perceptible rise" within the stressed vowel was a stronger cue to question identification, i.e., stronger than peak timing. But some objections can be raised against such a conclusion, which are: 1) target location was not explicitly defined nor tested for rise-shift stimuli, whose peaks had a flat shape (plateau); 2) the two stimulus series created from the interrogative base, such as the rise-shift and the inter-peak stimuli, had a different rise slope (shallower) than the declarative based stimuli. Neither slope nor shape was constant among the stimuli employed, then. Therefore, the effect of potential cues other than peak timing, such as pitch accent shape (e.g., slope, peak shape, etc.) was not accounted for. It is then likely that the results of that study might have been biased by the particular characteristics of the base stimuli, specifically the particular shape of the rise-fall accent, whose rise and fall gradient, as well as peak shape, differed between the two base stimuli. Also, the effect of another potential cue to the contrast, such as f_0 level, was not tested. In fact, f_0 differences between the nuclear accents of the two base stimuli were simply eliminated through the stylization procedure.

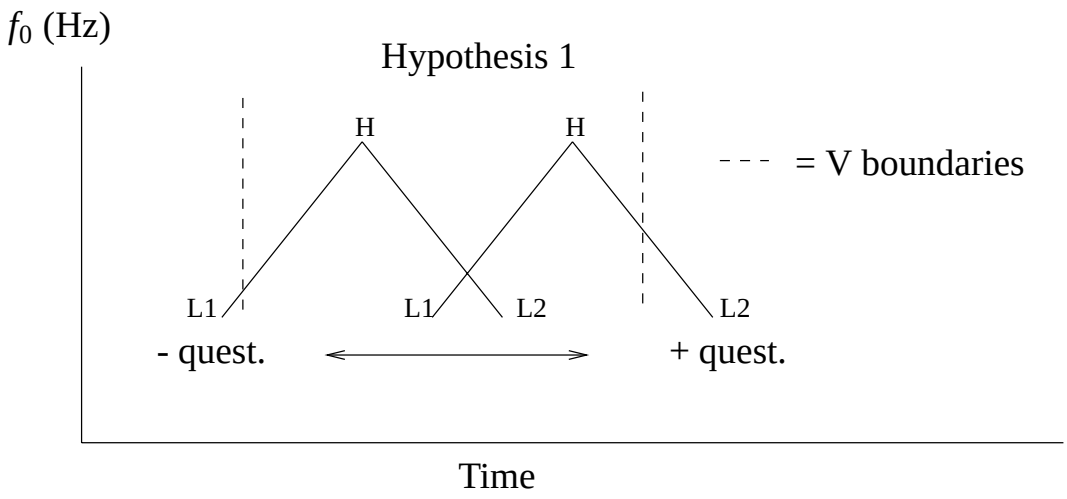

Figure 3.1: Schematic representation of hypothesis 1.

Hence, a more controlled study is presented first, in which I intend to replicate the findings of D'Imperio and House (1997) regarding the timing manipulation for those stimuli in which the entire rise-fall contour was shifted in time (such as the inter-peak series, see Figure 1.11) and then use those results as a reference for further comparisons (e.g., comparison with stimuli in which shape and pitch range are manipulated orthogonally with timing), with the aim to discover the mechanism of perceived target mapping in all conditions. The continuum created for this purpose was labeled "primary continuum" and the hypothesis tested through such manipulation will be referred to as hypothesis 1 (see Figure 3.1). By this hypothesis, when the LHL configuration is moved earlier within the stressed vowel, a lower percentage of question identification responses should be obtained. We also hypothesize that a category boundary will be found in the middle of the continuum.

Once I have shown that target alignment can induce a category shift (which we expect both from the results of D'Imperio and House (1997) as well as from studies on the perception of English and German pitch accents reviewed in §1.5 above), in §3.3.2 I will go on to test another hypothesis. The central working hypothesis (Hypothesis 2) will be that the shape of the accent peak modifies the coordinates of the perceived targets and, consequently, the identification of the question/statement contrast. One way to test such a hypothesis is to verify some alternative predictions, all aiming at understanding the mechanism of perceived target location in "plateau" versus "peak stimuli".

Therefore, a second stimulus series was created, labeled "plateau continuum". Such a continuum was intended to systematically control for the "plateau shape" of the rise-shift stimuli in D'Imperio and House (1997). In that study, plateau duration was not explicitly controlled, and was in fact varying from a stimulus number to the other, within the rise-shift and fall-shift series. Here, instead, plateau duration was always equal to 3 steps of the primary continuum. In this fashion, I will be able to test at least two hypotheses about target timing location. First, I will compare the results of the primary continuum (i.e., the peak stimuli) with those for plateau stimuli whose plateau starts at the peak of some of the primary continuum stimuli. The first null hypothesis stemming from our central hypothesis 2 (hypothesis 2a) is that the number of question responses will be the same for plateau stimuli and peak stimuli whose peak is timed at plateau onset. This hypothesis is depicted in Figure 3.2. In other words, I will test if the "target" of the rising transition of plateau stimuli is timed to occur at the end of the LH rise (which was implicitly assumed in D'Imperio and House (1997), see discussion in §1.5.1).

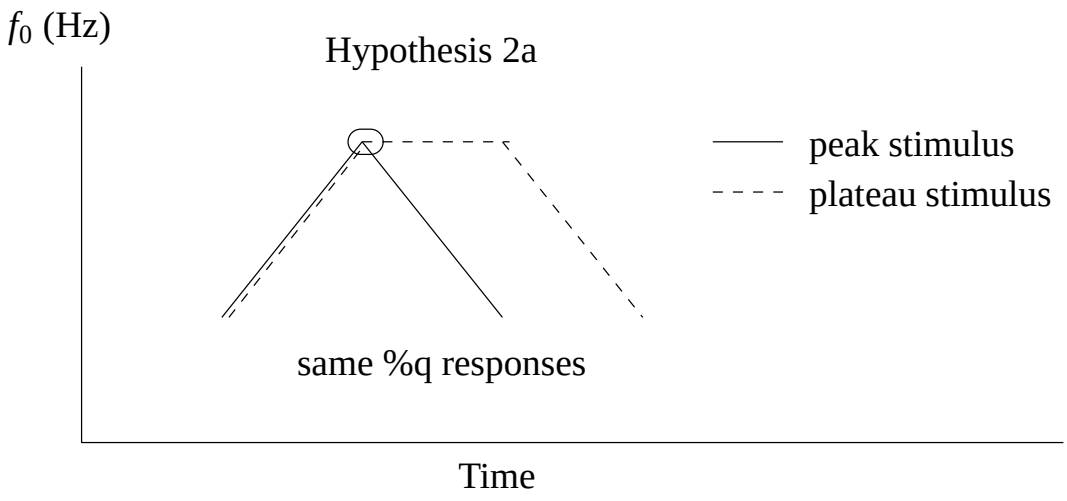

Figure 3.2: Schematic representation of hypothesis 2a.

If the hypothesis is not verified, then I 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. The hypothesis (hypothesis 2b) is illustrated in Figure 3.3.

Note, however, that both hypotheses 2a and 2b treat any difference in perceived target location between the two series as exclusively due to time domain effects. 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. This is quite plausible on the basis of two main findings. First, remember that in quite a few languages (e.g. Balinese, Korean, Japanese; see discussion in §2.5) questions seem to

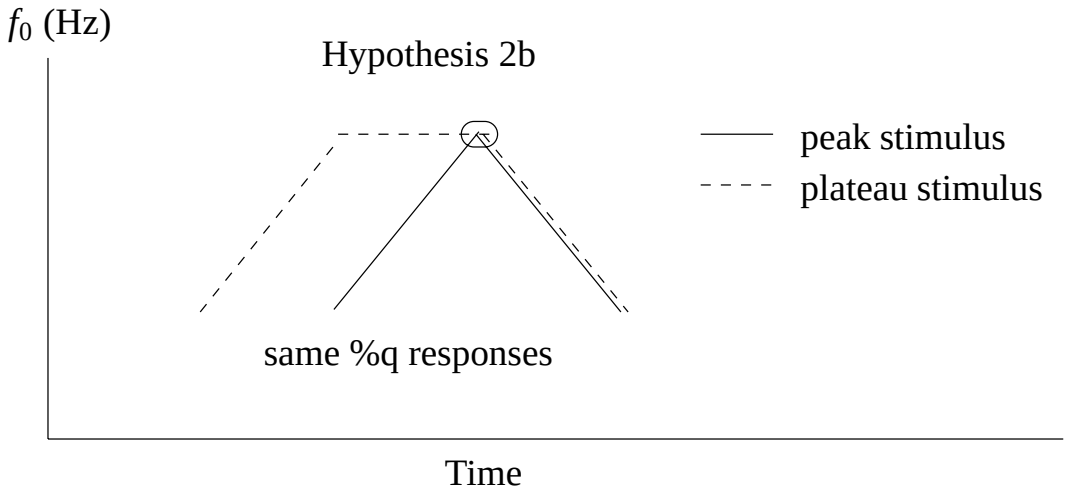

Figure 3.3: Schematic representation of hypothesis 2b.

be associated to higher f_0 peaks and/or expanded f_0 range. Moreover, it appears that in Hungarian, higher peaks successfully cue questions in perception (Gósy and Terken 1994).

In Chapter 2 we saw, though, that, from a production point of view, the Neapolitan situation is quite complex. In fact, while one of the speakers presented no f_0 difference at all between question and statement H peaks, the other showed a difference in the opposite direction to the one expected on the basis of results from other languages. That is, for this speaker, statement peaks were higher. I offered a plausible interpretation of this result by invoking an effect of emphasis. That is, the speaker that showed such a contrast seemed to mark statements with a higher degree of emphasis, which usually translates in an increase in peak height. However, we cannot totally dismiss the possibility of a perceptual effect of f_0 height on question/statement identification.

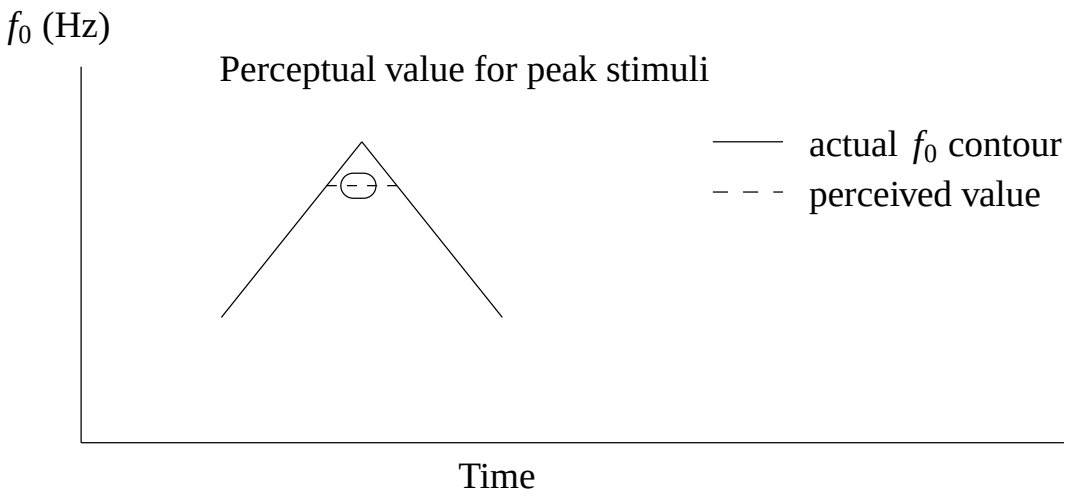

Figure 3.4: Schematic representation of predicted tonal value for peak stimuli perception.

Also, as discussed in §1.6, the results of a number of psychoacoustic studies appear to converge on the hypotheses that the perceived pitch of dynamic tone transitions (glissandi) is equal to the level reached only at roughly 2/3 of the transition. Therefore, it is likely that the perceived value of peak stimuli is lower in pitch, and therefore mapped onto a perceived target value that is different from that of a plateau stimulus. In other words, it is possible that, analogously to what appears to be true for pure tones, the actual peak target value perceived by the listener is lower than the one measured acoustically. This is shown in Figure 3.4.

One might compare such an effect to perceptual *averaging* of formant transitions (van Son 1993). Such an effect can be explained as a weighted mean of the frequency values associated to a formant throughout a vowel (i.e., the entire formant contour). In such a

case, though, the perceived formant target is not equal to a frequency relative to a specific location within the formant contour, since the averaging (though weighted) is made on the basis of all the frequency values. This means that by changing the values and the shape of the formant contour, the averaging process leads to quite different results.

Note that the literature on vowel perception presents controversial evidence for the hypotheses of perceptual averaging mentioned above vs. the hypothesis of perceptual compensation or "overshoot". Formant target "overshoot" occurs when the measured acoustic target is less extreme than the perceived target for particular frequency trajectories (Lindblom and Studdert-Kennedy 1967; Nearey 1989). Formant trajectory shape (and contextual variables) appear to trigger either averaging or overshoot. Therefore, formant shape is relevant for vowel identification.

The perceptual effect of transition shape for tonal contours is much more controversial, both from a production and a perception point of view. Nábeľk and Ovchinnikov (1997) remark that perceptual overshoot has never been found in tone perception studies, while some sort of averaging seems to be supported by earlier works (e.g. Nábeľk et al. (1970), Schouten and Peeters (1995)). However, these older studies also show that, if a perceptual averaging is indeed performed, it might be "blocked" when the glissandi are either preceded or followed by a plateau, since in that case the perceived pitch corresponds exactly to the value at the plateau. Nevertheless, one cannot dismiss the hypothesis that the plateau section of the contour might be "weighed" more in the averaging process, to a point that it outweighs the other values within the contour.

A difference in the perceived pitch value of plateau stimuli vs. sharp peak stimuli has also been found in synthetic speech perception. 't Hart (1991) reports on a similarity judgment for linearly vs. parabolically stylized tonal stimuli, finding that "flat peak" (plateau) linear stimuli are perceived as having the same pitch height as parabolic stimuli. But, crucially, flat peak linear stimuli are perceived as being higher than corresponding sharp peak linear stimuli. This suggests either that the pitch target of plateau stimuli was, so to speak, "overshot", or, more plausibly (given the psychoacoustic findings), that peak stimuli were perceived at a lower value. If a difference in perceived pitch between plateau and sharp peak stimuli exists, it is an empirical question whether it can affect the question/statement distinction. Also, if the score for question identification is found to be different between plateau and primary continuum stimuli here, one can question whether the nature of the difference is to be identified with a pitch level effect. The interest of this issue comes also from the fact that, in production, Neapolitan speakers do not seem to associate a clear pitch target difference to the two modalities.

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. One of our null hypotheses will then be that pitch accents with the same shape but different overall f_0 height would not induce different scores of question responses. Such a prediction is depicted in Figure 3.5. 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 value difference. This might also mean that the effect might not be due to an "averaging" of the f_0 transition within peak stimuli.

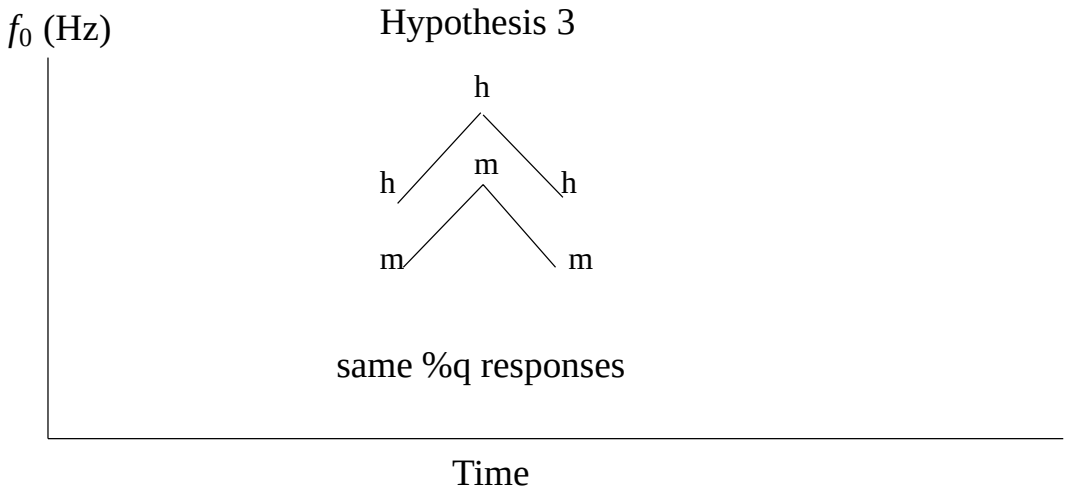

Figure 3.5: Schematic representation of hypothesis 3.

To summarize up to this point, the central hypothesis that peak shape modifies perceived target timing needs to be tested in two ways. First, we need to find evidence that target alignment does affect the perceptual target location that is mapped onto the question/statement contrast. Then, we need to make sure that the nature of the difference between plateau and peak stimuli scores is in the timing of their perceived values, and not in some other property, such as pitch height. For this to be true, it must be necessarily the case that peak stimuli differing only in the f_0 values of their acoustic targets will also receive same scores of question responses. This hypothesis was tested and results are reported in §3.3.3.

Note that, if the difference between plateau and peak stimuli perception has its basis in a general psychoacoustic effect, we expect to find it irrespective of the language spoken.

That is, we should find a comparable difference in listeners of different language groups. It will be interesting then to test the predictions with non-Italian speakers. This was done through the experiments presented in Chapter 4, in which the hypothesis that the shape of the peak affects perceived target is evaluated for American English listeners.

The shape of the accent, though, is also determined by dynamic characteristics of the rise and fall. By dynamic characteristics I mean those aspects of the transition between two tonal targets that include transition slope, transition duration, etc. Most obviously, an f_0 target can be reached by means of transitions whose velocity can differ. Such velocity differences are reflected in the gradient or slope of the pitch movement itself. Slope differences can be contextually determined by various factors, such as gestural overlap and tonal crowding (cf. Silverman and Pierrehumbert (1990), D'Imperio (1998b)), which in turn produce timing and/or melodic readjustments of the target tones. Some researchers have also suggested that slope can vary with speech rate changes in languages such as Dutch (Caspers and van Heuven 1993) and British English (Ladd, Faulkner, Faulkner, and Schepman 1999), in which pitch accent rises were investigated.

Remember that the slope results reported in §2.4.4 were not easily interpretable, since just one of the speakers, LD, showed a significant difference between questions and statements. Specifically, the difference was in the direction of steeper slopes for statements.

However, I argued that such a difference might have been caused by increased emphasis in the statements, which is related to the pitch height difference mentioned above. In any case, if such a difference was not an artifact of a difference in scaling, but indeed statements are characterized by steeper slopes, than we would expect slope to affect perception.

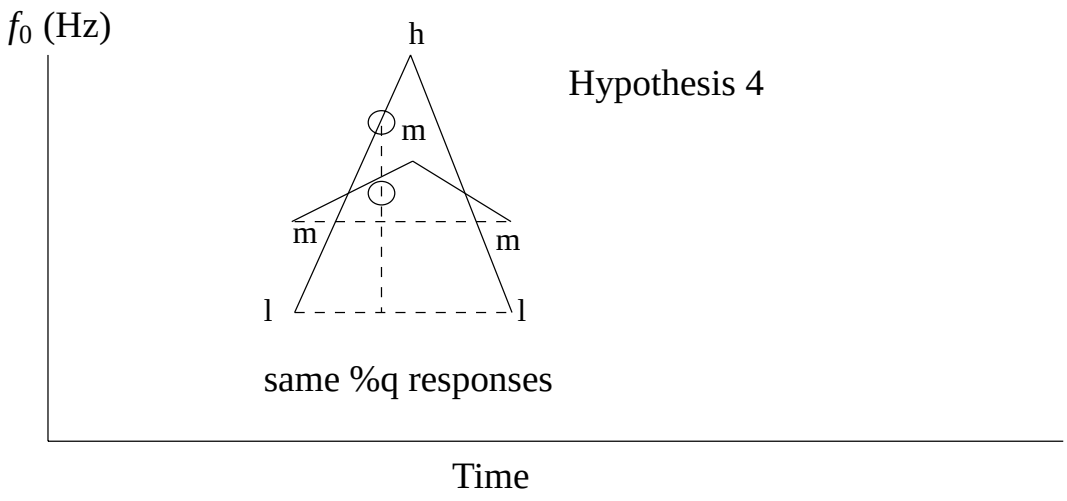

Figure 3.6: Schematic representation of hypothesis 4.

Namely, our null hypothesis will be that steeper or shallower LH rise and HL fall will produce the same percentage of question responses. The predictions of this hypothesis, which will be our hypothesis 4, are schematized in Figure 3.6.

Note that the hypothesis crucially relies on the assumption that contours with different slope but same timing of L1, H and L2, will have the same perceived timing of the tonal target. This assumption is based on a geometric property of the contour (Thales theorem). In trigonometry, it is easy to demonstrate that the time coordinate of a point at a certain fraction of the rising (or falling) slope of a triangle is at exactly the same time coordinate of the same fraction at the triangle base (see Figure 3.6).

Specifically, I assume that if, say, the perceived target value of the LH rise is at 2/3 of the transition, this will always be true of any contour.[2] In Figure 3.6, therefore, the location at 2/3 of the rising transition, for instance, corresponds on the time axis for both contours.

If we find that pitch target level has no effect on percent question responses (that is, if hypothesis 3 is confirmed), I will interpret the results as showing that the perceived target for those stimuli is the same. Therefore, an alternative hypothesis will be that any observed difference is to be attributed to some other factor. One could object, though, that the contours have different starting points for the rise and different end values for the fall, in that L1 and L2 have different f_0 (low in the "lhl" case and medium in the "mmm" case). That is why the analysis will include contours with same timing but higher L1 and L2 pitch (all high targets, i.e., "hhh" combination), in order to test any additional effects caused by such variable.

Finally, if I can sufficiently show that timing characteristics of the whole rise-fall contour affect perceived tonal target, I can go on to test the impact of independent alignment modifications of L1 and L2. This is an interesting question, since most alignment research has focused on the alignment of the H peak. In other words, I will test if the timing of L1 and L2, independent of H timing, might cause a shift in target perception. This might also suggest that the timing of the entire LH or HL transition is responsible for target timing perception, and not peak location *per se*.

[2] We know, though, that slight differences can be found, since in the psychoacoustic literature it is reported that the perceived pitch value is closer to transition end for longer transitions. But for our purposes, since the stimulus transition duration does not change dramatically here, I believe that the assumption is correct.

Most obviously, by shifting L timing, while keeping H timing constant, one modifies slope as well as timing. However, if we verify the null hypothesis (hypothesis 4) that stimuli with different slope will not obtain different results, any observed difference will have to be attributed exclusively to L1/L2 timing. Also, remember from Chapter 2 that both L1 and L2 target timing were sensitive to the question/statement manipulation. Therefore, one would also expect later L1 and L2 timing relative to vowel onset to produce a higher percentage of question responses.

Additionally, in Chapter 2 I proposed an analysis by which the first L in the rise-fall configuration (L1) is part of the pitch accent proper (either L*+H or L+H*, depending on the modality), while L2 is the end target of a phrasal HL fall. Such a phonological difference might predict a different behavior in terms of signaling the question/statement contrast. Hence, the final null hypothesis (hypothesis 5) will be that stimuli with same H target location but different L1 and/or L2 timing will score the same number of question responses. Such a hypothesis is congruent with two predictions. First (hypothesis 5a), stimuli with different L1 timing will produce the same number of question responses. This hypothesis is shown in Figure 3.7. We also predict that stimuli with different L2 timing will produce the same percentage of question responses (hypothesis 5b), which is represented in Figure 3.8.

In sum, I will test 5 hypotheses relating to the perception of rise-fall accent configurations, in which either the timing of the entire rise-fall, the timing of either L1 or L2, peak shape, rise and fall slope or f_0 level are manipulated. Such working hypotheses are related to psychoacoustic findings on one side, and to tonal target production on the other. The results will be discussed in terms of a theory of tonal target perception.

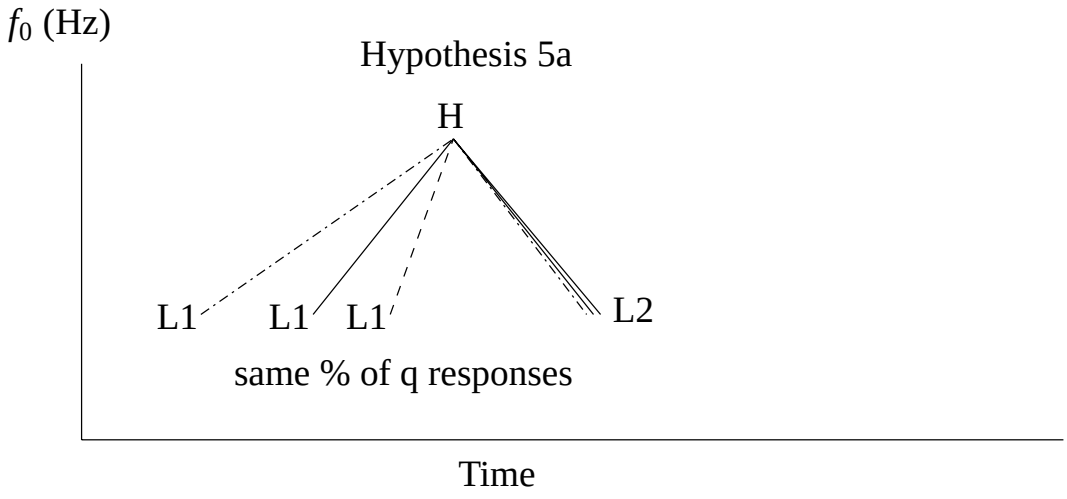

Figure 3.7: Schematic representation of hypothesis 5a.

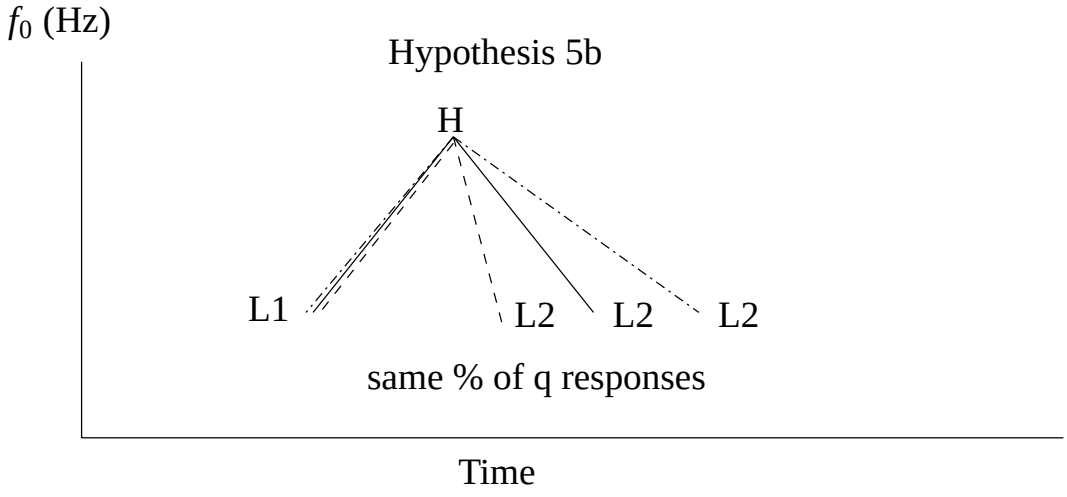

Figure 3.8: Schematic representation of hypothesis 5b.

3.2 Methods

3.2.1 Stimuli preparation

First, I intended to replicate the findings relative to the inter-peak series of D'Imperio and House (1997) (see §1.5.1), in which the timing of the peak of an interrogative base stimulus was shifted backwards, from vowel offset to vowel onset. Specifically, the hypothesis tested is that (hypothesis 1) such manipulation would lower the number of question responses in a question/statement identification task. To this end, the timing of the three tonal targets, i.e., L1, H and L2, was simultaneously manipulated to create a continuum of 7 stimuli, which will be referred to as the "primary continuum" (see Figure 3.9).

The base stimulus, from which the continuum was created, was selected from a set of read sentences produced by a native speaker of Neapolitan Italian (the author). The chosen utterance was *Vedrai il nono?* "Will you see the ninth?". The tonal structure of this utterance is comparable to the utterances of the production study, being characterized by a focal L*+H on the noun *nono*. No question utterance of the production study could be employed for the purpose of the perception experiment since all questions presented a marked post-focal/postnuclear region, with a !H* accent on the adverb *dopo* "afterwards". This region is crucially different between early focus statements and questions, in that questions allow postnuclear accents to be realized, while statements do not (D'Imperio 1997b; D'Imperio in press). Hence, the presence of such an additional cue to statement/question identification might have clearly influenced the results. Nevertheless, the timing and duration properties of the new utterance were comparable to those of the production study, having a similar tonal and segmental structure.

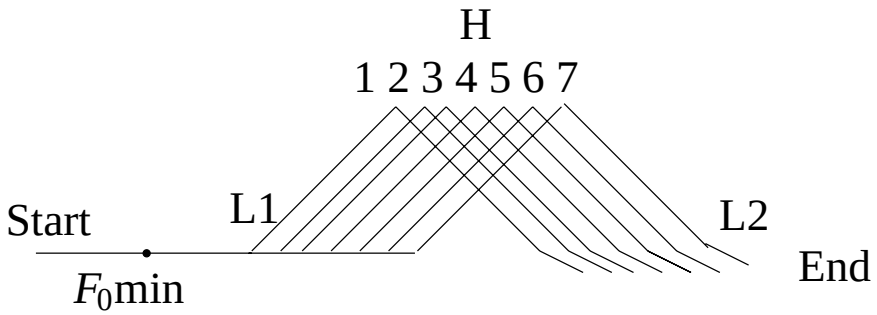

Figure 3.9: Structure of primary continuum stimuli.

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.[3]

The f_0 contour of the base utterance was extracted and its contour modified with the help of GIPOS, a speech analysis program developed at IPO, Netherlands. Such a program allows direct manipulation of the f_0 shape. Once the f_0 shape was altered with new target values (both timing values and/or f_0 values), the utterance was resynthesized (the details are presented below).

[3] As expected, most of the listeners (27 out of 30) showed between 80 and 100% question responses for this stimulus, so that only 3 listeners had to be discarded according to the mentioned criterion. An additional listener had to be discarded on the basis of a linguistic criterion (i.e., the listener was not born and raised in Naples), so that the data relative to 26 listeners will be presented here.

Note that in this experiment it was hypothesized that the duration and spectral characteristics of questions versus statement utterances would not be strong enough to bias identification. That is why I thought it would suffice to employ only a question base utterance in order to create the test stimuli, as opposed to employing both a question and a statement base. In order to assess such a claim, the listeners were presented with a resynthesized question base and a resynthesized statement base stimulus in which the fundamental frequency was flattened out over the entire utterance. The stimulus couple was presented twice, with the question preceding the statement in both repetitions for half of the stimulus pool, and with the statement preceding the question for the other half. The listeners were informed that they would listen to two synthetic speech utterances created from an original question and an original statement, but that the relative order of presentation was not known. After listening to the two repetitions, the listeners had to mark on an answer sheet whether what they heard was a sequence of a question followed by a statement or vice versa.

The pattern of responses was the following: 15 question-statement responses (55.56%) versus 12 statement-question responses (44.44%). I compared these scores to a random sampling, i.e., question = 50% and statement = 50%, by means of a χ^2 test. The results ($\chi^2 = 0.33; p = 0.56$) showed that the scores were not statistically different from a random sampling at the 0.01 level. Moreover, if the continuum were indeed biased towards question responses because of its spectral and duration properties, getting statement responses by simply varying the timing of the rise-fall configuration would be even more challenging. Therefore, any obtained effect in such a direction would prove to be quite strong.

For the stimuli preparation I employed PIOLA resynthesis (developed at IPO, Netherlands[4]), which has characteristics that are similar to PSOLA (*Pitch Synchronous Overlap and Add*, cf. Moulines and Charpentier (1990)). This resynthesis method allows for natural sounding stimuli. As shown in Figure 3.9, the prenuclear section of the contour was obtained from the interpolation of three points: 1) a starting point (start) at utterance onset; 2) a point which corresponds to the location of the f_0 minimum ($f_0$min) within the production utterances (this is the turning point at which f_0 either starts to slightly fall towards the elbow or to rise towards it) and 3) the actual elbow (L1). The region from start to $f_0$min was stylized with a straight line at 210 Hz. Such an f_0 value was chosen because it is intermediate between that of a typical question prenuclear region (which is generally around 220 Hz for this speaker) and a statement prenuclear region (which is around 200 Hz). By keeping f_0 height of start and $f_0$min fixed we ensured that the prenuclear contour would be the same for all stimuli and that it would not change shape (being, for instance, either falling, rising or flat) as a result of interpolation between a fixed f_0 height for start and a variable height for $f_0$min. By allowing the stimulus continuum to have a flat prenuclear contour, one can make observations that are independent from any effect of "head type". Additionally, instances of flat head utterances can be found in natural speech and do not sound unnatural. On the other hand, the fixed $f_0$min height was interpolated with a variable L1 height in the f_0 continuum described below, creating effects such as those depicted in Figure 3.10.

[4]Leo Vogten was the author of the algorithm used here.

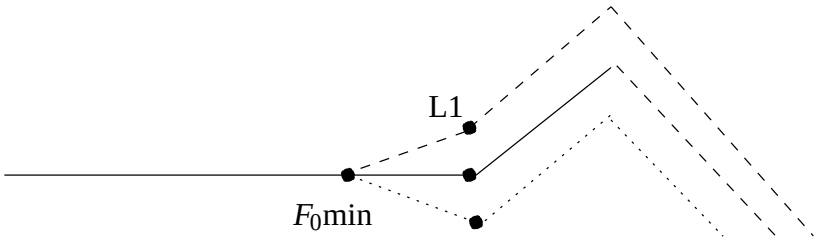

Figure 3.10: Example of $f_0$min-to-L1 interpolations.

The fundamental frequency and timing values of the primary continuum stimuli are shown in Table 3.1. 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. 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. Apart from L1, the manipulated accent consists of a peak (H) and a fall target (L2). The three target locations (L1, H and L2) were simultaneously shifted earlier, in 15 ms steps, from the original values of the natural utterance. For sake of clarity, time values for the continuum are given relative to vowel onset. Another reason for choosing such a reference location for the timing manipulations is that both statement and question peaks appear to be more consistently aligned with the left edge than with the right edge of the syllable (cf. §2.4.2). Such a generalization is tested in Chapter 5, where I investigate the effect of syllable structure (which translates into vowel duration differences) on the perceived alignment of tonal targets.

The timing values in the base stimulus were: L1 at +30 ms from vowel onset (v0), H at +150 ms and L2 at +270 ms (the vowel duration was 170 ms). Therefore, the only dimension that was varied in the primary continuum is rise-fall timing, while fundamental frequency was kept at "normal" levels for the entire accent configuration. These f_0 values will be referred to as "medium"("m") as opposed to the high ("h") and low ("l") values that will be employed to test hypotheses 3 and 4.

Note in Table 3.1, unlike the inter-peak stimuli of D'Imperio and House (1997), that the earliest H peak timing was already 60 ms into the stressed vowel. This was done to allow close investigation of the syllabic region within which statement and question peaks are measured in production. This allowed also smaller step duration than the one employed in D'Imperio and House (1997).

A straight-line was interpolated from L2 to the end of the stimulus (end), which was kept at the same f_0 value for all stimuli, i.e., 150 Hz. The L2-to-end interpolation created a slightly falling f_0, which is due to the sloping transition from the end of the phrasal HL- (i.e., L2) to the lower target for L%. Seven steps were thus obtained, resulting in seven stimulus types (see Figure 3.9).

Three secondary continua were created in order to test hypotheses 2, 3, 4 and 5. The "plateau" continuum was intended to test hypothesis 2, that is the hypothesis that the shape of the accent peak would affect the number of question responses by changing the perceived timing of the target. This continuum was obtained by creating a series of 4 stimuli whose shape was characterized by a 45 ms f_0 plateau. 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

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

offset (i.e., the plateau beginning) or HL fall onset (plateau end). Such tests were performed by comparing, first, the results for plateau stimulus 1 (in which plateau onset is at time step 1 (T1) of the primary continuum and plateau offset is at step 4 (T4) of the primary continuum) with results of peak stimulus 1 and, then (to test hypothesis 2b) with peak stimulus 4.

Rise offset in the plateau stimuli had values from step 1 to step 4 (i.e., T1, T2, T3 or T4) of the primary continuum, while fall onset could be at T4, T5, T6 or T7. The structure of the plateau stimuli is shown in Figure 3.11. Note that in each of the stimuli the beginning and end of the plateau always correspond with peak locations in the primary continuum.

Then, in order to test hypothesis 3, regarding the impact of f_0 level on the question/statement manipulation by changing target location in the melodic space, I created a continuum in which f_0 height of L1, H and L2 targets was varied between a medium value ("m"), a high value ("h") and a low value ("l"). 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

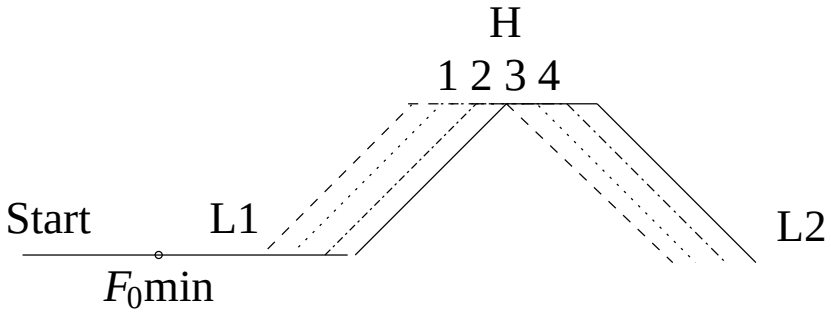

Figure 3.11: Structure of plateau continuum stimuli.

the primary continuum. I will refer to this continuum as the "f_0 continuum". The choice of manipulating f_0 only within a limited set of timing locations was motivated by avoiding a proliferation of stimuli, thus by a desire to keep the perception experiment within reasonable duration limits.

Within this continuum, f_0 targets could either be higher or lower or same as the corresponding values in the primary continuum. The combinations of values will be referred to as mmm (all medium values), hhh (all high values) and lll (all low values). Then, in order to test hypothesis 4 of an effect of slope of the rise and the fall, the lhl combination (i.e., low value for L1, high value for H, low value for L2) was compared to mmm as well as hhh. The precise f_0 values for each combination are reported in Table 3.2.

Note that the hhh combination was shifted up by 20 Hz relative to mmm stimuli, while the lll combination was shifted down by 20 Hz. Additionally, in the lhl combination, the f_0 excursion was incremented from a value of 70 Hz to 110 Hz. It is important to note here that the f_0 values employed for such combinations are all commonly found in

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

Table 3.2: f_0 values for L1, H and L2 within the f_0 continuum.

natural speech. Combinations of low/medium and high/medium f_0 height were excluded since the aim was to exaggerate the f_0 height effect in order to get a perceptual effect, without compromising stimulus naturalness. While in the production dynamics some of the intermediate combinations are indeed observable, it is not always useful to mimic all of them in a perception study because the chance of them producing perception effects could be quite low. Hence, production differences need often to be exaggerated in order to obtain such effects.

Finally, a continuum was created in order to test the impact of both L1 and L2 timing, while keeping H timing constant. Also in this continuum, which I will refer to as the "low timing continuum", L1 and L2 occurred at 3 possible timing values, that is T1, T4 and T7. At each location, the timing manipulation was compared to the corresponding primary continuum combination. For instance, at T4, the following combinations for L1 timing manipulation were compared (each number indicates the timing of L1, H and L2 in order): 4-4-4 (from the primary continuum), 1-4-4 and 7-4-4.

All the allowed timing configurations are shown in Table 3.3. Note that not all L1-H-L2 combinations could be obtained at each timing location. This was due to the fact that the existing timing values for endpoints were limited in terms of relative distance from each other. Also, perceptible effects needed to span more than one time step. This restricted the number of timing combinations and caused stimuli at certain time locations to allow for more combinations than others.

As already noticed above, varying the timing of L1 and L2 independently of H creates most obviously some slope effects. For instance, stimulus 1-7-7 has a shallower rise than stimulus 7-7-7, since the location of L1 is relatively earlier in stimulus 1-7-7. This indicates that the effects of L1 timing and slope of the rise cannot be completely separated. However, if the null hypothesis 4 is supported by the data, that is if no slope effect is found, then we can safely conclude that any effect found in this continuum will be due to L1 or L2 timing manipulation alone.

Some additional stimuli were included in order to test the effect of overall shape of the pitch accent configuration on the question/statement response, which are not directly relevant for the hypotheses formulated above. However, the results for these stimuli will be shown in order to strengthen the argument that the shape of the accent and not just the timing of the peak can affect the perceived target mapping. Most of these stimuli (apart from hlh) result in a triangular shape which, unlike all other stimuli, is highly non-symmetric, and is the result of f_0 manipulations either at the peak or at one of the L targets. Those combinations are shown in Table 3.4. This set will be referred to as "shape set". Through the manipulations described above, 45 stimulus types were obtained. The natural question base stimulus was added to the set, for a total of 46 stimuli.

L1 timing	H timing	L2 timing
1	1	1
4	1	1
1	1	4
1	1	7
4	4	4
1	4	4
7	4	4
4	4	7
4	4	1
7	7	7
1	7	7
4	7	7
7	7	4

Table 3.3: Timing combinations for the low timing continuum.

	L1 f_0(Hz)	**H f_0(Hz)**	**L2 f_0(Hz)**
hll	230	260	160
llh	190	260	200
lhh	190	300	200
hlh	230	260	200

Table 3.4: f_0 values for L1, H and L2 for the "shape" continuum.

3.2.2 Experimental Procedure

Thirty listeners from the University "Federico II" of Naples, Italy, participated in the experiment. The listeners, who were paid for their participation, were all brought up in Naples and spoke Standard Italian with a Neapolitan accent. The stimulus list was recorded on tape. In order to ensure that there would be no order effects, two tapes were created by reversing the order of the blocks in the first list. The stimuli were all separated by a three-second pause. To help the listeners follow the order of presentation, a tone was played at the end of each block.

The stimuli were presented binaurally through headphones in a studio at the University "Federico II" of Naples. The listeners 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.

3.3 Results

3.3.1 Primary continuum

The results for the primary continuum are shown in Figure 3.12. The scores were calculated as mean values for each listener, specifically by counting how many times the listener identified a stimulus type as a question across its 5 repetitions. The result was then divided by the number of repetitions. This yielded results between 0 and 1. Figure 3.12 shows mean question scores pooled for all listeners (y axis) across stimulus Time Step (x axis). Here, stimulus 1 has the earliest timing (T1) while stimulus 7 has the latest (T7).

As expected from the results for the "inter-peak" continuum reported in D'Imperio and House (1997), shifting the L1-H-L2 configuration earlier within the accented vowel decreased the number of question responses. Specifically, we see an orderly decrease from question to statement judgments as a function of rise-fall timing from the end of the continuum to the beginning of the continuum. As Figure 3.12 shows, at early locations within the timing continuum, such as at T1 and T2, statement responses were largely dominant

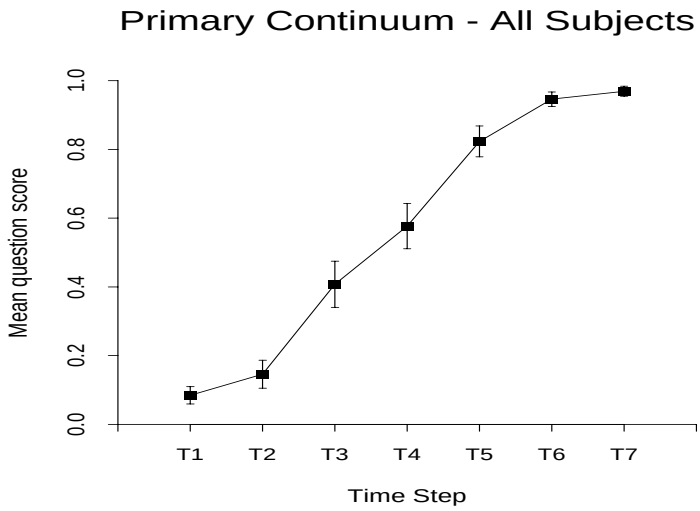

Figure 3.12: Primary continuum mean scores for all listeners. Standard error is indicated by vertical bars.

(with a mean question score of 0.08 for T1 and of 0.15 for T2), while at the opposite end of the continuum question responses were largely dominant (with a mean score of 0.95 for T6 and 0.97 for T7). Since the variance was not constant throughout the timing continuum (it was in fact greater at the center of the continuum and smaller at its edges, as the standard error bars in Figure 3.12 show), performing an Analysis of Variance on the raw data could yield erroneous results. This is because one of the assumptions of the Analysis of Variance is that the variance of the sample scores is homogeneous. If such an assumption is violated, a nonlinear transform of the data can make the variance more homogeneous across factor levels. An arcsine transform is sometimes suggested in these cases (cf. Woods et al.

(1986)).[5] The data thus transformed were used as the basis for a one-way Analysis of Variance[6] with Timing as a factor. The timing manipulation yielded a significant result, as expected [F(6,175) = 61.11; $p < 0.01$].

It must be noted that the response curve in Figure 3.12 does not superficially appear to be as steep as the one obtained for "rise-peak" stimuli in D'Imperio and House (1997) (see §1.5.1). This can be explained with the smaller step size duration employed here, which was only 15 ms, while it was greater (35 ms) in the previous study. The other difference with the earlier study is that, here, peak timing was shifted starting only at 60 ms into the stressed vowel, as mentioned above, and not immediately from vowel onset (which was the procedure adopted in the earlier study). Thus, the rise-fall configuration was shifted only within the region including those values for statement and question peak timing actually found in production. Earlier peak values, from vowel onset up to 60 ms into the vowel, appear to be associated with broad focus statement peak accent location (H+L* accent). Therefore, including such values might have created noise in the response curve.

Additionally, unlike the results of D'Imperio and House (1997) (see §1.5.1), the category shift occurs early, i.e., between T3 and T4, with the score for the center stimulus (T4) being already above chance. That is, given the characteristics of the continuum (which

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

[6]Note that, though the use of the logit model might be suggested in the case of categorical data, a classical Analysis of Variance can be used when a large number of samples is available for each combination of factor values. This is because the "Central Limit Theorem" implies that the distributions of the means are quasi Gaussian.

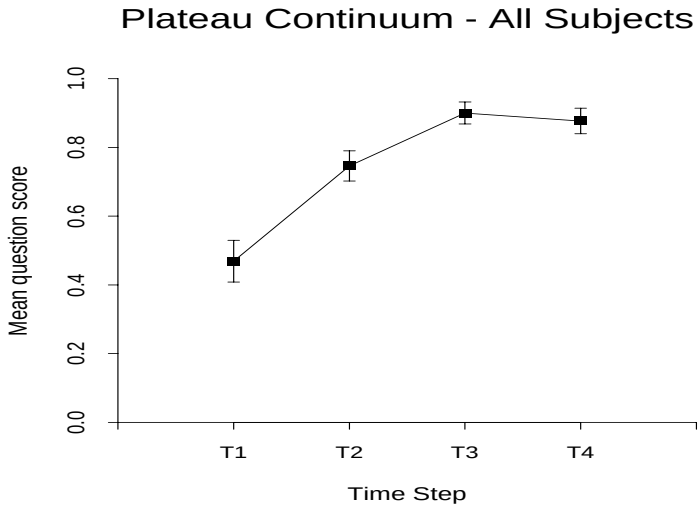

Figure 3.13: Mean scores for the plateau continuum (pooled results). Standard error is indicated by vertical bars.

covers the later section of the accented vowel), a majority of question responses were already obtained at relatively early timing locations. The center stimulus in D'Imperio and House (1997) was highly ambiguous, instead.

3.3.2 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.

Figure 3.13 presents mean question results for the plateau continuum. These results appear immediately strikingly different from those of the primary continuum (Figure 3.12). 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). Also, notice that the highest data variability (represented by standard error bars here) is obtained at the beginning and not at the center of the continuum (see Figure 3.13).

In order to test hypothesis 2a, stating that T1 peak stimuli would score the same percentage of question responses as T1 plateau stimuli, I 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 me reject hypothesis 2a.

I then tested hypothesis 2b, i.e., that peak stimuli with peaks timed at the plateau offset would receive the same question scores as plateau stimuli. In order to facilitate a direct comparison, I included scores for both sets in Table 3.5. Interestingly, by simply observing the score for plateau stimulus 1 and comparing it with peak stimulus at T4 (which is the stimulus corresponding to the timing of the plateau offset for plateau stimulus 1), note that they are quite similar (the result for the plateau stimulus is only somewhat smaller).

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

In addition, the variance of the data at plateau stimulus 1 is very similar in magnitude to that of peak stimulus 4 (see error bars in Figure 3.13). Analogously, the score for all other stimuli in the plateau continuum resembles the score of peak stimuli with timing corresponding to the end of the stimulus plateau (see Figure 3.13).

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 the plateau scores and the scores of peak stimuli timed at plateau offset (though those scores were different from a statistical point

of view), which can be interpreted in terms of closer timing of perceived plateau target to HL fall onset (i.e., the end of the plateau). Hence, in Chapter 5, results for a larger set of plateau stimuli will be shown in order to further test hypothesis 2b.

The next question is if plateau stimuli are perceived as simply having a higher pitch than the corresponding peak stimuli. This will be tested in the next section.

3.3.3 f_0 continuum

Results for the f_0 continuum stimuli are presented here. This continuum was designed to test hypothesis 3, by which the overall pitch level difference within the LHL configuration would displace target location in the melodic domain and, as a consequence, affect the question/statement identification scores. In Figure 3.14 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 f_0-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. Remember 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 (MD) range within the accent

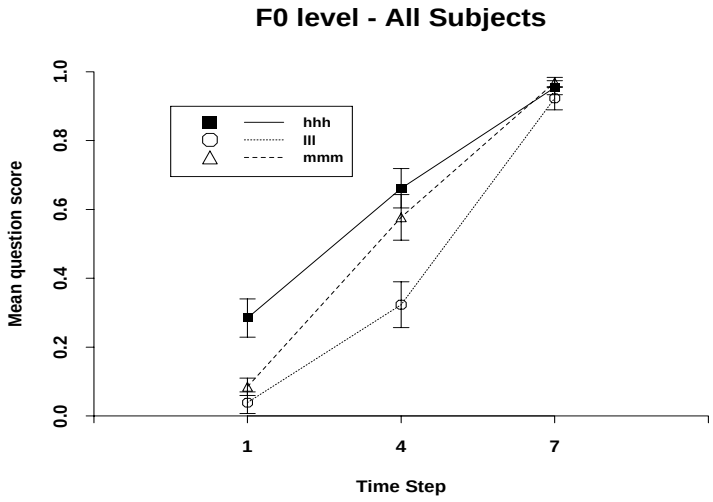

Figure 3.14: Mean scores for the f_0 continuum (pooled results). Standard error is indicated by vertical bars.

(a 99 Hz difference was found, on average, between L1 and H f_0, as one can see in the lower panel of Figure 2.10). Hence, we must notice that such a sensible difference does not produce a score modification that is as sensible as that produced by timing manipulations, especially at T1 and T7. In fact, the timing manipulation alone (from T1 to T7) changes the score dramatically. Therefore, one can tentatively conclude that the f_0 level effect is relatively small and that bigger differences seem to be needed in order to show such an effect.

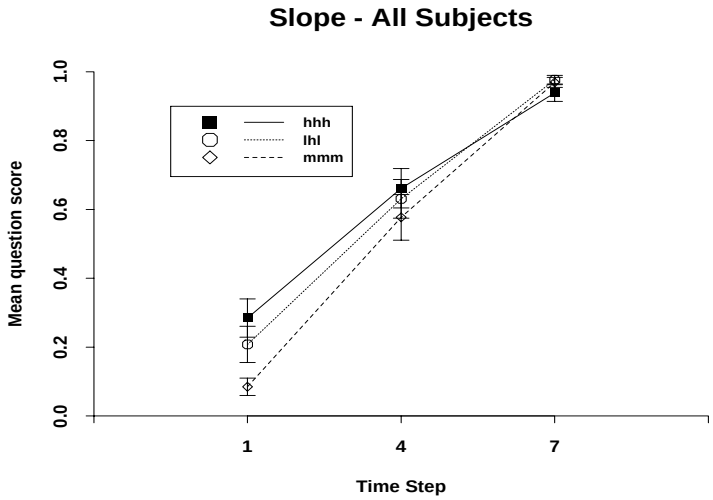

Figure 3.15: Mean scores for the slope continuum (pooled results). Standard error is indicated by vertical bars.

Consequently, for this work it is safe to assume that the f_0 differential will play only a minor role in the question/statement contrast. Moreover, since stimuli with peaks at either a normal (mmm) or high (hhh) f_0 value were the most similar, they will be employed in order to test the hypothesis of an effect of rise and fall slope (hypothesis 4).

3.3.4 Slope continuum

Figure 3.15 shows mean scores for the mmm and lhl stimuli, in which the slope of the rise and fall are different. The hhh stimulus combination was added in order to test if the f_0 height of L1 and L2 might cause an additional effect. The data are again arranged by peak timing location.

Note from Figure 3.15 that the scores are remarkably similar. A two-way ANOVA, with Stimulus Type and Time Step as factors, revealed only a significant main effect of Time Step [$F(2,234) = 186.2; p < 0.01$], and no effect of Stimulus Type [$F(2,234) = 1.23; p = 0.3$] nor any interaction [$F(4,234) = 0.73; p = 0.6$]. The hypothesis of an effect of slope on the results (hypothesis 4) can therefore be rejected.

3.3.5 L1 and L2 timing continua

Remember that the null hypothesis 5a claimed that stimuli with the same H timing and different L1 timing would obtain the same percentage of question responses. Analogously, the null hypothesis 5b claimed that stimuli with the same H timing but different L2 timing would score the same. Table 3.6 shows mean scores for each of the L1 and L2 timing manipulation, separately for each H target timing.

Separate one-way Analyses of Variance to assess the effect of L1 at each H timing location were performed on the arcsine data, with Stimulus Type as factor, as well as Tukey post-hoc analyses (confidence interval = 0.01). Remember that not all L1 timing combinations were obtained at each of the 3 timing locations (T1, T4 and T7) because of intrinsic limitations of the chosen timing continuum.

L1 timing	**H timing**	**L2 timing**	Mean Q Score	SD
1	1	1	0.08	0.13
4	1	1	0.09	0.17
1	1	4	0.16	0.16
1	1	7	0.33	0.25
4	4	4	0.58	0.34
1	4	4	0.41	0.33
7	4	4	0.73	0.23
4	4	7	0.80	0.20
4	4	1	0.23	0.25
7	7	7	0.97	0.07
1	7	7	0.91	0.17
4	7	7	0.91	0.15
7	7	4	0.90	0.16

Table 3.6: Mean question (Q) scores and standard deviation (SD) for all the L1 and L2 timing manipulations.

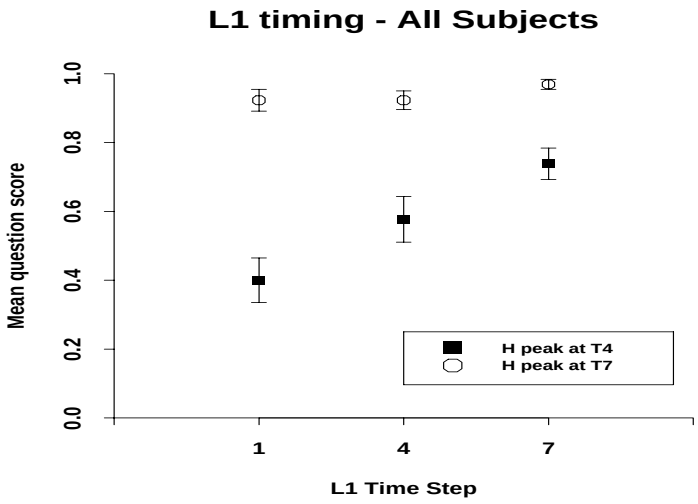

Figure 3.16: Mean scores for L1 timing manipulation at H timing 4 and 7 (pooled results). Standard error is indicated by vertical bars.

As it can already be inferred from the raw results, the L1 timing manipulation was not significant at H timing T1 [$F(1,50) = .002; p = 0.96$], nor it was at T7 [$F(2,75) = .82; p = 0.44$]. The manipulation was instead significant at T4 [$F(2,75) = 5.02; p < 0.01$]. However, at T4, the only comparison that produced a significant difference according to a post-hoc test was that between 1-4-4 and 7-4-4. Essentially, an earlier L1 clearly produces fewer question responses and a later L1 produces more question responses only when the peak is at T4. This effect can be clearly seen in Figure 3.16, where only the results for the T4 and T7 continua are shown.

Then, in order to test hypothesis 5b, regarding an effect of L2 timing manipulation, I performed another set of ANOVAs at separated H timing locations. The overall trend

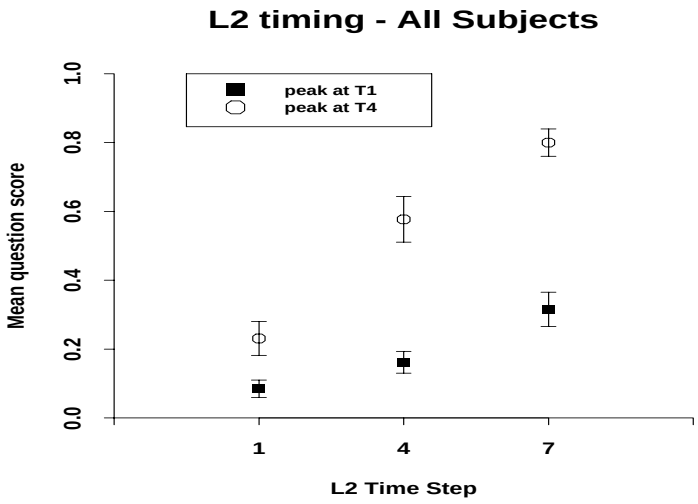

Figure 3.17: Mean scores for L2 timing manipulation at H timing 1 and 4 (pooled results). Standard error is indicated by vertical bars.

is, then, that earlier L2 produces fewer question responses and later L2 produces more question responses, as it can be also seen in Figure 3.17. Here we can directly compare the direction of the effect within the T4 and T7 continua. Unlike the L1 timing manipulation, the analyses revealed a significant effect already at H timing T1 [$F(2,75) = 8.9; p < 0.01$]. Here, a post-hoc test revealed a significant difference only between the 1-1-1/1-1-7 stimulus pair. The manipulation was also significant at T4 [$F(2,75) = 20.7; p < 0.01$]. In this manipulation, stimulus 4-4-1 obtained significantly different results from both 4-4-4 and 4-4-7, according to a Tukey test. No noticeable effect was produced by the L2 manipulation [$F(1,50) = 2.18; p = 0.14$] at the opposite end of the continuum, i.e., at T7.

Therefore the hypothesis that stimuli with varying L1 and L2 timing would score same percentages of question responses can be rejected, though the magnitude of the effect appears to be different according to continuum location and to whether the low precedes or follows the H peak.

3.3.6 Additional stimuli

Regarding some of the shape effects induced by f_0 height manipulation, the scores of an additional set of stimuli was analyzed. In most of these stimuli, the shape of the contour is highly non-symmetric (apart from stimulus hlh). The shape of those configurations are schematized in Figure 3.18, in which each "triangle" represents a different contour. The timing for all three targets was the same, and was again shifted in three crucial locations of the continuum, i.e., T1, T4 and T7. The structure of the stimuli in the primary continuum (mmm) is added for ease of comparison.

Scores for the shape set are shown in Table 3.7. The scores are horizontally ordered by timing location (T1, T4 and T7) of the manipulation, and vertically according to score magnitude at T4. Specifically, higher scores at T4 correspond to higher row entries. The manipulation was also coded in terms of f_0 height "features", so that each target (L1, H and L2) was coded as being either "-h" or "+h" (second column from the left). Primary continuum stimuli scores (mmm) are added for comparison.

First, note that the scores are not the same, despite the identical timing of the stimuli. It might simply be concluded from observing the table that stimuli with [+high] L1 are perceived as having early target timing, and therefore cuing statements, while stimuli with [+low] L1 target are perceived as having a later target, therefore cuing questions. Or, more

Stimuli	Target tone value	Score at T1	Score at T4	Score at T7
lhh	- + +	0.44	0.89	0.99
llh	- - +	0.23	0.81	0.98
mmm	- - -	0.08	0.58	0.97
hlh	+ - +	0.09	0.30	0.78
hll	+ - -	0.04	0.05	0.65

Table 3.7: Means for the shape set scores (pooled results). Scores with decreasing value at T4 are shown from higher to lower row entries. Primary continuum scores (mmm) are included for comparison.

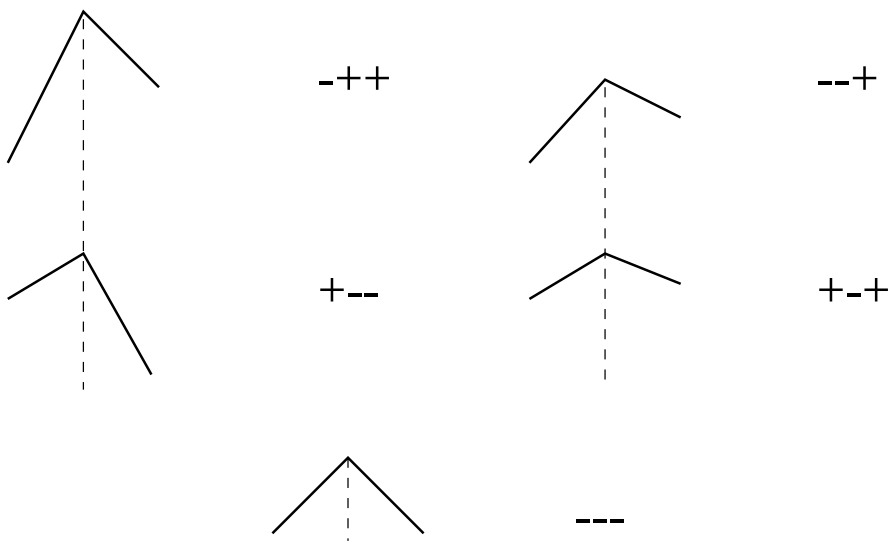

Figure 3.18: Schematic representation of the configurations in the shape set.

simply, that high L1 values are associated with statements (as we saw in Chapter 2) and low L2 values are associated with questions. But the situation is not so simple, given the fact that, as we saw in §3.3.4 above, scores for hhh and lhl stimuli were the same. Also, if from the other three stimuli we are erroneously led to conclude that [+high] for L2 is translated with a later timing (and therefore more question responses) by the listener, the same cannot be said of stimulus hlh, which scores significantly lower than lhh and llh. Hence, an alternative explanation must be sought.

Note that the orderly progression at T4 is (for the most part) reflected by the scores at T1 and T7. This suggests that the main determinant of the ordering applies at all timing locations within the vowel. As to what determines the pattern of responses, I will propose that is a complex process of averaging at both rising and falling contour f_0 values. Take for instance stimulus lhh. Imagine that the listener operates such averaging in order to obtain a specific f_0 value at either the first half (the section *before* the vertical dashed line in Figure 3.18) or the second half (the section *after* the vertical dashed line in Figure 3.18) of the stressed vowel. By this token, the average value for the rise in stimulus lhh is quite low, since the low value of L1 pulls the average down, while the average value for the fall is high. This might be mapped as simply as: [+low] in the first temporal unit and [+high] in a second unit. The same mapping is obtained for llh, and this is confirmed by the results. Instead, hll would get the opposite mapping, and it should result in a different pattern of results, which is what I found. The stimulus hlh is expectedly more ambiguous, scoring halfway between the other stimuli, since it will obtain [-high]/[-low] for both rise and fall averaging.

Of course, as mentioned already above, this is a very simplified view of the mapping of shape onto target localization. Simply, the results presented above lead me to propose that the whole rise-fall contour is abstractly translated to one "gestalt" value, that is the result of a weighed averaging process. This proposal will not be explored further in this thesis, but will be object of future research.

3.4 Discussion

The perception experiment presented in this chapter aimed at testing some hypotheses about tonal target perception in Neapolitan Italian, by assuming that the question/statement opposition is mapped onto a contrast in perceived target location in a bidimensional space. The dimensions of the space are the f_0 and the timing of the inflection points.

First, I 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. I 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.

I 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, I 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. I 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, I 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). This might also mean that the target H location for the phrasal HL- is perceptually more important than the H target of the LH pitch accent.

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. Therefore, we would expect that just by shifting the overall f_0 values of the LHL configuration we would obtain different scores. However, f_0 had relatively little effect when stimuli with peaks (and lows) characterized by a 20 Hz difference were compared. Especially, what I 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.

In the stimuli resynthesized here the baseline values were left unchanged throughout the stimuli continua since I wanted to avoid effects of "head" type, etc. Hence, this means that the difference in height between the baseline and the "l" peaks (equal to 50 Hz) might have been too small in order to map the perceived target onto the category "question". In

other words, it might be the case that accent configurations whose values are underneath a certain threshold (such as the lll stimuli) are simply categorized differently from all other accent configurations above a certain abstract reference line. Of course, I cannot estimate the exact f_0 level of such a threshold with the results of this study, but it will be interesting to test such a hypothesis with further investigations. The results for stimuli with different shape of the contour seemed in fact to point to an "averaging" of the entire rise and falling section of the contour, so that a [+high] or [+low] kind of feature value are assigned to a specific temporal unit.

When manipulating LH rise and HL fall slope by changing both H target f_0 value and L1/L2 f_0 value, no effect was found (see §3.3.4). Since there was only a small effect of peak height, I believe it is justified to conclude that slope does not have any effect in the perception of the contrast, and that it does not displace the perceived alignment of the target.

It is possible that with a more drastic manipulation of f_0 height within the H target we could have obtained the kind of results discussed by Nábeläk et al. (1970). That is, when the tonal transitions were longer, the authors found that the pitch of the perceived target tended to approximate the end of the transition. But while such dramatic manipulations are allowed psychoacoustically, since the tonal transitions have no linguistic function, they are not feasible in real speech. Target f_0 values do change as a consequence of emphasis, for instance, but the observed effects are usually within a few semitones. It is unlikely, therefore, that by increasing the peak value above 300 Hz (which is the "h" value employed here), we would have obtained a different result. Also, above that threshold, the voice of

the speaker (at least for Neapolitans)[7] tends to sound unnaturally high. Therefore slope effects can be discarded from a model of perceived target location, at least for Neapolitan Italian.

It is still unclear whether the invariant features of pitch accent production are the temporal location of its tonal targets or more holistic, dynamic features, such as overall shape, transition slope and/or transition duration. Recently, Ladd et al. (1999) have shown evidence suggesting that speakers might keep tonal targets invariantly aligned with specific locations in the segmental string. By such a view, actual pitch accent shape and duration are not independently controlled, hence are not linguistically significant. The literature on tonal perception presents more controversial evidence regarding the role of contour slope and shape in the identification of linguistic categories. D'Imperio and House (1997) speculate that the different perception results for two of their stimulus series, which could not be explained in terms of peak timing difference, might have been caused by a difference in rise slope characterizing the stimuli.

Earlier, Bruce (1977, pp. 115-116) argued that the seeming effect of slope in his perception experiments on Accent I and II of Swedish was only an epiphenomenon and proposed that what listeners attend to is the temporal location of the tonal excursion midpoint ("the point in time at which half of the range of the fall has been covered"). Interestingly, if I compare stimuli having the same midpoint within either the fall or the rise transition (keeping everything else the same), I find also a very similar score.

[7] Many Japanese women, for instance, can raise their f_0 above 600 Hz and still sound natural (Mary Beckman, p.c.).

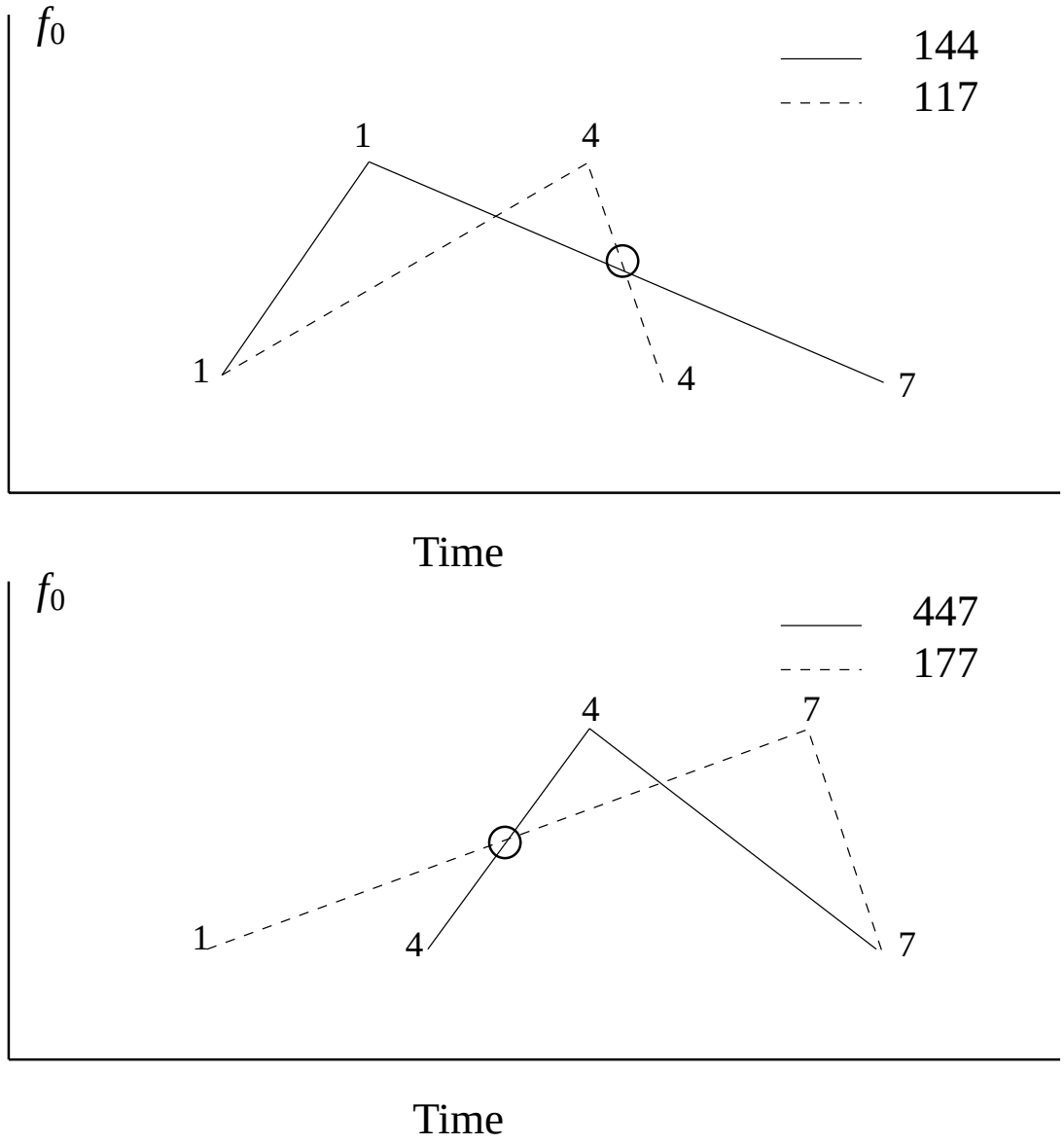

Figure 3.19: Schematic representation of stimuli 144 and 117 (upper) and 177 447 (lower) and, respectively, the midpoint of the fall and of the rise.

Given the characteristics of the continuum employed for the L1/L2 timing manipulation, the only pairs I can compare are 1-4-4 and 1-1-7 for the midpoint of the fall and 1-7-7 and 4-4-7 for the midpoint of the rise. The relationship between these stimuli in shown in Figure 3.19. The results were very similar for both pairs, i.e., 0.41 and 0.33 respectively for 1-4-4 and 1-1-7, and 0.91 and 0.80 for 1-7-7 and 4-7-7. The problem with this kind of evaluation for a rise-fall, though, is that when midpoint of the rise stays constant, the midpoint of the fall does not (and fall gradient is different) and vice versa. Hence, before suggesting that such an effect does or does not exist, one needs to perform a sensitive experiment in order to avoid the danger of missing a regularity because of the paucity of stimuli to be compared.

Unlike Bruce (1977), and unlike what was supported by this study, House (1995) reports on the relevance of fall gradient for Thai. In this work, the listeners seemed to employ slope gradient in a tone matching task, but only for stimuli in which the relevant contour was not followed by a pause. Specifically, House designed a perception experiment consisting of matching two tones with either different end frequency, but same gradient, or with same gradient but different end frequency. When the fall was followed by a pause, listeners tended to match end frequency, while when there was no pause (or a shorter pause) listeners tended to match fall gradient. However, such results could also be interpreted in terms of different timing of the L target for the fall. In House's stimuli, in fact, the peak location was left unvaried, while the timing of the end of the fall was varied to obtain slope effects. It maybe that the timing of the end of the fall (analogously to L2 in my experiment) is more salient when it is followed by a pause than when it is not, but this is only speculative.

The L1/L2 manipulation presented in my experiment served the purpose of testing the relevance of timing effects that depend only on some and not all targets within the LHL configuration. In a sense, the results of this manipulation can be interpreted, similarly to Bruce's position, by claiming that what matters to the listeners' ears is not the slope of a rising or falling contour but the timing of the target L1 or L2. When hypotheses 5a and 5b were tested, in fact, I could conclude that there is an effect of L timing location, though the effect is not generalizable to all locations within the timing continuum. Hence, peak timing is not the only factor determining the pattern of responses.

Interestingly, the effect of the L target manipulation was asymmetric, in that L2 manipulation was significant at two timing locations for H, i.e., 1 and 4, while L1 manipulation was significant only at T4. This result can be taken as supporting evidence for the different phonological analysis of the LH (pitch accent) rise and the phrasal HL fall. The novel result is that the contribution of the falling section of the rise-fall configuration appears to be more salient than the rising portion. Asymmetries between rise and fall perception have also been found (regarding "accentuation boundaries") for Dutch, but the rises and falls employed there were both "accent lending" movements, and none was analyzed as an edge phenomenon (Hermes 1997).

Alternatively, it might also be that the time difference between T1 and T4 for L1 does not span a linguistically significant boundary, while the time difference between T1 and T4 for L2 does. In fact, in chapter 2 we saw that L1 tends to be timed roughly with syllable onset for statements and with vowel onset for questions. Since the L1 timing continuum started only at -60 ms from vowel onset, once we reach T4 we are still in the area that is more consistent with L1 timing for statements (-15 ms) than for questions (at least for speaker MD, who produced also the resynthesized stimuli employed for the perception experiments).

Such an asymmetry, therefore, might not be merely psychoacoustic in nature. Since both L1 and L2 are located within a consonant (onset for L1 and coda for L2), we cannot attribute the effect to an asymmetrical memory load and processing of f_0 characteristics, which would be in line with House's "Spectral Constraint Hypothesis" (House 1990). In fact, a displacement in time of a L on one side of the continuum should have the same effect as the displacement of a L target on the other side, since both are located in regions of "spectral instability". Hence, if the effect is linguistic in nature, and particularly related to Neapolitan Italian regularities, we should not necessarily find it in other language groups. This will be explored in Chapter 4.

The one clear effect that was found is the one differentiating peak stimuli from plateau stimuli. Such a difference cannot be attributed to a mere difference in f_0 level of the "perceived" target in the two classes of stimuli. This does not necessarily mean, however, that there is not a melodic representation of tonal target height altogether. In fact, the results suggest that a complex computation of f_0 height is performed by the listener, in which an averaging of the f_0 values throughout the entire rising and falling transition is performed. Furthermore, the psychoacoustic literature appears to support the averaging effect.

In sum, my results confirm the previous findings that listeners can employ shifts in LHL timing to recover a specific linguistic category. What is proposed here is that an "invariant" target location for the pitch configuration is first recovered, which is then mapped to the linguistic category. The two mappings do not necessarily need to be performed sequentially, and in fact they probably are not. I believe that the perceptual target is a different "face" of the same entity to which the linguistic contrast is related. The target location depends on both f_0 and timing coordinates. However, timing appears to be more relevant in the final computation. From the results of the shape set, we saw that f_0 values seem to "pull" the target to earlier or later timing locations within the stressed syllable, and, consequently, producing different scores of question responses.

I have also proposed that the role of f_0 height is dependent on the magnitude of the L1-to-H excursion, or on the difference between the f_0 value at the rise-fall onset and the baseline value. In any case, the nature of the computation appears to be very complex. Future research will aim at uncovering the mechanisms associated to this phenomenon.

CHAPTER 4

PERCEPTION EXPERIMENT II: PERCEPTUAL "TARGETS" AND PITCH ACCENT IDENTIFICATION IN AMERICAN ENGLISH

4.1 Rationale

One of the assumptions of this dissertation is that intonation is part of the phonological knowledge of native speakers. Much in the same way as American English (as well as Italian) speakers distinguish /l/ and /r/, while Japanese do not, we can talk about tonal contrasts that are either present or absent in a language. For instance, the phonological system of language A might present an opposition between a H and an L tone, while language B might not. Furthermore, the details of the phonetic specification of an "intonational phoneme" might differ from language to language. So, the peak of a LH rising tone might be timed to occur at stressed vowel offset in language A, while being timed at vowel onset in language B. We might assume that this knowledge is part of the representation for implementing a phonological contrast in the native language.

An example of language-specific knowledge is represented by details of temporal alignment relative to starred tones, i.e., those tones that (as mentioned in §1.3 above), according to the autosegmental-metrical framework of intonation, are associated with the stressed

syllable. As already discussed in §1.3, different languages appear to implement the exact alignment of starred tones in quite different ways. This linguistic knowledge can be assumed to influence the perception of tonal contrasts. In other words, just as it has been assumed that the familiarity with a certain inventory of phonetic segments might affect segmental perception, we can assume that the tonal properties of a language might influence tonal perception.

An enormous literature on cross-linguistic speech perception exists for segmentals, such as consonants and vowels (see Strange (1995) for a review of the topic). Recent cross-linguistic studies on the perception of intonation have been mainly restricted to investigating the notion of "foreign accent" (Bond, Stockmal, and Muljani 1998) and on the interpretation of semantic differences. Among earlier studies, some have tackled the issue of how speakers of different languages perceive intonation detail, both from a linguistic and a psychoacoustic point of view (cf. Hadding-Koch and Studdert-Kennedy (1976)). Such studies have employed synthetic stimuli, since they allow for detailed control and independent manipulation of the acoustic characteristics of the signal.

Trubetzkoy had already proposed that the native phonological system operates similarly to a filter when it comes to perceiving non-native contrasts (Trubetzkoy 1939). A linguistic "transfer" from native to non-native perception has then been assumed by much recent cross-linguistic research. Furthermore, constraints on auditory perception have been claimed to be at the origin of the distinctive features that are commonly employed by the languages of the world (Stevens 1972).

Therefore, our capacity to perceive speech signals might be a result of both genetic factors and our experience with the language we speak or with any of the languages we

have learned and used in our life. The effect of linguistic experience appears, though, to be highly dependent on the phonetic distinction being made, on task characteristics, and individual differences (Beddor and Gottfried 1995).

The assumption I make here is that tonal alignment is a phenomenon whose nature is similar to that of those segmental contrasts requiring close coordination of the laryngeal and the supralaryngeal systems. Such a similar case is exemplified by Voice Onset Time (VOT). Early studies (see Strange (1995) for a review) found that a VOT continuum from -150 to +150 ms is divided into either two or three categories by native speakers of different languages. For instance, English speakers divide the continuum as voiced vs. voiceless with a boundary between +20 and +40 ms, while Spanish speakers divide the same continuum at a much earlier location, i.e., -5 VOT. In other words, only stimuli from -150 to -5 ms VOT (prevoiced stimuli) are perceived as voiced by Spanish listeners, while English listeners continue to perceive voiced stops until at least 25 ms later in the continuum. Speakers of Thai present yet another category, since for them both aspiration and voicing are distinctive. Therefore, they present two boundaries within the continuum (at -20 and +40 VOT).

It is plausible to assume that temporal alignment of tones is governed by similar regularities. That is, we can assume that within a range of tonal alignment values for a certain tonal entity (either monotonal, as a H level, or bitonal, as a LH rise), the languages of the world select some specific value. In our case, the alignment of the LH tone is contrastive in Neapolitan, since both L and H tones are associated with the same syllable for both L+H* and L*+H. American English shares with Neapolitan the contrastive alignment opposition

between L+H* and L*+H. However, there are some crucial differences between Neapolitan and American English. First, the target alignment difference appears to be more subtle in Neapolitan than in American, since peak targets for questions are aligned, on average, only 60 ms later than statement peaks, as we saw in Chapter 2 above.[1].

The only study on the L+H* vs. L*+H contrast in American English that has details about peak alignment (though the methodology of that study is different form the one employed here, and results might not be readily comparable) is Pierrehumbert and Steele (1987)[2]. In that study, the results relative to the production of the L*+H/L+H* contrast for fours speakers of American English are reported. From the data shown, speakers present an alignment contrast in the order of 120 ms, which is the double of the difference presented by Neapolitan Italian speakers.[3] Also, it appears that the peak targets for both pitch accents are reached later relative to the stressed syllable than in Neapolitan Italian.

As mentioned also in §1.3, for Pierrehumbert and Beckman (1988), "association lines represent temporal overlap between substantive elements and structural positions" (p. 153), where the substantive elements in question can be starred tones. On the other hand, the exact location of the unstarred tone in a bitonal accent appears to be dependent on factors such as rate and and intrinsic length of the segments of which the stressed syllable is made (Pierrehumbert and Steele (1989) share such a view). This causes, for instance, the H

[1]Precisely, speaker MD presented a 37 ms difference, while LD presented a difference of 81 ms.

[2]This study is more extensively discussed in §1.5

[3]The exact alignment contrasts are not reported in the paper, but can be derived from the peaks in the bimodal distributions shown. The measures appear to be, for L+H* and L*+H respectively: 140 vs. 240 ms for TWB; 180 vs. 320 ms. for SAS; 180 vs. 340 ms for HDT; 140 vs. 220 ms for RLB.

tone of the American L*+H to be realized outside of the stressed syllable when the syllable is short, as in the example shown in Figure 4.1. Here, the utterance *Rigamarole is monomorphemic* (with a L*+H on *rigamarole*), produced by an American English speaker, is shown.[4] As you can see, the peak for the L*+H pitch accent is realized well beyond the stressed syllable boundary (specifically, two syllables away). In fact, in order to describe L*+H in American English, Beckman and Ayers Elam (1994, p. 23) say:

> *Because it is the L target in the "scooped" accent that is associated to the stressed syllable, and not the H,* **the high pitch target is specified only as occurring somewhat later than the L***, and the timing of the peak f0 relative to the segments is not controlled.* [emphasis is mine]

.

In Neapolitan Italian it appears that the latency of the H from L in a LH bitonal accent does not vary so extremely as a result of stressed vowel duration. In nuclear position, for instance, the peak of the HL rise is always reached very closely to the offset of the stressed vowel. The upper panel of Figure 4.2 shows the question utterance *Firmiamolo?* "Should we sign it?", where the stressed syllable [mja] is antepenultimate in the word, thus short (D'Imperio and Rosenthall 1999). The question pitch accent L*+H (which in the Figure is split in a L* and a H target for ease of comparison) has its H target realized only slightly beyond the stressed syllable boundary, i.e., within the following consonant [m], though more segmental material is available further along. This, however, might be an outcome

[4]This example was taken from the EToBI (English ToBI) instruction materials available at http://ling.ohio-state.edu/Phonetics/EToBI/homepage.html.

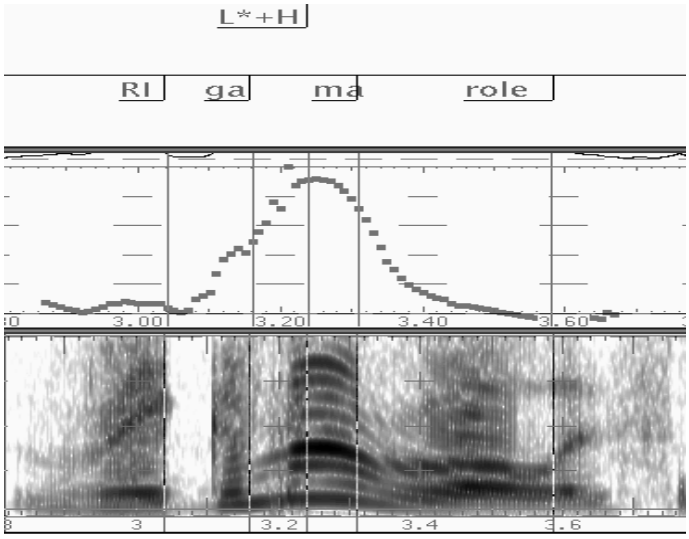

Figure 4.1: Tone labels, f_0 curve and spectrogram for a segment of the English utterance *Rigamarole is monomorphemic*.

of the overlap with the phrasal HL-. When the same pitch accent is in prenuclear position in a long focus constituent, the H peak is realized later, since the HL- phrasal event occurs at the last stressed syllable within the constituent. Even so, the H target location follows the L not more than one syllable away from the stressed syllable, as it can be seen in the lower panel of Figure 4.2. Here a fragment for the utterance *Vedrai [la bella mano di mammola] domani?* "Will you see Mammola's hand tomorrow" (square brackets indicate scope of focus) is shown. Specifically, the word *bella* is depicted here, in order to show the alignment of the prenuclear L*+H with the stressed syllable [bEl]. Note that the H occurs only at the end of the postaccentual syllable.

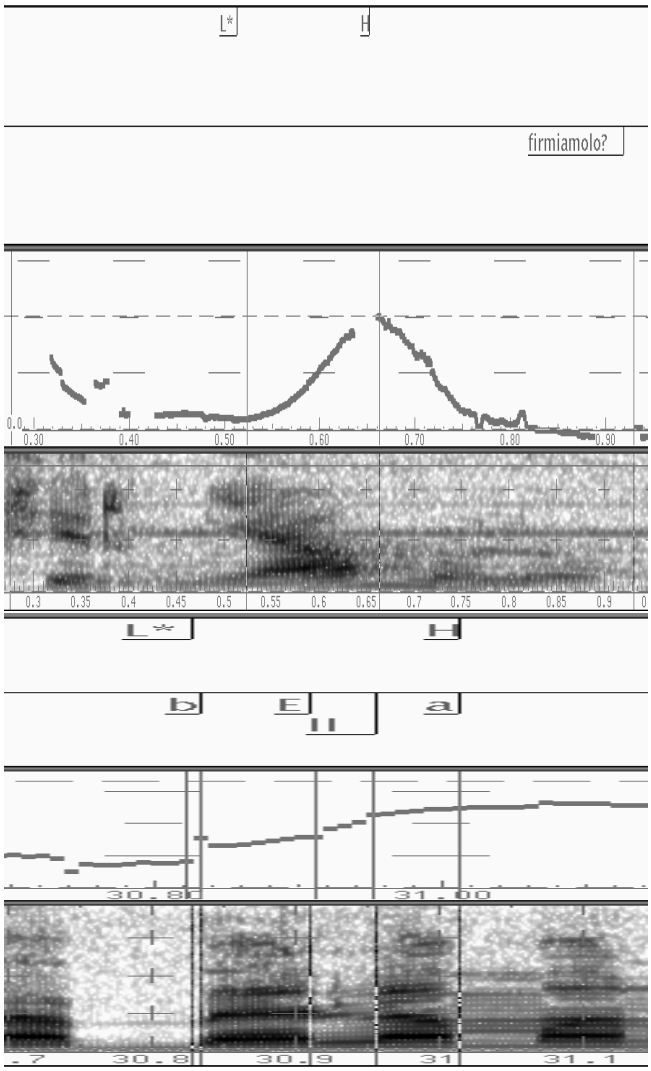

Figure 4.2: Tone labels, segment labels, f_0 curve and spectrogram for the utterance *Firmiamolo?* (upper) and for a fragment of the utterance *Vedrai la bella mano di mammola domani?* (lower).

Finally, the pragmatic meaning of L*+H in American English (Ward and Hirschberg 1985) is different from that of L*+H in Neapolitan Italian[5] Finally, in Neapolitan the question "tune" is always characterized by a HL phrasal fall, either immediately following the focal accent (as in Figure 4.2, upper) or realized further away (as in Figure 4.2, lower).

In sum, we can say that the American English and the and Neapolitan Italian tonal systems share the phonological contrast, which is an alignment contrast, between L+H* and L*+H, though the contrast differs on various levels. First, the details of the temporal implementation of the starred and unstarred tone differ between the two languages, with American English realizing the contrast in later sections of the segmental string. Second, the temporal contrast between the accent peak location is more subtle in Neapolitan Italian. Additionally, it appears that both tones are tightly aligned to the stressed syllable in Neapolitan and are immediately followed by a HL tone when the rising accent is in nuclear position. Finally, the pitch accents cue different pragmatic contrasts.

Among the cross-linguistic differences listed above, I would like to emphasize the temporal ones and relate them to issues in the segmental perception literature. Several studies show that the difficulties that L2 learners have in perceiving a non-native contrast are not due to a loss of the sensory capacity to discriminate the contrast; rather, it appears that the listeners learn to be "attuned" ("selective perception") to those contrasts that are present in their own language. Also, it appears that listeners can be trained to perceive certain phonetic contrasts in a rather rapid way when the training successfully reduces uncertainty.

[5]Nevertheless, the uncertainty conveyed by American L*+H can be argued to share similarities with the pragmatic value of a yes/no question, in that it denotes uncertainty on the part of the speaker relative to what is being predicated.

One way of accomplishing this is by providing immediate feedback to the listeners, after presentation of extreme exemplars of specific categories. Usually, listeners are able to label certain contrasts with reference to arbitrary categories after discrimination training.

Additionally, learners of a second language seem to differentiate non-native contrasts by attributing a different perceptual weight to the acoustic parameters than those attributed by native speakers (cf. Strange (1995) for a review of relevant studies). It is likely that the same strategy would be employed by listeners who are only trained to perceive a non-native contrast within an experimental setting. Furthermore, degree of difficulty in perceiving a non-native contrast appears to vary according to the specific segment being tested and how it relates to the segmental inventory of the language (Strange 1995). Psychoacoustic "salience" of the segment appears to be particularly important.

Listeners can also be instructed to attend to signals in either a linguistic or a psychoacoustic mode, which generally leads to different results. It has been shown that when listeners attend to signals psychoacoustically their discrimination capacity is increased in the middle of the stimulus continuum, while categorization is poor (cf. Beddor and Gottfried (1995)). For this reason, the experiment described here includes both a linguistic and a psychoacoustic task.

The main assumption of this study is that the temporal alignment contrast between L+H* and L*+H is perceived differently by native speakers of Neapolitan Italian and American English, based on the different way they seem to exploit tonal alignment detail in their intonation systems. On the other hand, I also assume that both systems exploit a similar set of phonetic intonation cues. As an example, f_0 scaling might be employed by both sets of speakers to signal degree of local emphasis, while temporal alignment can be used as

a broad index of tune-text association (though the exact mapping of a specific alignment or scaling contrast to the pragmatic meaning can be different). Despite the differences between the intonational system of Neapolitan Italian and American English, we can then assume that both sets of speakers might be able to perceive the same phonological contrast.

Consequently, I formulated two main hypotheses. Hypothesis A is that American English listeners will be able to perceive the temporal alignment contrast that characterizes the question/statement opposition in Neapolitan Italian. Therefore, like for the Italian results, I expect that hypothesis 1 (stating that when the LHL configuration is moved backwards within the stressed vowel, a lower percentage of question identification responses should be obtained) will be confirmed for the American results. I also expect, consequently, that the same effect of peak shape will be found, so that hypothesis 2a and 2b will be rejected as for the Italian results (for the hypothesis formulation, see §3.1), as well as an effect of L1 and L2 timing but no effect of overall pitch level and slope. The second main hypothesis (hypothesis B) is that listeners will respond similarly to the linguistic task (in which either question or statement meaning has to be identified) and the psychoacoustic task (in which a specific acoustic detail of the stimulus has to be identified).

4.2 Methods

4.2.1 Stimuli preparation

The set of stimuli employed was the same as the one for the experiment described in Chapter 3. See §3.2.1 for details.

4.2.2 Experimental Procedure

Thirty-seven American undergraduate students from the Ohio State University participated in the experiment. The listeners were all native speakers of American English and were paid for their participation. The stimuli were played through headphones in a double-walled sound-proof booth. The listeners were seated in front of a computer, and in the first part of the test (linguistic task) they were instructed to click either a statement or a question button after hearing the stimulus. In order to ensure that there would be no order effects, button position on the computer screen was randomly varied from trial to trial. As to the specific instructions, the listeners were told that Neapolitan Italian differentiates questions from statements through intonational means alone, and were given some examples from the experimenter.

A practice list was played at the beginning of the test containing extreme exemplars of the synthetic stimuli as well as natural exemplars of questions and statements produced by the same speaker[6], for a total of 20 stimuli. After the listener made a choice relative to the practice stimulus heard, feedback was immediately provided by leaving the right button lit on the screen before going on to the following trial.

Once the practice was concluded, a percent correct appeared on screen. Only those listeners reaching at least 80% correct responses could proceed to the actual test. However, if the criterion was not met with the first practice, another practice (with the same characteristics) could be carried out. If the criterion was not met after the second trial, the listener was excluded from the main test.

A similar procedure was followed for the psychoacoustic test. In the second part of the experiment, the listeners' attention was directed to the temporal difference between "early" and "late" alignment of the pitch peak. We expected that focusing attention on one of the acoustic details of the stimuli would have improved sensitivity as well as identification, as shown for segmental studies (Beddor and Gottfried 1995). The difference between early and late timing was also shown with a schematic pitch track on the instruction sheet. Crucially, the listeners had no knowledge of the fact that such phonetic contrast cues the question/statement opposition in Neapolitan Italian.

The stimulus group was played 5 times in 5 differently randomized blocks (to avoid order effects) with each stimulus occurring once per block. The stimuli were presented

[6]The inclusion of multiple tokens by one talker (or more) is useful in order to include variation of phonetic parameters that might be irrelevant to the purpose of identification, such as loudness, speaking rate, etc. The ecological validity of the task increases by including such stimuli, since listeners are forced to respond in terms of relevant phonetic information while ignoring irrelevant detail.

binaurally through headphones. After listening to a stimulus, the listeners performed a two-alternative forced choice task (either choosing between question and statement or choosing between early and late peak timing). Once the choice was made, the next stimulus was played.

Different sets of listeners participated in the linguistic and the psycholinguistic task. Of the original 17 listeners recruited for the linguistic task, 13 reached a criterion of at least 80% correct after the short training period and were allowed to participate in the actual experiment. As to the psychoacoustic task, of the original 20 listeners, only 11 reached the set criterion and could complete the entire experiment.

4.3 Results

4.3.1 Linguistic task

Primary continuum

Despite the variability observed within the data, the results indicate that naive subjects can identify the temporal alignment contrast of Neapolitan L*+H vs. L+H*. Figure 4.3 shows the results for the primary continuum. The scores were calculated in the way described in §3.3.1.

Also here, similarly to the Italian results reported in §3.3.1, shifting the L1-H-L2 configuration backwards within the accented vowel decreased the number of question responses. As Figure 4.3 shows, at early locations within the timing continuum, such as at T1 and T2, statement responses were largely dominant, while at the opposite end of the continuum question responses were largely dominant. Also, the variance (see standard error bars in

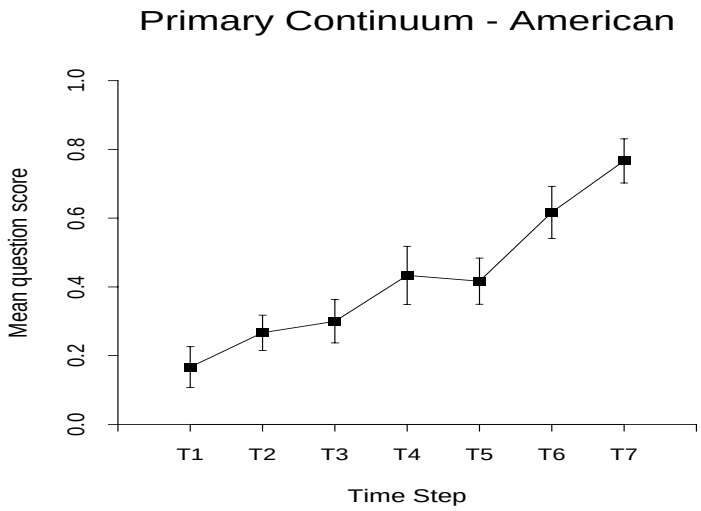

Figure 4.3: Primary continuum mean scores for American listeners in the linguistic task. Standard error is indicated by vertical bars.

Figure 4.3) was globally larger than for the Italian results, which indicates higher uncertitude on the part of the listeners. However, also here, the largest variance was recorded at the middle of the continuum. The ANOVA was then performed, as for the Italian results, on arcsine transformed data. The timing manipulation yielded a significant result [$F(6,77) = 9.95$; $p < 0.01$].

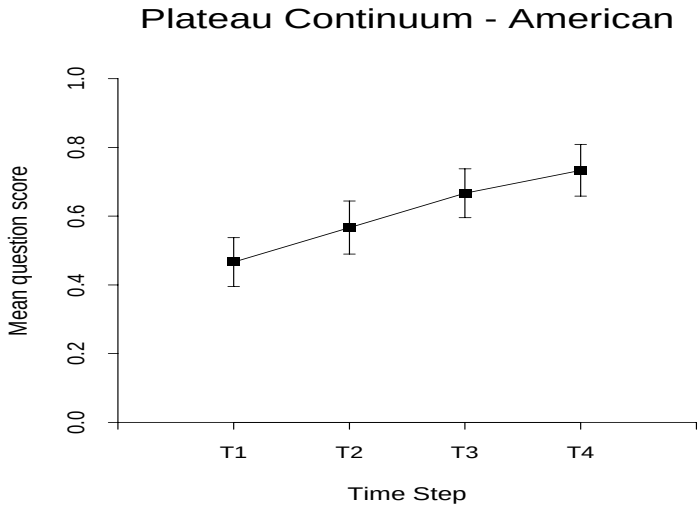

Figure 4.4: Plateau continuum mean scores for American listeners. Standard error is indicated by vertical bars.

Unlike the Italian results for the same continuum, the category shift occurs later, i.e., between T5 and T6. This is expected, given the later alignment of peaks in the American English L+H*/L*+H contrast as emerging from Pierrehumbert and Steele (1989). The identification function was also less steep than for the Italian results, which suggests higher inconsistency in labeling stimuli which were close to the boundary. However, endpoint stimuli were identified above chance.

Plateau continuum

Figure 4.4 presents mean question results for the plateau continuum. Notably, similarly to the Italian results, also here a great number of question responses was already obtained

at early timing within the plateau continuum, with the score for stimulus 2 that is already above chance. This is taken as supporting evidence that the effect is in some sense "universal" and not due to language-specific differences in tonal structure. Also, notice that data variability is high both at the beginning and at the end of the continuum (see error bars in Figure 4.4).

Since the effect is very similar to the one found for Italians, I also tested the hypothesis (hypothesis 2a in Chapter 3) stating that T1 peak stimuli would score the same percentage of question responses as T1 plateau stimuli. Hence, I 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. Also here, the Shape manipulation was significant [$F(1,96) = 35.1; p < 0.01$] while the Timing manipulation was not, though very close to the rejection threshold [$F(3,96) = 3.9; p = 0.01$]. The interaction was not significant. Thus, hypothesis 2a can be rejected also for the American data.

Then, I also tested hypothesis 2b, stating that peak stimuli timed with the plateau offset of corresponding plateau stimuli would receive same question scores. From a simple inspection of the data, it was already noticed that the plateau results resemble the results of peak stimuli timed with plateau offset (see Figure 4.4).

However, while the effect of Shape was significant also for this comparison in the Italian results, here it was not significant [$F(1,96) = 0.84; p = 0.36$]. The Timing manipulation was instead significant [$F(3,96) = 6.9; p < 0.01$] and no interaction was found, [$F(3,96) = 0.38; p = 0.8$]. Hence, hypothesis 2b could not be rejected for the American results, that is, the perceived target for plateau stimuli can be identified with that of stimuli with peak timed at plateau offset. Then, the question if plateau stimuli are perceived as simply having a higher pitch than peak stimuli timed with plateau onset was also asked here. The impact of f_0 height is investigated in the next section.

f_0 continuum

Figure 4.5 shows mean question scores plotted against timing for the f_0 continuum. Remember that this continuum was designed to test the effect of overall pitch level on question/statement response, as a result of perceptual target displacement.

Similarly to the Italian listeners, the American English subjects appeared to be influenced by overall f_0 level. Specifically, they identified hhh stimuli mostly as questions, while lll stimuli received less question responses. The differences between mean scores appear more marked here, though. Also, note that at T7 lll stimuli did not receive the same percentage of question responses as the other stimuli, which instead was the case for the Italian results. Overall, though, the effects observed in Figure 3.14 are remarkably similar to those found in Figure 4.5.

The results of a two-way ANOVA revealed a significant effect of f_0 level [$F(2,108) = 11.1; p < 0.01$] as well of Timing [$F(2,108) = 12.25; p < 0.01$], while no interaction was found [$F(2,108) = 1.07; p = 0.4$]. However, much like the Italian results, a post-hoc Tukey

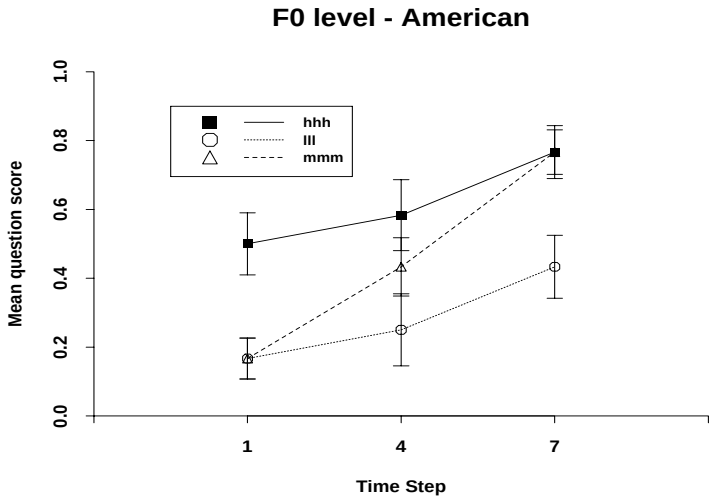

Figure 4.5: Mean scores for the f_0 continuum (pooled results). Standard error is indicated by vertical bars.

test (confidence level = 0.01) showed that the only significant difference of f_0 level was the one between lll and hhh stimuli, while mmm stimuli did not differ from either hhh or lll stimuli. The results seem to bear out the prediction that native American English speakers use f_0 height in order to identify the question/statement contrast. On the other hand, no significant difference seem to exist between stimuli with peaks at either a normal (mmm) or high (hhh) f_0 value, similarly to the Italian results. Hence, also here, such peak values will be employed in order to test the hypothesis of an effect of rise and fall slope (hypothesis 4).

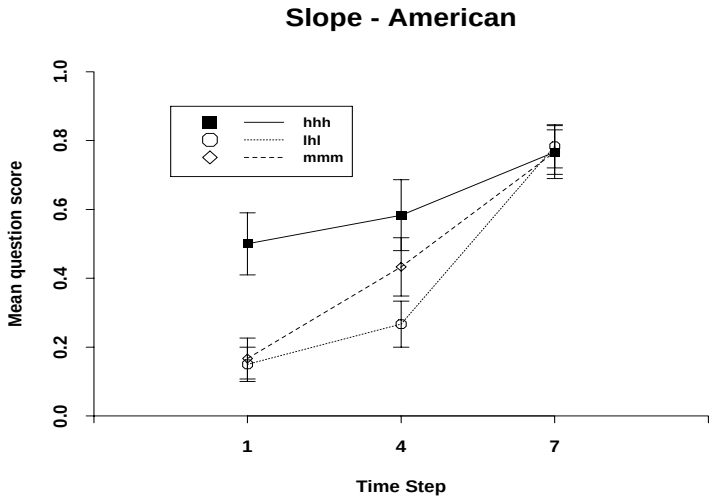

Figure 4.6: Mean scores for the slope continuum (pooled results). Standard error is indicated by vertical bars.

Slope continuum

Figure 4.6 shows mean scores for the mmm and lhl stimuli, in which the slope of the rise and fall are obviously different. The hhh stimulus combination was added, as in the Italian perception study, in order to test whether L1 and L2 f_0 height might add to the effect.

Results are shown in Figure 4.6. Note that the findings for the lhl combination go in the opposite direction relative to the Neapolitan results shown in Figure 3.15. Remember that for Neapolitan lhl obtained more question responses (rather than fewer) and the differences were biggest at T1. Here we find the opposite situation, instead. Also, note that the scores are more different between the hhh series and the other series at T1, unlike in the Italian results.

A two-way ANOVA revealed in fact a significant effect of both Stimulus Type [$F(2,108) = 5.14; p < 0.01$] and Timing [$F(2,108) = 27.09; p < 0.01$], but no interaction [$F(4,108) = 1.19; p = 0.3$]. However, when a posthoc analysis was run on Stimulus Type, no level resulted to be significantly different from any other (Tukey, confidence level = 0.01). The hypothesis of an effect of slope on the results (hypothesis 4) can therefore not be rejected.

L1 and L2 timing continua

The effect of L1 and L2 timing was also studied for the American listeners. Specifically, hypothesis 5a of Chapter 3, claiming that stimuli with same H timing and different L1 timing would obtain the same percentage of question responses, was tested. Analogously, hypothesis 5b, claiming that stimuli with the same H timing but different L2 timing would score similarly, was tested.

Also here, separate one-way Analyses of Variance to assess the effect of L1 at each H timing location were performed on the arcsine data, with Stimulus Type as factor. Unlike the Neapolitan data, though, and as it can be inferred already from the raw results, neither L1 nor L2 timing manipulations were significant at any of the H timing locations, with p values always above 0.1. Results for the L1 timing manipulation at H timing 1 and 4 and for L2 timing manipulation at H timing 4 and 7 are shown, respectively, in the upper and lower panel of Figure 4.7. Note that the pattern of responses for L1 timing is slightly different from that of Neapolitan listeners, shown in Figure 3.16.

L1 timing	**H timing**	**L2 timing**	Mean Q Score	SD
1	1	1	.17	.21
4	1	1	.27	.21
1	1	4	.30	.23
1	1	7	.43	.22
4	4	4	.43	.29
1	4	4	.47	.27
7	4	4	.43	.28
4	4	7	.40	.31
4	4	1	.38	.28
7	7	7	.77	.22
1	7	7	.60	.26
4	7	7	.75	.21
7	7	4	.67	.16

Table 4.1: Mean question (Q) scores and standard deviation (SD) for all the L1 and L2 timing manipulations.

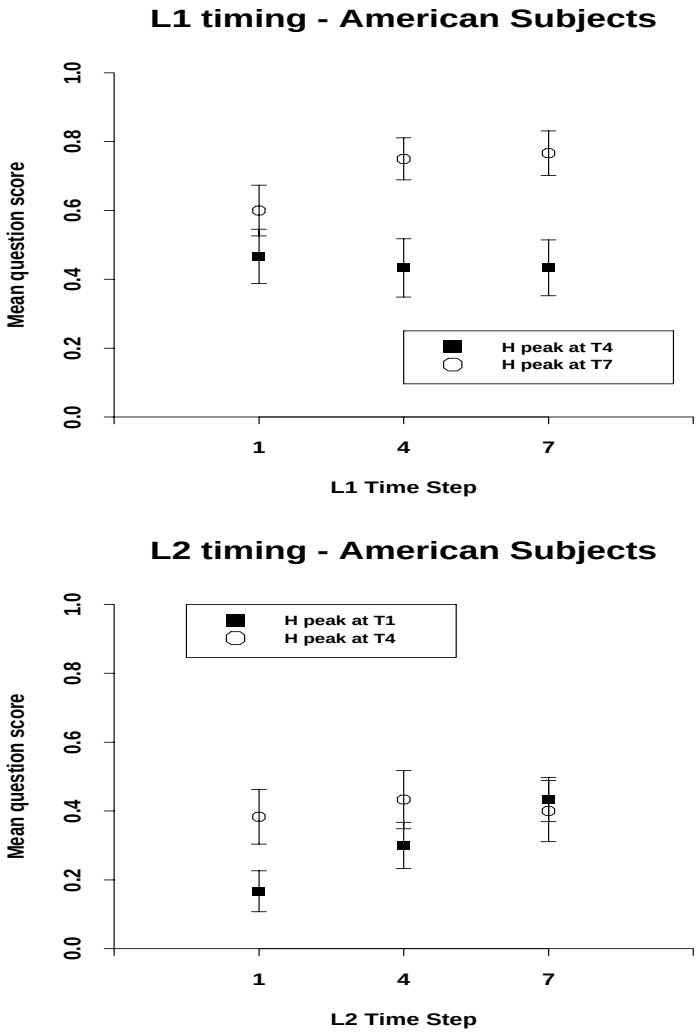

Figure 4.7: Mean scores for the L1 timing manipulation at H timing 4 and 7 (upper) and for the L2 timing manipulation at H timing 1 and 4 (lower). Standard error is indicated by vertical bars.

For Neapolitans, earlier L1 translated into fewer question responses at T4 peak timing, while for Americans there is a similar trend at T7. Analogously, the trend for fewer question responses at earlier L2 timing, which is visible at T1 peak timing in the lower panel of Figure 4.7 was found (as a clear effect) at at both T1 and T4 peak timing for Neapolitans (see Figure 3.17).

4.3.2 Psychoacoustic task

Unlike the results of the linguistic task, neither of the manipulations resulted to be significant. It will suffice to report here the results of the primary continuum stimuli, which are shown in Figure 4.8. This Figure shows mean late peak judgments on the y axes and stimulus timing on the x axes. Unlike the response function for the linguistic task, listeners here did not show a clear crossover boundary for any of the timing values for the L1-H-L2 configuration, with stimuli identified around chance level at both end of the continuum (.42 at T1 and .58 at T7). The results of a one-way ANOVA showed in fact no significant effect of Timing [$F(6,70) = 0.9; p = 0.47$]. Also, as the error bars in Figure 4.8 show, the variance was constant throughout the timing continuum, which was not the case for the linguistic results.

4.4 Discussion

The results of this section showed that, when it comes to identifying a question vs. a statement in Neapolitan Italian, American English listeners (who had never learned Neapolitan Italian before) are capable to accomplish the task, and can be trained for such a purpose

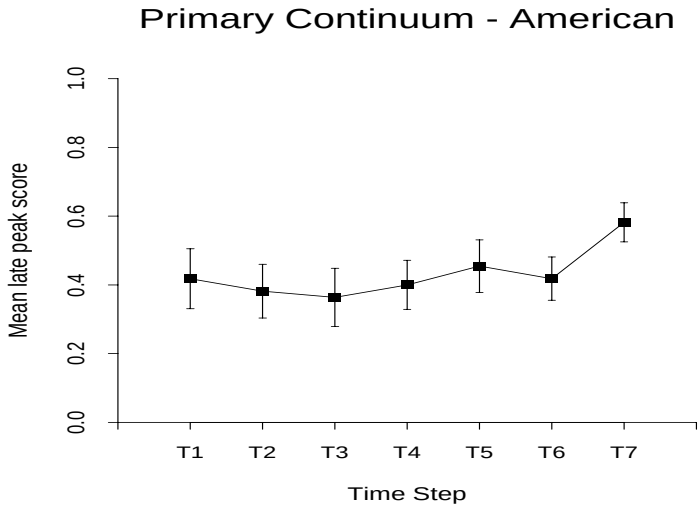

Figure 4.8: Primary continuum mean scores for American listeners in the psychoacoustic task. Standard error is indicated by vertical bars.

rather rapidly. The experiment also showed that American English listeners can employ a subtle alignment contrast, similar but not identical to the one exploited by their intonation system. Therefore, hypothesis A is confirmed. This result is reminiscent of some results reported within the segmental perception literature, in which listeners are trained to perceive contrasts which are implemented differently in their native language.

Several studies have shown that listeners, when confronted with the task of perceiving a non-native contrast, tend to rely on acoustic cues that are used phonemically in their own language (cf. Bohn (1995) for a review and an alternative view). As to the f_0 continuum, listeners seemed to associated high overall f_0 with the question meaning (i.e., they were sensitive to gross phonetic shape of the patterns).

Some of the observed f_0 level effects were not expected on the basis of "transfer" from native contrasts. For instance, since in American English L+H* is often used as emphatic contrastive accent, it is usually associated to larger f_0 excursions. Therefore, we would have expected that the larger excursion of lhl stimuli would pattern with earlier peak responses, hence cue more statement responses. But this was not the case. Also, given the fact that the L target for American English L*+H is typically lower than that for L+H*, we would have expected the lower f_0 value of L1 in lhl stimuli to pattern with late peak responses, and therefore cue more questions. But this expectation was not met.

Nevertheless, lhl stimuli received a great number of question responses at T7, despite the low final contour. This is similar to the result obtained by Hadding-Koch and Studdert-Kennedy (1976), which, like previous studies, showed that both the direction of the terminal section ("terminal glide") as well as the shape and the f_0 level of the intonation contour are relevant when perceiving question vs. statement intonation in American English. Specifically, in Hadding-Koch and Studdert-Kennedy (1976) question responses on the part of American listeners appeared to be related to both terminal contour direction as well as to the pitch level of the "turning point" (which corresponds to the H peak here). This might account for the difference between lhl and lll stimuli.

Also, the late crossover for the timing continuum (between stimuli 5 and 6) might be due to the fact that the American listeners were responding in terms of the alignment characteristics of the L+H*/L*+H contrast proper to their intonational phonology. In fact, the H target for English L*+H is aligned much later than for the Neapolitan L*+H. That is, in American English, the peak is usually timed beyond the stressed syllable boundary, while the L valley is generally located within the stressed vowel. Late category boundary was also found in Pierrehumbert and Steele (1987), which investigated the perception of L+H* vs. L*+H in American English (see §1.5 for details). Also, the fact that f_0 level cues are actively employed to identify a question might be related to the L+H*/L*+H characteristics, which have usually a quite marked peak. This fact might explain why, even at late timing locations, lll stimuli never received the same score of question responses than the other stimuli.

Although within much early literature on segmental perception it had been claimed that perceptual sensitivity to nonphonemic contrast diminishes as an effect of linguistic experience, more recently it has been shown that, with the appropriate experimental procedure, listeners can be trained to perceive subphonemic phonetic contrast. As Pisoni et al. (1982, p. 313) already stated:

> *There is little solid empirical evidence that the underlying sensory or perceptual apparatus, as yet undefined, has been "retuned" or modified in any permanent manner as a result of selective early experience.*

This appears to be true even after a short training period (cf. Pisoni et al. (1982)).

Note that no effect of L1 or L2 timing was found for American listeners, while it was found for Italians. Only a speculative account can be offered here, which is based on recent findings. That is, American speakers seem to align the L of a L+H* rise (which is the accent that most resembles both L+H* and L*+H of Neapolitan, in that the H peak is always located within the stressed vowel) at the onset of the stressed syllable(Ladd, Faulkner, Faulkner, and Schepman 1999). Therefore, it might be more difficult to train them to perceive the timing difference for the L than for the H peak.

Like for the Italian listeners, plateau stimuli were identified mostly as questions, which lends additional support to the idea that this effect has a "universal" nature. That is, the wider "peak" has the effect of displacing perceived target location (either in the time dimension, as it is claimed here, or in the melodic dimension) for the rise-fall, relative to sharp peak stimuli. Such an effect cannot be explained in terms of parsing a language-specific tonal structure, since obviously American English and Italian do not share the same phonological system neither the same phonetic implementation mechanisms.

Unlike the present study, Hadding-Koch and Studdert-Kennedy (1976) found that the identification function of the psychophysical responses to their stimuli was much the same as the one characterizing "semantic" (linguistic) responses.[7] Here, though, we found the responses to the psychoacoustic task (early vs. late peak responses) to be all around chance level. Hence, hypothesis B (claiming that the responses of the linguistic task and of the psychoacoustic task would be the same) is not supported by the data. This could have been caused by an inherent difficulty of the task (it is not easy to discern a timing difference,

[7] Nevertheless, the psychophysical responses showed higher uncertainty, which is similar to what was found here.

while pitch level differences can be more easily singled out). Nevertheless, this might show that when listeners must consciously concentrate on the time alignment difference, they cannot form long-term memory representations that can be employed to compare the stimuli. That is, when the auditory property is not readily available to them, and the linguistic contrast is not suggested (so that simple transfer cannot be applied), listeners have trouble forming abstract categories. In fact, though listeners completed the practice trials successfully, they were not capable to extend identification to stimuli different from the ones employed within the limits of the practice.

CHAPTER 5

PERCEPTION EXPERIMENT III: THE EFFECT OF SYLLABLE STRUCTURE AND BASE STIMULUS ON TARGET PERCEPTION IN ITALIAN

5.1 Rationale

It is well known that, in Italian, stressed vowels in open (and penultimate) syllables are long (cf. D'Imperio and Rosenthall (1999) for a review of the literature and a new theoretical account of the phenomenon). The difference in duration between stressed vowels in open vs. closed syllables is quite marked. In the production study presented in Chapter 2, for instance, the difference was in the order of 30.5 ms for statements and 26 ms for questions (for both speakers).

In Chapter 2, I asked if such a structural difference would affect the alignment of the tonal targets in the LHL configuration. The results appeared to support the hypothesis of an invariant alignment of the L1 and L2 targets relative to the left edge of the syllable. That is, whereas timing relative to the left edge of the syllable was constant across syllable types, when these targets were measured relative to accented vowel offset or following vowel onset, a difference between open and closed syllables was found. Specifically, the tendency was for a tonal target to be closer to the onset of the following vowel when the

accented syllable was closed. Regarding the H target, the evidence was mixed, but more suggestive of an invariant alignment with the left edge of the syllable than with the right edge.

Analogously to the question asked in production, one could ask whether perceived target location is computed relative to the left or the right edge of the syllable. If tonal targets are aligned relative to the left edge of the syllable (either syllable or vowel onset), we expect the difference in duration induced by syllable structure to have no effect on perceived target. If, on the other hand, perceived tonal target location is computed relative to the entire stressed syllable or stressed vowel duration, we expect to find such an effect. The first hypothesis (hypothesis 1) tested here is therefore that question response scores will be the same independent of stressed syllable structure. Specifically, I only tested the target alignment manipulation (where L1, H and L2 are manipulated simultaneously) by creating primary continua whose base would be, in one case, an utterance with a closed stressed syllable and, in the other, an utterance with an open stressed syllable.

Remember also that, in both perception experiments described in Chapter 3 and Chapter 4, the base utterance employed to create the resynthesized stimuli was always a question. It was assumed, for the purposes of this dissertation, that there would be no perceptual bias induced by such a choice. Indeed, it was shown that, when the intonation is flattened out over the entire utterance, listeners cannot reliably identify an original question vs. a statement.

However, dismissing an effect of base stimulus in such a way might be judged to be rather premature. Therefore, the present experiment will include base stimulus (question or statements) as an additional variable. The second null hypothesis (hypothesis 2) is, hence, that question response scores will be the same independent of the modality of the base utterance.

Finally, I intended to replicate the results for the plateau stimuli reported in Chapter 3 and at the same time extend the plateau stimuli series by adding a step to the left end and one to the right end of the continuum (increasing the series to 6 stimuli). This will allow to estimate the shape of the response curve with more accuracy than when only 4 steps are employed.

5.2 Methods

5.2.1 Stimuli preparation

In order to test hypothesis 1 and 2, I employed four stimulus series in which target alignment was manipulated in the same way as for the primary continuum stimuli of Chapter 3 (see §3.2.1). All the main series contained peak stimuli, therefore I will refer to them as "peak continua". In order to create the open (syllable)/question series, I employed the very same stimuli of the primary continuum series. The only difference was the addition of a stimulus to the right end of the continuum (T8). All the other series were also made of 8 time steps, in order to allow for a better evaluation of the shape of the response curve.

A question utterance with a closed syllable target was used as the base of the closed (syllable)/question continuum. The chosen utterance was *Vedrai il nonno?* "Will you see

grandpa?". The focal accent was the same as the one for the primary continuum, i.e., L*+H on the noun *nonno*. Then, for the open (syllable)/statement series, the utterance employed was *Vedrai il nono.* "You will see the ninth", while the base utterance for the closed (syllable)/statement stimuli was *Vedrai il nonno.* "You will see grandpa". For the statement utterances, the focal accent was, naturally, L+H*. The four natural stimuli were added to the stimulus set.

Remember that the duration of the stressed vowel in the base utterance for the primary continuum was 170 ms. In the other series, stressed vowel duration was equal to 136 ms for the closed/question series, 154 ms for the closed/statement series and 190 ms for the open/statement series. Despite the difference in vowel duration, the step sizes for the target alignment manipulation were always the same in all four series. Hence, absolute tonal alignment values (relative to vowel onset) were kept constant for all series (for details, see §3.2.1).

Regarding the new plateau continuum, two new stimuli were added to the series whose details are described in §3.2.1. Specifically, those stimuli were, respectively, at 1 step to the left of stimulus 1 (stimulus 0), and at 1 step to the right of stimulus 4 (stimulus 5), which was the last stimulus in the continuum. Step size was 15 ms, as in the original primary continuum, while plateau duration was 45 ms (like for the original plateau continuum). In this fashion, rise offset in in plateau stimuli had values equal to either T1, T2, T3, T4 and T5 (for stimuli 1, 2, 3, 4 and 5) of the primary continuum, while fall onset could be at T3, T4, T5, T6, T7 or T8. Through such manipulations, a total of 42 stimuli were created, which made the stimulus set.

5.2.2 Experimental Procedure

Twenty-nine listeners from the University "Federico II" of Naples, Italy, participated in the experiment. The listeners, who were paid for their participation, were all brought up in Naples and spoke Standard Italian with a Neapolitan accent. The stimulus set was played directly from a laptop computer that was placed in a studio at the University of Naples. Similarly to the experiment described in Chapter 4, the listeners were seated in front of the computer and, after ten practice trials, they listened to the actual stimuli.

The listeners were instructed to perform a two-alternative forced choice task, in which they had to identify the stimulus heard as either a question or a statement. They indicated their choice by clicking with the mouse one of the two buttons appearing on the screen. Each button was labeled as either *domanda* "question" or *affermativa* "statement". In order to ensure that there would be no order effects, button position (left or right) on the computer screen was randomly varied from trial to trial. The stimulus group was played 5 times in 5 differently randomized blocks with each stimulus occurring once per block. The stimuli were presented binaurally through headphones in a studio at the University "Federico II" of Naples.

As in Experiment I, responses to the natural open/question stimulus identification were employed to set a criterion for subject inclusion in the data analysis. Hence, listeners showing less than 80% of question responses for the natural stimulus were assumed to be unable to perform the task in a reliable fashion. Of the original twenty-nine subjects, twenty-four met the set criterion.

5.3 Results

5.3.1 Peak continua

As mentioned above, the primary continuum of Experiment I corresponds to the open/question continuum of the present experiment (apart from the addition of a stimulus at one end of the continuum). The results for this continuum, as well as those for the closed/question continuum, are shown in the upper panel of Figure 5.1 and in Figure 5.3. The scores were again calculated as mean values for stimulus type per subject. Figure 5.1 (upper) shows mean scores pooled for all subjects (y axis) across stimulus type (x axis), for both open and closed/question continua. The lower panel of Figure 5.1 shows responses for the two statement base continua.

Similarly to the results of the primary continuum in Experiment I, we see an orderly increase from statement to question judgments as a function of timing, as expected. As the upper panel of Figure 5.1 shows, at early locations within the timing continuum, such as at T1 and T2, statement responses were largely dominant (with a mean score of 0.08 for closed questions and 0.1 for open questions at T1), while at the opposite end of the continuum question responses were largely dominant (with a mean score of 0.97 for open questions and 0.94 for closed questions at T8). Analogously, as the lower panel of Figure 5.1 shows, at early locations within the statement base continua, statement responses were dominant, while at the opposite end question responses were dominant. Clearly, though, at this end of the continua, fewer question scores were found than for both question continua (specifically, a mean of 0.81 for open statements and of 0.88 for closed statements).

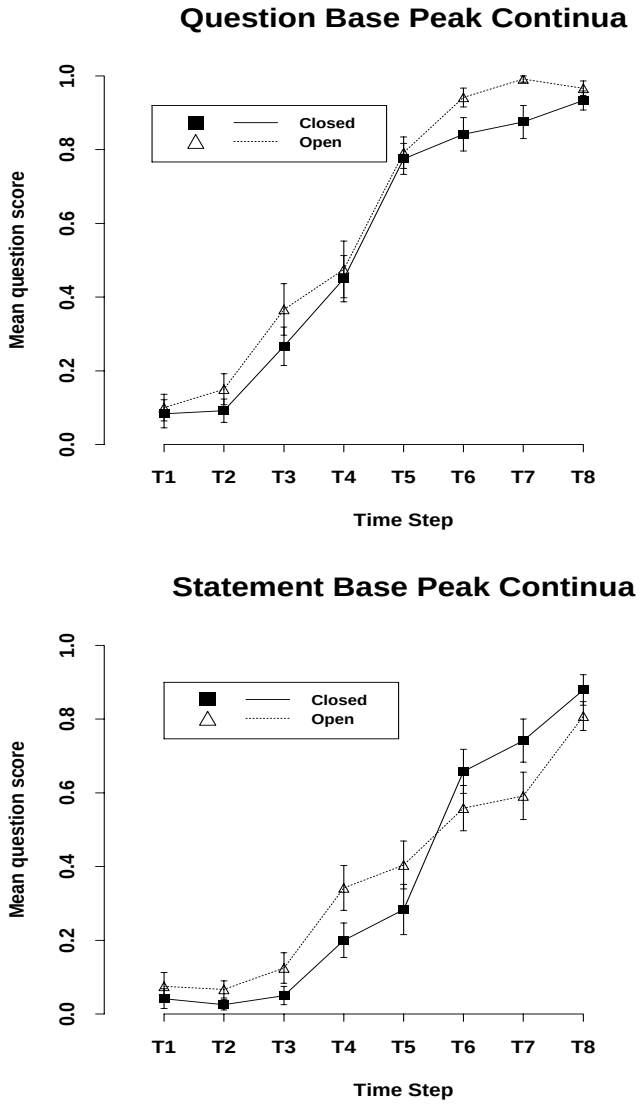

Figure 5.1: Mean question scores for the open and closed question base continua (upper) and for the open and closed statement base continua (lower), for all subjects. Standard error is indicated by vertical bars.

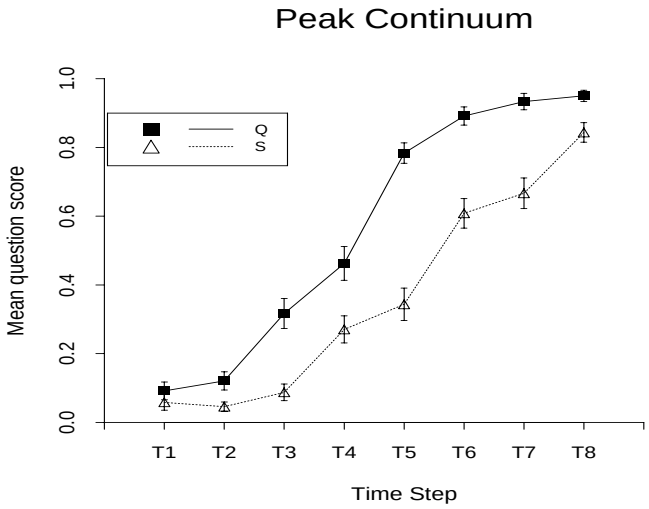

Figure 5.2: Mean question scores for the question vs. statement (open and closed) base continua, for all subjects. Standard error is indicated by vertical bars.

Also here, the variance was not constant throughout the timing continua, being greater at the middle than at the continua edges (see Figures 5.3 and 5.4). Therefore, the statistical analysis was performed on the arcsine transformed data. A three-way ANOVA, with Timing, Open/Closed and Question/Statement as main factors, was performed on the pooled results for all timing continua. The Timing manipulation yielded a significant result in all cases, as expected [$F(7,736) = 199.16; p = 0$]. The Open/Closed manipulation was not significant [$F(1,736) = 4.43; p = 0.04$], though the value was quite close to the cutoff point. Interestingly, the Question/Statement base factor was instead significant [$F(1,736) = 150.14; p < 0.01$].

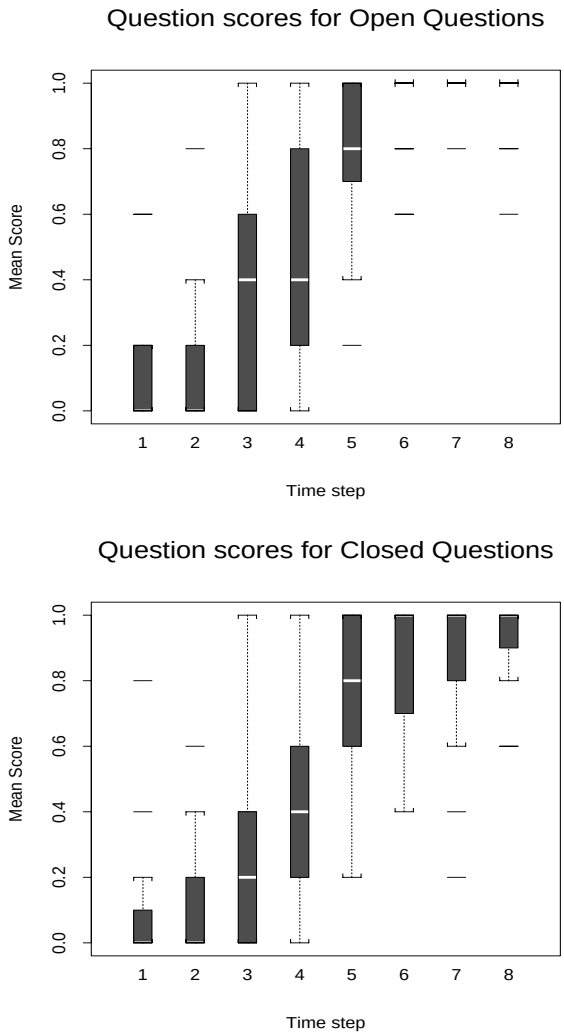

Figure 5.3: Boxplot for the open (upper) and closed (lower) question continuum scores for all subjects. The horizontal line within the box represents the median, while the height of the box is equal to the interquartile distance (IQD). The dotted lines extend to the extreme values or a distance equal to 1.5 x IQD from the center, whichever is less.

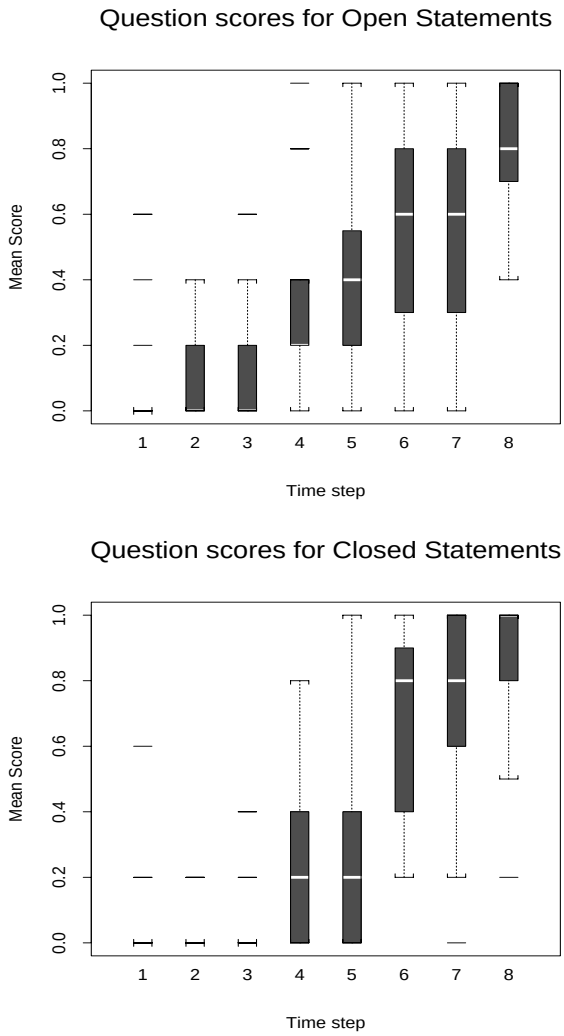

Figure 5.4: Boxplot for the open (upper) and closed (lower) statement continuum scores for all subjects. The horizontal line within the box represents the median, while the height of the box is equal to the interquartile distance (IQD). The dotted lines extend to the extreme values or a distance equal to 1.5 x IQD from the center, whichever is less.

Such an effect can be more easily evaluated from Figure 5.2, which compares results for the question vs. statement continua (collapsing across syllable type) against each other. Among the interactions, only the Timing by Question/Statement was significant [$F(7,767) = 7.9; p = 0$].

In order to further evaluate the similarity of the response functions for the open vs. closed syllable series, a χ^2 test was applied to the data. Specifically, the two statement base series results were compared first (where the open/statement series was the observed data and the closed/statement series the expected data). There was a significant deviation from predicted values [$\chi^2 = 30.59; df = 7; p = 0$]. The same comparisons for the question series showed instead no significant difference [$\chi^2 = 10.4; df = 7; p = 0.17$].

The ratio of question responses for each series was also plotted as a function of timing for each subject. To statistically evaluate such data and calculate 50% response points of the identification function, the data were smoothed using a logistic regression (see §2.4.4 for details about logistic regression modeling). Then, the 50% response point was recorded for each individual function. These values were employed to evaluate whether a boundary shift is produced as a result of each manipulation.

First, I calculated the boundary value for the primary continuum results of Experiment I (described in Chapter 3), in order to test if the same boundary would be obtained when the continuum was inserted in a different experimental context. The overall mean boundary value was equal to step 3.57 (standard deviation = 1.08), which corresponds to 115.05 ms into the stressed vowel.

Then, crossover points were recorded for each individual function relative to the syllable structure manipulation. Such values were used as the basis for a two-way analysis of

Subj.	QBound. (ms)	SBound. (ms)	Subj.	QBound. (ms)	SBound. (ms)
1	111/118	122/136	13	120/119	172/154
2	86/99	97/117	14	133/123	150/167
3	124/131	162/145	15	122/127	169/155
4	125/110	143/136	16	131/133	163/167
5	103/91	105/119	17	142/177	176/179
6	119/119	152/149	18	145/163	176/135
7	90/106	121/138	19	127/143	178/155
8	102/119	113/161	20	107/128	101/140
9	86/61	98/106	21	132/125	142/142
10	107/118	140/129	22	131/134	163/161
11	98/101	135/124	23	95/121	146/136
12	105/116	129/142	24	127/133	172/148

Table 5.1: Means for the 50% crossover boundary (inflection point) of the logistic function fitted to the question and statement continua for each subject. Means are given in ms, from stressed vowel onset. Open continua values are on the left of each couple (closed continua values are on the right). QBound = Question Boundary; SBound = Statement Boundary.

variance, with Open/Closed and Question/Statement as factors. The hypothesis tested was that differences in syllable structure would not produce a boundary shift, though production data showed a seeming alignment difference in open versus closed syllables.

The results of the statistical analysis did not appear to support the hypothesis, since neither the effect of open vs. closed continua boundaries [F(1,92) = 0.55;$p = 0.46$] nor the interaction of the main effects [F(1,92) = 0.54;$p = 0.46$] were significant. On the other hand, as for the results of the ANOVA on the question score results, the question vs. statement base effect was significant [F(1,92) = 32.07;$p < 0.01$]. The overall mean boundary values were 3.69 (st. dev. = 1.13) for the open question continuum, 3.74 (st. dev. = 1.53) for the closed question continuum, 5.56 (st. dev. = 1.85 and 1.16) for the open as well as for the closed statement continuum. These step values corresponded, respectively, to 115.35 ms, 116.1 ms and 143.4 ms into the stressed vowel. Note that the boundary value for the open/question continuum was very similar to the one for the primary continuum, which means that, since the continua are the same except for one stimulus, the primary continuum results are replicated. Mean boundary values for all continua, for each individual subject, are given in Table 5.1.

5.3.2 Plateau continuum

Figure 5.5 presents mean question results for the plateau continuum. As in Chapter 3, we notice a greater number of question than for the peak continua already at early timing locations.

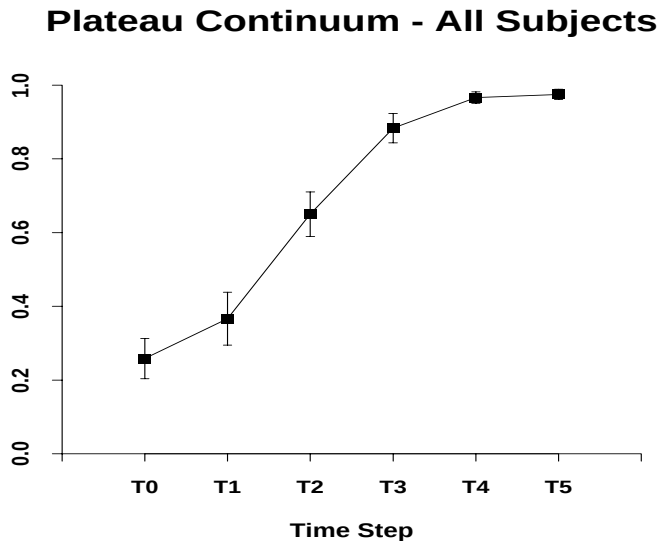

Figure 5.5: Mean scores for the plateau continuum for all subjects. Standard error is indicated by vertical bars.

Plateau Stimulus	Mean score	Primary Stimulus	Mean Score
T0	0.29	T3	0.37
T1	0.38	T4	0.47
T2	0.61	T5	0.79
T3	0.86	T6	0.94
T4	0.94	T7	0.99
T5	0.95	T8	0.97

Table 5.2: Mean scores for the plateau continuum (pooled results, left) and mean open/question continuum scores for stimuli timed with plateau offset (right).

Since more stimuli were added to the plateau continuum employed in the present experiment, I tested once more the hypothesis (hypothesis 2a of Chapter 3) that T1 (open/question) peak stimuli (with peak timed at plateau onset) would score the same percentage of question responses as plateau stimuli. I therefore performed a two-way ANOVA on the arcsine transformed results for plateau stimuli (only T1 to T5) and peak stimuli (from T1 to T5), with Timing and Shape (peak or plateau) as independent variables. The results were significant for both Timing and Shape manipulation [Timing: $F(4,240) = 50.28; p = 0$; Shape: $F(1,240) = 151.4; p = 0$]. The interaction resulted also significant, [$F(4,240) = 4; p < 0.01$]. Hence, the results made me reject once more hypothesis 2a.

Then, I went on to test again hypothesis 2b, i.e., that peak stimuli timed with plateau offset of corresponding plateau stimuli would receive same question scores. Scores for the open/question continuum and plateau stimuli for this comparison are shown in Table 5.2.

Similarly to the results presented in Chapter 3, the results for the two series are quite similar (results for the plateau stimulus are once more only somewhat smaller). Unlike the results presented in Chapter 3, though, and confirming the similarity of the results presented here, the results of a two-way ANOVA on the arcsine transformed data for all plateau stimuli and peak stimuli from T3 to T8 were not significant for Shape [$F(1,288) = 6.11; p = 0.014$] while Timing was, again, significant [$F(5,288) = 75.65; p = 0$]. The interaction of the main effects was, also, not significant [$F(5,288) = 0.7; p = 0.6$]. The results therefore show that the perceived target of plateau stimuli overlaps with peak stimuli timed at plateau offset.

5.4 Discussion

The results presented in this Chapter appear to lend support to the hypothesis that syllable structure does not affect the percent of question responses in the perceptual task employed here. That is, it appears that the presence vs. absence of a coda segment and/or a longer vs. shorter stressed vowel do not shift the category boundary between statement and question responses. However, the results of a χ^2 test show that the response functions for open vs. closed/statement series are different. In fact, the shape of the response function for closed statements appears to be slower than the one for open statements up to the crossover point, while becoming faster after such a point. Question response functions were instead not found to be different. Syllable structure does seem to have an effect, therefore, though it is not a clear one.

Additionally, note that the response function obtained for the open/statement series was quite shallow. Usually, stimulus naturalness is taken to be responsible for the greater/lower categorical nature of the response function, with enhanced naturalness usually resulting in response functions that are more categorical in shape (Beddor and Gottfried 1995).

To account for such data is rather difficult at this point. One could speculate that the crossover boundary is a result of alignment computation relative to vowel onset and that, therefore, syllable structure differences are not relevant. This might be especially true for questions, for which the peak is already located rather late within the stressed syllable.

This result contrasts with the one obtained by Rietveld and Gussenhoven (1995) for Dutch (cf. §1.5). In that study, the perception of two accentual categories was affected by the presence vs. absence of a coda consonant within the stressed syllable, in that voiced codas shifted the Point of Subjective Equality, as calculated by the authors, to the right.

Here, instead, the category boundary is shifted to the right only as a result of base stimulus type. That is, statement base series had later crossover boundaries than question series. Such a results cannot simply be accounted for with the duration difference of the stressed vowel in the statement vs. question series. In fact, such a duration difference existed also between the two question series, but no category shift was obtained there. The effect must therefore be due to indices other than mere duration. That is, the original formant values, spectral balance and tilt, which are likely to be different for the statement bases and question bases, might have caused the boundary shift. The late boundary location in the experiment reported in Pierrehumbert and Steele (1987) (see §1.5 for details) was also accounted for in terms of a bias induced by some property of the original stimulus that might have worked as a secondary cue favoring one answer over the other. Furthermore, a small effect of base type can be inferred from the results of D'Imperio and House (1997), discussed in §1.5.1 above. However, the quest for such putative cues will be left to future research.

The results also show earlier timing of "perceptual" boundaries than production boundaries. In the production study it was pointed out, for instance, that peak location for open syllable questions was around 135 ms, while here the crossover boundary for the same

series was around 115 ms. This might be due to the fact that the rise-fall of the base utterances employed for the perception experiment was always phrase-final here (while it was not in the production study), but this is only speculative.

Finally, the study offered more supporting evidence for the hypothesis (hypothesis 2b of Chapter 3) that plateau stimuli receive the same question scores as stimuli with peak timed at plateau offset. A similar result was reported for American listeners in Chapter 4. I take this to mean that the perceived target of plateau stimuli is timed with the onset of the HL fall (i.e., the end of the plateau).

CHAPTER 6

CONCLUSION

6.1 Summary of the findings

The goal of this dissertation was to investigate the temporal alignment of tonal targets as it relates to target perception. This was accomplished by formulating predictions within the autosegmental-metrical framework of intonational phonology and by describing the categories of interest (i.e., pitch accents and edge tones) in terms of a ToBI-style transcription system, specifically tailored to Neapolitan (see also Grice et al. (in press)). The use of these models was crucial in formulating specific predictions employed to design the perception experiments of Chapter 3, 4 and 5. Also, the use of PSOLA resynthesized utterances helped to avoid unnaturalness within the perceptual stimuli employed, thus contributing, together with the straightforward listeners' task, to the acquisition of reliable data.

Some of the previous studies on the perception of alignment (which are not numerous) have been marred by either methodological flaws, unsupported assumptions or non-explicit predictions about what constitutes the "same" percept. In this thesis, instead, I employed a contrast that is very clear and common to all speakers of the variety of Italian under investigation.[1]

[1]Unlike the Neapolitan L*+H, the American English L*+H category, for instance, seems to be quite

Results from the perception experiments show that listeners of both Neapolitan Italian and American English employ contrastive alignment of the entire rise-fall contour in order to perceptually identify pitch accent contrast. The results also suggest that the alignment of the sole initial or final L (within the LHL sequence) affects listeners judgments (though the results were significant only for Neapolitans, while a trend was observed for American listeners). The results are briefly summarized below.

Neapolitan Italian speakers can mark the pragmatic distinction between a yes/no question and a narrow focus statement by employing systematic temporal alignment differences of the acoustic targets associated to the tones of the rise-fall configuration, as shown in Chapter 2. Specifically, L1, H and L2 are aligned later, relative to the left edge of the stressed syllable, in yes/no questions than in narrow focus statements. By contrast, no systematic difference was found for the slope of the LH rise nor for the slope of the HL fall. The exact contribution of f_0 height to signaling the contrast could not be determined, though. In fact, while speaker LD seemed to mark the difference by producing higher peaks for statements, MD did not produce any difference. Also, the higher peaks of LD statements might be accounted for in terms of markedness. In fact, narrow focus statements are in a sense intrinsically emphatic. Questions, instead, present the same pitch accent for broad and narrow focus utterances.

After determining the range of variability of alignment and scaling for the two speakers, the alignment of all three targets (L1, H and L2) was modified through resynthesis of an original question. The stimuli thus created successfully shifted the perception of a question

---

marked, to the point that the phonology of some speakers might lack it completely, as suggested by Pierrehumbert and Steele (1987).

to the perception of a statement when the alignment of the three targets was shifted earlier within the stressed vowel, as shown in Chapter 3. This confirms results for the inter-peak stimuli shown in D'Imperio and House (1997). In the present study, though, target alignment was shifted only within a limited section of the stressed syllable, so that peak locations would all be included within a critical region for question and statement peaks (i.e., starting from 60 ms into the stressed vowel). This way, I avoided the creation of stimuli whose alignment could be mapped onto an additional and unrelated pitch accent category, such as H+L*, as well as avoiding the perception of accent peaks on the syllable preceding the accented one. Moreover, the step size employed here was much smaller, and the prenuclear contour was stylized so as to minimize the contribution of "head-type" to question identification.

The manipulation of target alignment had a significant effect also on responses of a different language group, as shown in Chapter 4. Namely, American English speakers, when asked to identify a question or a statement for the same set of stimuli, identified more statements at earlier target timing. The phonology of American English shares with that of Neapolitan a contrastive use of alignment, as well as sharing a contrast between two rising accents, L+H* and L*+H. However, specific phonetic details of such categories, such as a globally later target alignment when compared to the Neapolitan L+H*/L*+H pair, appear to influence the location of the crossover boundary between perceived questions and statements.

I also showed that by modifying the shape of the peak within the rise-fall one can affect question identification, as shown in Chapter 3. That is, when the accent peak is in the shape of a pitch plateau, more question responses are recorded relative to stimuli

with a sharp peak timed at plateau onset. On the other hand, much more similar results were obtained when the plateau stimuli scores were compared to those for stimuli with sharp peak timed at plateau offset. I interpreted this result as an effect of peak shape on perceived target location. That is, the perceived target location for the LH pitch rise appears to be closer to plateau offset than onset. This phenomenon, which was replicated for stimuli used in the experiment presented in Chapter 5, cannot be explained in terms of language-specific constraints on prosodic parsing, since it was also found for a different language group. Namely, as shown in Chapter 4, American English listeners identified more question responses (hence, more "late peaked", L*+H accents) for plateau stimuli. What is more, also for American listeners, plateau stimuli scored like stimuli with a sharp peak timed at plateau offset. This suggests that the observed effect is psychoacoustic in nature and is potentially cross-linguistically shared.

The f_0 height manipulation obtained a similar pattern of results for American and Italian listeners. Specifically, the f_0 effect was found for extreme values, i.e., different scores were found for globally higher and globally lower stimuli. However, no difference between stimuli at a medium level and stimuli at either a lower or a higher level was reported. Overall, the temporal alignment effect was much stronger for both language groups. On the other hand, the slope manipulation produced slightly different results. Specifically, while for Neapolitans steep stimuli received scores equal to those of globally higher and globally medium stimuli, the American listeners showed a trend for steep stimuli to receive less question responses. This is difficult to account in terms of the acoustic details typical of the American L*+H, in which the f_0 excursion is quite marked.

Finally, the results of Chapter 5 suggest that syllable structure effects, which are measurable in production (see Chapter 2), do not clearly affect the recoverability of the question/statement contrast. That is, the syllable structure manipulations were not able to shift the category boundary location between perceived questions and statements. This result suggests that no look-ahead mechanism is employed when computing perceived target location. I propose, then, that question and statement tonal targets are computed relative to the left edge of the stressed syllable, so that the location of the right boundary of the stressed vowel and its duration do not affect the identification process. This can be taken to suggest that Neapolitan aligns tones relative to the left edge of the TBU. On the other hand, a category boundary shift was found when stimuli were resynthesized from either a question base or a declarative base utterance. This suggests that cues other than target alignment are employed when computing perceived pitch accent contrast (e.g., spectral tilt, spectral balance, and so forth). The nature of such cues was beyond the scope of this thesis, hence it will be object of future research.

In sum, this thesis proposes that temporal alignment, both as a production and a perception mechanism, is a crucial phonetic and phonological mechanism that cannot be ignored and that can help better define the nature of tonal targets and its denotations as objects in the "real world". This has implications for phonological theory, which will be discussed below.

6.2 Implications for intonational phonology

Since Bruce's discovery (Bruce 1977) of the alignment contrast of the HL sequence for Accent I (early alignment) and Accent II (late alignment) in Stockholm Swedish, the common understanding within linear approaches to intonational phonology has been that tones (either L or H) are mapped onto f_0 targets (connected by linear interpolation). Pierrehumbert's model of English intonation (Pierrehumbert 1980) is very similar to Bruce's, in that particular tone sequences are mapped onto f_0 targets by means of (context sensitive) phonetic implementation rules. Such targets are then interpolated. The first issue tackled in this thesis was the nature of such tonal targets. Once we determine a consistent way of measuring the exact alignment and scaling of those targets in the f_0 contour, it is desirable that the measured target's coordinates would somehow map onto the coordinates of a "perceived" target.

This issue is connected to the debate about the relationship between alignment and association. If we take alignment to mean the temporal specification of the phonological tones relative to the segmental string, mediated by the prosodic structure to which the segments are linked, it follows that the phonology of a language must be capable of describing the details of such a mapping. This is also because precise alignment contrasts are employed by listeners in order to recover pitch-accent contrast, as shown in previous works (Pierrehumbert and Steele 1987; Kohler 1987; D'Imperio and House 1997) as well as in Chapters 3, 4 and 5 of this dissertation.

Also, it appears that the nature of such "temporal" detail is both universal and language specific. We saw in Chapter 2 that Neapolitan Italians align H peaks of L*+H, on average, just 60 ms later than H peaks of L+H*, while it appears that the temporal distance of American English peaks for the analogous pitch accent contrast is much less subtle. These cross-linguistic alignment differences are reminiscent of language-specific VOT values for segmental voicing contrast. For instance, if we take the phonetic denotation of /p/ in Italian, this would correspond to a phonetic instantiation of /b/ in initial word position in English. That is because Italian lacks the positive VOT of English voiceless stops, while English lacks the negative VOT of Italian voiced stops. Therefore, the boundary between the voiced/voiceless contrast is drawn at different locations within a timing continuum for the laryngeal gesture in the two languages. Analogously, the phonetic denotation of the category L*+H in Neapolitan might be more similar to an instantiation of the American L+H*, and identified as such by American listeners, unless the phonetic realization of the target tones is sufficiently late for a L*+H to be identified. The results presented in Chapter 4 are congruent with such a view. The later crossover boundary for American English listeners might then be due to the details of the phonetic alignment contrast between L+H* and L*+H in their language. Those details are at the base of the perception contrast.

By following this reasoning, one could argue that any comparison between, say, the English L*+H and Neapolitan L*+H is spurious, since those are different phonological categories denoting different "objects" in the real world (see Pierrehumbert et al. (in press) for a general discussion about such a viewpoint). That is indeed a paradoxical situation, which might prevent us to draw any comparisons between the two languages. However,

these kinds of comparisons can still be useful in order to investigate the issue of "natural" phonological categories in a language, that is categories that arise from universal constraints on the production and the perception system.

Hence, drawing the line between which part of alignment is "phonological" and which part is "phonetic" cannot be done *a priori*, according to the position taken here. Some of the knowledge of one's own language lies in quantitative detail specifying things such as "the exact temporal location of an f_0 peak giving rise to the percept of a L*+H pitch accent in language X". Such detail can be argued to be phonetic in nature, but since it is learned, it must also reflect a process of relative abstraction. It is also true that much of what is generally believed to be part of phonetics, might be at the base of language-specific patterns, such as neutralization processes (Steriade 1997), or non-linearities in the articulatory-to-acoustics mapping (Stevens 1972), just to cite some well-known examples.

Some authors have recently defined the concepts of "alignment" and "association" as being two different phonological mechanisms. Gussenhoven (2000), for instance, defines "alignment" as the mechanism controlling the "location of morphological constituents relative to the morphological or phonological structure of the expression they are part of". Association is instead defined as the structural link between a tone and a TBU which "establishes a temporal relation" between the tone and specific segments linked to the TBU. Hence, as Gussenhoven observes, all tones must be subject to alignment constraints, but not all tones need to participate in an association relation. That is a consequence of narrowly defining association in terms of a "special" link between tone and a minimal TBU, such as the special relationship entertained by a pitch accent and the stressed syllable to which it

is associated. However, we know that also edge tones can participate in the association relationship (Pierrehumbert and Beckman 1988), though they will more often associate with a higher level phonological constituent, such as the intermediate phrase or the intonational phrase. What is more, according to recent views, some of these edge tones (i.e., phrase accents) seem to participate in a secondary association to minimal TBUs, such as stressed vowels, or morae (Gussenhoven 2000; Grice et al. 2000).

Hence, drawing a boundary between association and alignment might not be that necessary, if we believe that both phonetic-like and phonologic-like constraints on alignment figure as part of the language-specific, intrinsic knowledge of a tonal category of any kind. Here the boundary between association and alignment is therefore purposely not drawn in an *a priori* fashion. Association "is", in a sense, contrastive alignment and alignment is both a phonetic and a phonological fact.

I then propose that the phonology of Neapolitan Italian specifies both tones in a LH pitch accent to be aligned relative to the segments of the stressed syllable they belong to. That is, not only the starred tone is aligned to the stressed syllable, but also the trailing tone.[2] This alignment constraint might prevent the H of L*+H in Neapolitan to occur beyond the stressed syllable boundary.

It is also possible, though, that the actual location of the H tone depends on a secondary association of the HL- phrase tone with the last stressed syllable of a focus constituent. That is, the H of the HL- tone might force the H target of the L*+H to occur where it is by "taking over" and determining peak location within the LHL configuration. This might

[2]It has been claimed, instead, that the trailing tone of bitonal accents, at least for English, might be aligned relative to the preceding starred L tone, i.e., at a fixed distance from it (cf. Pierrehumbert (1980)).

be due to pressure to realize the HL- tone at the last stressed syllable of the intermediate phrase. Note that in American English the phrase accent appears to be realized only at the right edge of the nuclear accented word (Pierrehumbert and Beckman 1988), leaving space for the H of L*+H to be realized beyond the limits of the stressed syllable when the word contains one or more postaccentual syllables (see Figure 4.1 above). In Neapolitan Italian, instead, the H peak for the nuclear word is never aligned to occur with postaccentual syllables (see D'Imperio (in press)). In sum, the H peak of a nuclear L*+H in Neapolitan is both "associated" in the traditional sense with the stressed syllable as well as being aligned to its right-periphery. Below, I will speculate further on the concept of association as it relates to starred tones.

Also, I postulate that the phrasal HL-, which is shared by statements and questions, is associated both with the intermediate phrase and (as a secondary association) with the last stressed syllable of the same domain. An alignment constraint will determine the exact location for the end of the fall, which appears to be the right edge of the stressed foot. However, such alignment was shown consistently only by speaker MD, and not by speaker LD (see Chapter 2). More data is therefore needed in order to solve this issue.

6.3 Perception as the basis of phonological description

Phonological theory, since SPE (Chomsky and Halle 1968), has been mainly based on production features of speech, while considerably (if not totally) neglecting perception. Very recently, though, perception has started to shape our thinking about why certain phonological regularities appear in a language (Steriade 1997). The findings presented in this thesis appear to support the view that perceptual constraints on alignment must be part of the native speaker's intrinsic knowledge of her/his own language.

The results for the perception experiment presented in Chapter 3 show that for a L*+H to be identified, the "perceived" target for the rise must be late, so late that if the pitch peak has a plateau shape such target will be "pulled" towards the plateau offset. This leads me to propose that the starred tone within bitonal pitch accents of Neapolitan Italian is aligned with the first mora of the stressed syllable[3], which I take to be the most perceptually salient mora.

This alignment constraint accounts for the fact that in the stressed closed syllables of the corpus in Chapter 2 (which are bimoraic and the second mora is linked to the sonorant coda) the f_0 peak for L*+H was measured beyond the boundaries of the stressed vowel for MD, so as to allow the listener to recover the starred L within the monomoraic vowel. Further evidence comes from the fact that, if a stressed open syllable is monomoraic, such as an open syllable in antepenultimate position (D'Imperio and Rosenthall 1999), the peak of L*+H will occur just beyond the syllable boundary, so that it is not in a position for a

[3] Independent evidence for mora structure in Italian comes from lengthening under stress and foot structure, cf. D'Imperio and Rosenthall (1999).

L+H* to be parsed. In other words, the perceptual constraint allowing to contrast L+H* and L*+H "outranks" the constraint on realization of the H peak within the boundaries of the stressed syllable.

This situation can be seen in Figure 4.2, shown above, where the intonation contour for the utterance *firmiamolo?* "Let us sign it?", was produced with a L*+H on [mja]. Notice that, since the [a] is short and monomoraic, the H is reached later than expected from the findings for open syllables followed by a nasal reported in Chapter 2 (where it was found that the H peak of L*+H is aligned with the stressed vowel offset or slightly earlier). Also, the H in this utterance is not even reached within the postaccentual vowel, which is moraic, thus excluding the hypothesis of an alignment of the H with a segment dominated by a mora. The H target is instead located within the nonmoraic onset of the postaccentual syllable.

Analogously, one can propose that for the Neapolitan L+H*, the starred H might be aligned with the first mora of the stressed syllable. It is possible that listeners attend to a globally low level f_0 within the first mora in order to identify the L+H* of statements, and a high level or simply absence of a low level within the same location in order to identify the L*+H of questions. A similar view of perceived target processing was proposed in §3.3.6 and §3.4 above.

From the point of view of production, it seems that the location for the alignment of the L* of L*+H corresponds, quite stably, with the stressed vowel onset (see, for instance, Figure 4.2). If the location of the L* is what is attended to by the listeners, the shape of the peak should not matter, but we saw that this is not the case. Plateau stimuli did in fact shift question perception scores suggesting that the perpetual target for such configurations

takes into account not only the location for the beginning an/or the end of the LH rise location, but also some other location closely timed with the HL fall. I take this to suggest that tonal target perception is more dynamic in nature, and that more than a single target's alignment is taken into consideration when a specific pitch accent is identified. Also, if the starred tone is defined as the "strong", stably anchored tone of a bitonal accent, then it is interesting that recovering the location for its target is not sufficient in order to identify the entire pitch accent.

A detailed model of tonal target perception needs therefore to take into consideration dynamic factors, such as peak shape within the accent as well as the relative alignment of the entire melodic transition within the stressed syllable. An "averaging" process might then apply, as it is suggested by the psychoacoustic literature on tone perception for both speech and non-speech stimuli. In other words, the listener must compute the pitch level across the syllable in order to perceive a question or a statement as such. But this is a question for future research.

6.4 Conclusion

In conclusion, the relationship between acoustic tone targets and perceived targets for the question/statement (L+H*/L*+H) contrast in Neapolitan Italian was tested experimentally through a set of perception experiments and for two languages, i.e., Neapolitan Italian and American English. The experiments show that details of temporal alignment of target tones as well as the shape of the peak within a rising accent affect the identification of the contrast, for both languages. Also, for Italian, syllable structure detail affects acoustic target

alignment but does not modify the crossover boundary between perceived questions and statements. These findings have implications for both labeling and intonational phonology. First, since peak shape is perceptually relevant for pitch accent identification, labeling as well as speech recognition should be sensitive to that. Second, some of the effects reported appear to be cross-linguistically shared. Nevertheless, language-specific differences are also found, such as the later crossover boundary for alignment perception in the American results. Finally, I proposed that both psychoacoustic and "learned" constraints on perception must shape the intonational phonological categories of a language, and cannot be merely relegated to lower-level, quantitative effects.

APPENDIX A

ADDITIONAL LATENCY DATA

Graphical presentation of additional latency data, which were left out in the main body of the dissertation, are included below.

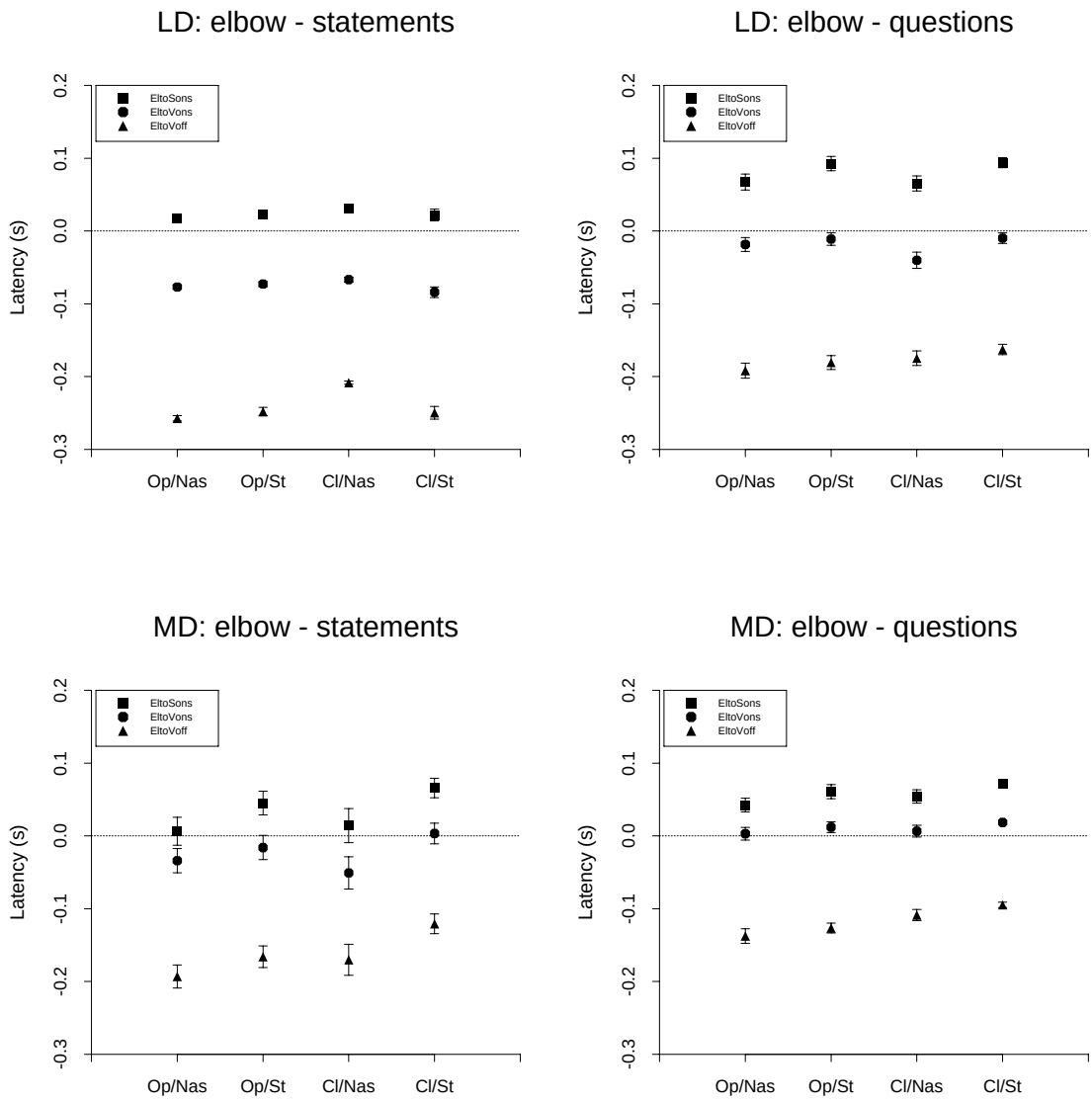

Figure A.1: Mean L1 (elbow) latency from syllable onset (EltoSons), vowel onset (EltoVons) and vowel offset (EltoVoff) for LD (upper) and MD (lower) (Op/Nas = open syll., nasal; Op/St = open syll., stop; Cl/Nas = closed syll., nasal; Cl/St = closed syll., stop). The dotted line is the reference point for each latency measurement. Standard error is indicated by vertical bars.

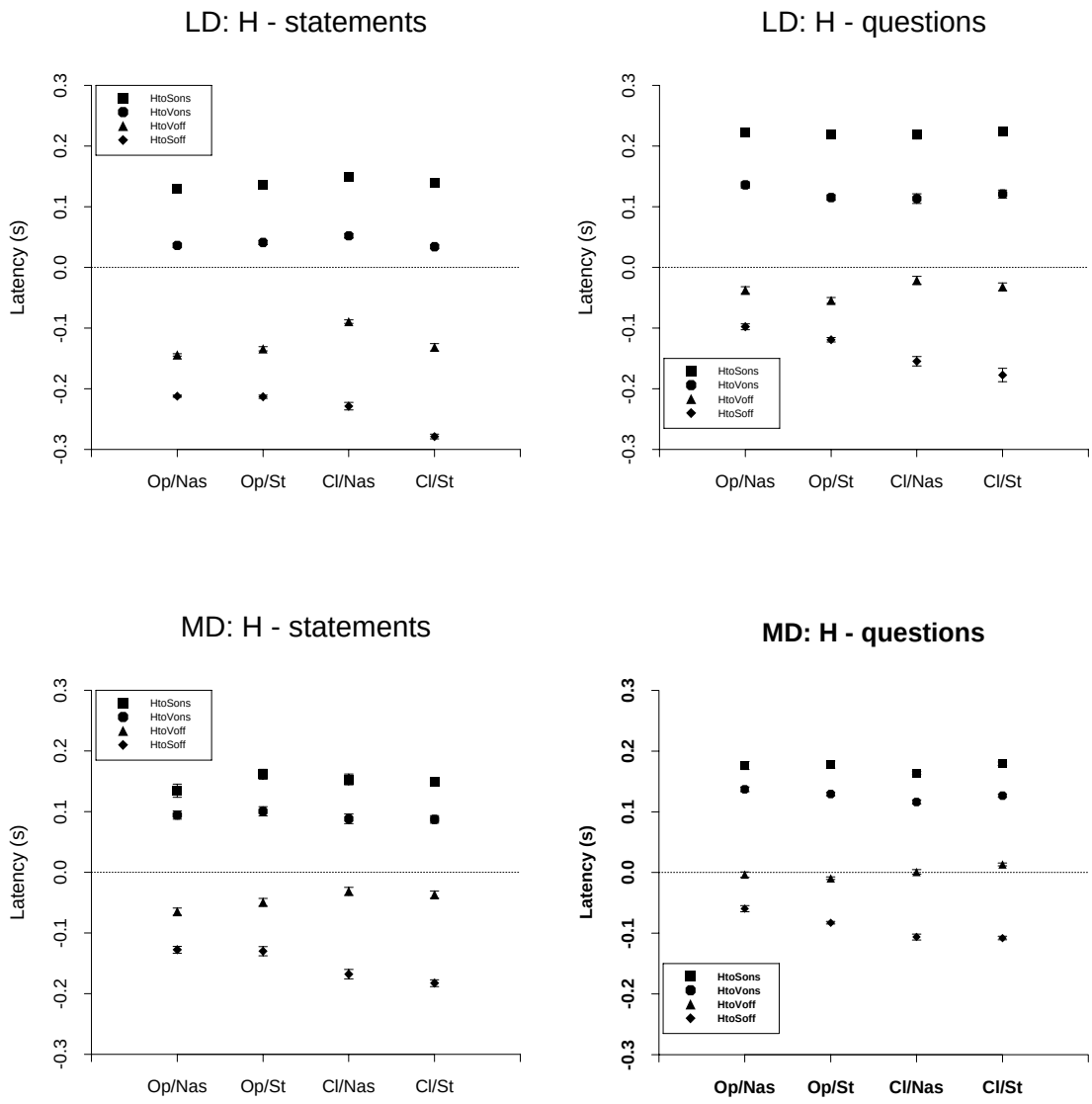

Figure A.2: Mean H (F0max) latency from vowel onset (HtoVons), vowel offset (HtoVoff) and syllable offset (HtoSoff) for LD (upper) and MD (lower) (Op/Nas = open syll., nasal; Op/St = open syll., stop; Cl/Nas = closed syll., nasal; Cl/St = closed syll., stop). The dotted line is the reference point for each latency measurement. Standard error is indicated by vertical bars.

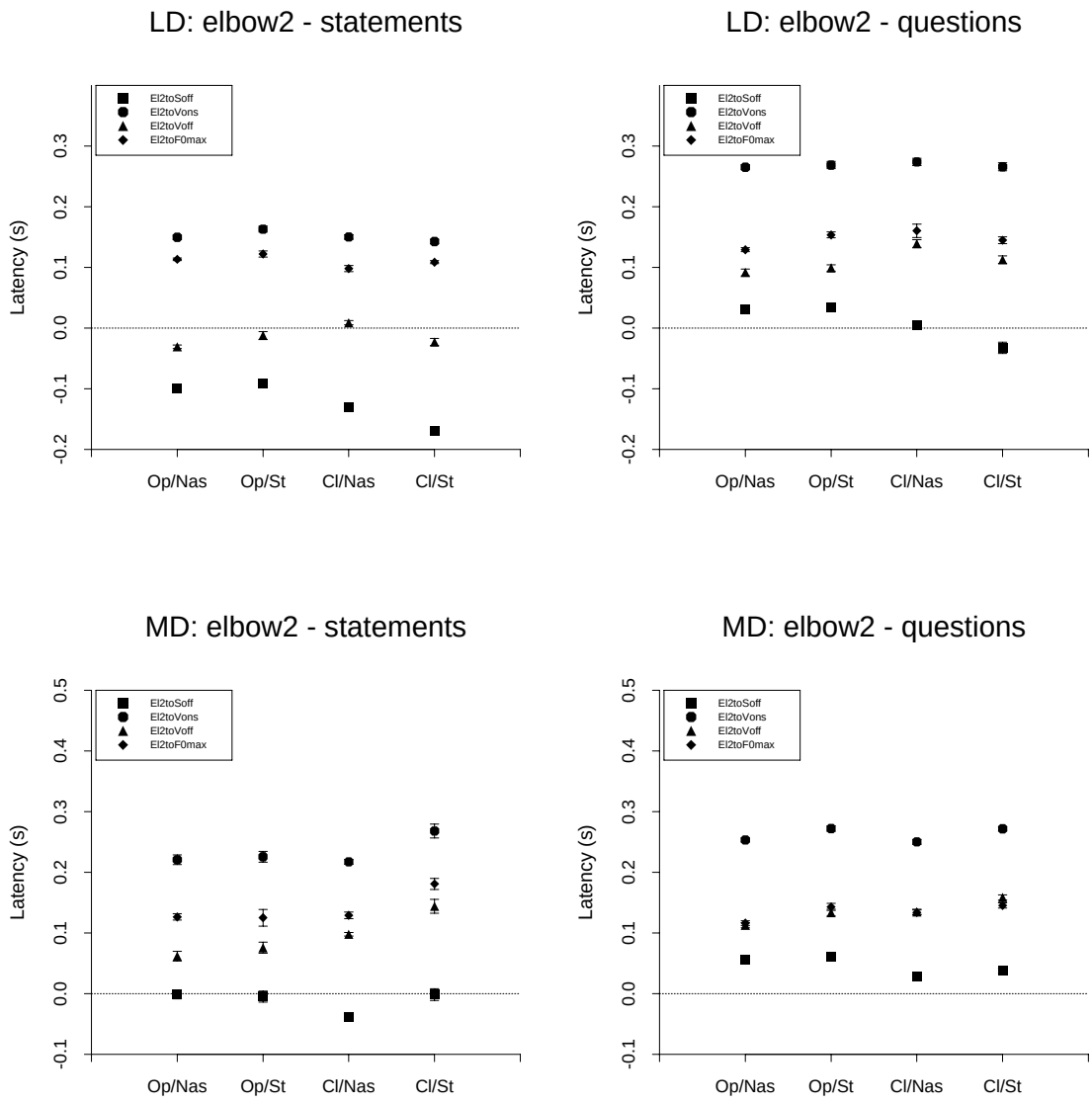

Figure A.3: Mean L2 (elbow2) latency from vowel offset (El2toVoff), vowel onset (El2toVons), syllable offset (El2toSoff) and F0max (El2toF0max) for LD (upper) and MD (lower) (Op/Nas = open syll., nasal; Op/St = open syll., stop; Cl/Nas = closed syll., nasal; Cl/St = closed syll., stop). The dotted line is the reference point for each latency measurement. Standard error is indicated by vertical bars.

BIBLIOGRAPHY

Ainsworth-Darnell, K. and M. D'Imperio (Eds.) (1997). *The Ohio State University Working Papers in Linguistics – Papers from the Linguistics Laboratory*, Volume 50. OSU.

Anderson, M. D., J. B. Pierrehumbert, and M. Y. Liberman (1984). Synthesis by rule of English intonation patterns. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, New York, pp. 2.8.2–2.8.4.

Arvaniti, A., D. Ladd, and I. Mennen (1998). Stability of tonal alignment: The case of Greek prenuclear accents. *Journal of Phonetics 26*, 3–25.

Arvaniti, A., D. R. Ladd, and I. Mennen (2000). What is a starred tone? Evidence from Greek. See Broe and Pierrehumbert (2000).

Avesani, C. (1990). A contribution to the synthesis of Italian intonation. In *Proceedings of the International Conference on Spoken Language Processing*, Kobe, Japan, pp. 833–36.

Avesani, C. (1995). ToBIt: un sistema di trascrizione per l'intonazione italiana. In *Atti delle 5e Giornate di Studio del Gruppo di Fonetica Sperimentale (A.I.A.)*, Povo (TN), Italy, pp. 85–98.

Beckman, M. E. (1996). The parsing of prosody. *Language and Cognitive Processes 11*(1/2), 17–67.

Beckman, M. E. and G. Ayers Elam (1994). Guidelines for ToBI labelling, vers. 3.0, March 1997. Manuscript and accompanying speech materials. The Ohio State University (http://ling.ohio-state.edu/Phonetics/EToBI/homepage.html).

Beckman, M. E. and J. Edwards (1994). Articulatory evidence for differentiating stress categories. In P. Keating (Ed.), *Papers in Laboratory Phonology III: Phonological Structure and Phonetic Form*, pp. 7–33. Cambridge: CUP Press.

Beckman, M. E. and J. B. Pierrehumbert (1986). Intonational Structure in Japanese and English. *Phonology Yearbook 3*, 255–310.

Beddor, P. and T. Gottfried (1995). Methodological issues in cross-language speech perception research with adults. In W. Strange (Ed.), *Speech Perception and Linguistic Experience*, pp. 207–232. Baltimore, MD: York Press.

Bohn, O.-S. (1995). Cross-language speech perception in adults: First language transfer doesn't tell it all. In W. Strange (Ed.), *Speech Perception and Linguistic Experience*, pp. 279–303. Baltimore, MD: York Press.

Bolinger, D. (1958). A theory of pitch accent in English. *Word 14*, 109–149.

Bolinger, D. (1961). Contrastive accent and contrastive stress. *Language 37*, 83–96.

Bond, Z., V. Stockmal, and D. Muljani (1998). Learning to identify a foreign language. *Language Sciences 20*(4), 353–367.

Botinis, A., G. Kouroupetroglou, and G. Carayiannis (Eds.) (1997). *Intonation: Theory, Models and Applications. Proceedings of an ESCA Workshop*, Athens, Greece.

Broe, M. and J. B. Pierrehumbert (Eds.) (2000). *Papers in Laboratory Phonology V: Acquisition and the Lexicon*. Cambridge: Cambridge University Press.

Bruce, G. (1977). *Swedish Word Accents in Sentence Perspective*. Lund: Gleerups.

Bruce, G. (1982). Developing the Swedish intonation model. In *Working Papers 22*, pp. 51–116. Lund University, Department of Linguistics and Phonetics.

Bruce, G. (1983). Accentuation and timing in Swedish. *Folia Linguistica 8*, 221–238.

Bruce, G. (1987). How floating is sentence accent? In K. Gregersen and H. Basbøl (Eds.), *Nordic Prosody IV*, pp. 41–49. Odense: Odense University Press.

Bruce, G. (1990). Alignment and composition of tonal accents: Comments on Silverman and Pierrehumbert's paper. See Kingston and Beckman (1990), pp. 107–114.

Bruce, G. and E. Gårding (1978). A prosodic typology for Swedish dialects. In E. Gårding, G. Bruce, and R. Bannert (Eds.), *Nordic Prosody*, pp. 219–228. Lund: Gleerups.

Caldognetto, E. M. and E. Fava (1972). Studio sperimentale delle caratteristiche elettroacustiche dell'enfasi su sintagmi in Italiano. In *Atti del VI Congresso Internazionale di Studi 'Fenomeni morfologici e sintattici nell'italiano contemporaneo'*, pp. 441–456. Roma: Bulzoni.

Caputo, M. R. (1994). L'intonazione delle domande sí–no in un campione di italiano parlato. In *Atti delle IV Giornate di Studio del Gruppo di Fonetica Sperimentale (A.I.A.)*, Torino, Italy, pp. 9–18.

Caputo, M. R. (1996). Presupposizione, fuoco, modalità e schemi melodici. In *Atti del XXIV Convegno Nazionale dell'Associazione Italiana di Acustica (A.I.A.)*, Trento, Italy, pp. 49–54.

Caputo, M. R. and M. D'Imperio (1995). Verso un possibile sistema di trascrizione prosodica dell'italiano: cenni preliminari. In *Atti delle IV Giornate di Studio del Gruppo di Fonetica Sperimentale (A.I.A.)*, Povo (TN), Italy, pp. 71–83.

Caspers, J. and V. J. van Heuven (1993). Effects of time pressure on the phonetic realisation of the Dutch accent-lending pitch rise and fall. *Phonetica 50*, 161–171.

Chomsky, N. and M. Halle (1968). *The Sound Pattern of English*. New York: Harper and Row.

Cruttenden, A. (1993). The de-accenting and re-accenting of repeated lexical items. In *Working Papers 41*, Volume 41, pp. 16–19. Lund University, Department of Linguistics and Phonetics.

D'Imperio, M. (1995). Timing differences between prenuclear and nuclear pitch accents in Italian. *JASA 98*(5), 2894.

D'Imperio, M. (1996). Caratteristiche di timing degli accenti nucleari in parlato italiano letto. In *Atti del XXIV Convegno Nazionale dell'Associazione Italiana di Acustica*, Trento, Italy, pp. 55–60.

D'Imperio, M. (1997a). Breadth of focus, modality and prominence perception in Neapolitan Italian. See Ainsworth-Darnell and D'Imperio (1997), pp. 19–39.

D'Imperio, M. (1997b). Narrow focus and focal accent in the Neapolitan variety of Italian. See Botinis, Kouroupetroglou, and Carayiannis (1997), pp. 87–90.

D'Imperio, M. (1998a). Narrow focus and focal accent in the Neapolitan variety of Italian. Ms. The Ohio State University.

D'Imperio, M. (1998b). Sul ruolo dei correlati acustici della prominenza nella percezione della varietà napoletana di italiano. In *Proceedings of the "VIIIe Giornate di Fonetica Sperimentale"*, Pisa, Italy, pp. 43–54. G.F.S. (A.I.A.).

D'Imperio, M. (1999). Tonal structure and pitch targets in Italian focus constituents. In J. Ohala (Ed.), *Proceedings of the 14th International Congress of Phonetic Sciences*, Volume 3, San Francisco, USA, pp. 1757–1760.

D'Imperio, M. (in press). Focus and tonal structure in Neapolitan Italian. *Speech Communication*.

D'Imperio, M. and D. House (1997). Perception of questions and statements in Neapolitan Italian. In G. Kokkinakis, N. Fakotakis, and E. Dermatas (Eds.), *Proceedings of Eurospeech'97*, Volume 1, Rhodes, Greece, pp. 251–254.

D'Imperio, M. and S. Rosenthall (1999). Phonetics and phonology of main stress in Italian. *Phonology 16*(1), 1–28.

Frota, S. (1997). Association, alignment, and meaning: the tonal sequence HL and focus in European Portuguese. See Botinis, Kouroupetroglou, and Carayiannis (1997), pp. 127–130.

Frota, S. (2000a). *Prosody and Focus in European Portuguese. Phonological Phrasing and Intonation (PhD Dissertation, University of Lisbon)*. New York: Garland.

Frota, S. (2000b). Tonal association and target alignment: implications for intonation theory. Paper presented at Laboratory Phonology VII, Nijmegen, July 2000.

Goldsmith, J. (1976). *Autosegmental Phonology*. Ph. D. thesis, MIT (also 1979, published by Garland Press).

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

Grice, M. (1991). The intonation of interrogation in two varieties of Sicilian Italian. In *Proceedings of the XIIth International Congress of Phonetic Sciences*, Volume 5, Aix-en-Provence, France, pp. 210–213.

Grice, M. (1995a). *The intonation of interrogation in Palermo Italian: implications for intonation theory*. Niemeyer, L.A. series.

Grice, M. (1995b). Leading tones and downstep in English. *Phonology 12*, 183–233.

Grice, M., M. D'Imperio, M. Savino, and C. Avesani (in press). Towards a strategy for ToBI labelling varieties of Italian. In S.-A. Jun (Ed.), *Prosodic Typology and Transcription: A Unified Approach*. (Collection of papers from the ICPhS 1999 satellite workshop on "Intonation: Models and ToBI Labeling". San Francisco, California).

Grice, M., D. Ladd, and A. Arvaniti (2000). On the place of phrase accents in intonational phonology. *Phonology 17*(2), 145–187.

Grice, M. and M. Savino (1995). Low tone versus 'sag' in Bari Italian intonation; a perceptual experiment. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Volume 4, Stockholm, Sweden, pp. 658–661.

Grice, M. and M. Savino (1997). Can pitch accent type convey information status in yes–no questions? In *Proceedings of a Workshop Sponsored by the Association for Computational Linguistics*, Madrid, Spain, pp. 29–38.

Grønnum, N. (1991). Prosodic parameters in a variety of regional Danish standard languages, with a view towards Swedish and German. *Phonetica 47*, 188–214.

Gussenhoven, C. (1984). A semantic analysis of the nuclear tones of English. In *On the Grammar and Semantics of Sentence Accents*, pp. 193–265. Dordrecht: Foris.

Gussenhoven, C. (2000). The boundary tones are coming: On the nonperipheral realisation of boundary tones. See Broe and Pierrehumbert (2000), pp. 132–151.

Hadding-Koch, K. and M. Studdert-Kennedy (1976). An experimental study of some intonation contours. In D. Fry (Ed.), *Acoustic Phonetics*, pp. 431–441. Cambridge: C.U.P.

Herman, R. (1997). Syntactically-governed accentuation in Balinese. See Ainsworth-Darnell and D'Imperio (1997), pp. 69–99.

Hermes, D. (1997). Timing of pitch movements and accentuation of syllables in dutch. *The Journal of the Acoustical Society of America 102*(4), 2390–2402.

Hermes, D., F. Beaugendre, and D. House (1997). Temporal alignment of accentuation boundaries in Dutch. See Botinis, Kouroupetroglou, and Carayiannis (1997), pp. 170–180.

Hirst, D. (1988). Tonal units as constituents of prosodic structure: The evidence from English and Frech intonation. In H. van der Hulst and N. Smith (Eds.), *Autosegmental Studies on Pitch Accent*, pp. 151–165. Dordrecht: Foris.

House, D. (1990). *Tonal Perception in Speech*. Lund: Lund University Press.

House, D. (1995). The influence of silence on perceiving the preceding tonal contour. In K. E. . P. Branderud (Ed.), *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Volume 1, Stockholm, Sweden, pp. 122–125.

House, D. (1996). Differential perception of tonal contours through the syllable. In *Proceedings of the International Conference on Spoken Language Processing*.

Jakobson, R., G. Fant, and M. Halle (1952). *Preliminaries to Speech Analysis*. Cambridge, MA: The MIT Press.

Johnson, K., E. Flemming, and R. Wright (1993). The hyperspace effect: phonetic targets are hyperarticulated. *Language 69*(3), 505–528.

Jun, S.-A. and M. Oh (1996). A prosodic analysis of three types of wh-phrases in Korean. *Language and Speech 39*, 37–61.

Kingston, J. and M. E. Beckman (Eds.) (1990). *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*. Cambridge: Cambridge University Press.

Kohler, K. (1987). Categorical pitch perception. In *Proceedings of the XIth International Congress of Phonetic Sciences*, Volume 5, Tallin, Estonia, pp. 331–333.

Kohler, K. J. (1990). Macro and micro F0 in the synthesis of intonation. See Kingston and Beckman (1990), pp. 115–138.

Kori, S. and E. Farnetani (1983). Acoustic manifestation of focus in Italian. *Quaderni del Centro di Studio per le Ricerche di Fonetica 2*, 323–328.

Ladd, D. R. (1980). *The Structure of Intonational Meaning*. Bloomington: Indiana University Press.

Ladd, D. R. (1983). Phonological features of intonational peaks. *Language 59*, 721–759.

Ladd, D. R. (1996). *Intonational Phonology*. Cambridge: Cambridge University Press.

Ladd, D. R., D. Faulkner, H. Faulkner, and A. Schepman (1999). Constant "segmental anchoring" of f_0 movements under changes in speech rate. *The Journal of the Acoustical Society of America 106*(3), 1543–1554.

Ladd, D. R., I. Mennen, and A. Schepman (2000). Phonological conditioning of peak alignment in rising pitch accents in Dutch. *The Journal of the Acoustical Society of America 107*(5), 2685–2695.

Leben, W. (1973). *Suprasegmental Phonology*. Ph. D. thesis, MIT.

Liberman, M. (1975). *The Intonational System of English*. Ph. D. thesis, MIT.

Liberman, M. and A. Prince (1977). On stress and linguistic rhythm. *Linguistic Inquiry 8*, 249–336.

Lindblom, B. E. F. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America 35*(11), 1773–1781.

Lindblom, B. E. F. and M. Studdert-Kennedy (1967). On the role of formant transitions on vowel recognition. *The Journal of the Acoustical Society of America 42*(4), 830–843.

Madden, J. P. and K. M. Fire (1997). Detection and discrimination of frequency glides as a function of direction, duration, frequency span, and center frequency. *The Journal of the Acoustical Society of America 102*, 2920–2924.

Maekawa, K. (1991). Perception of intonational characteristics of wh- and non-wh questions in tokio japanese. In *Proceedings of the XIIth International Congress of Phonetic Sciences*, Aix-en-Provence, France, pp. 202–205.

McCarthy, J. and A. Prince (1993). Generalized alignment. *Yearbook of Morphology*.

Moon, S.-J. and B. Lindblom (1994). Interaction between duration, context, and speaking style in English stressed vowels. *The Journal of the Acoustical Society of America 96*(1), 40–55.

Moore, B. C. J. (1989). *An Introduction to the Psychology of Hearing*. London: Academic Press.

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

Nábeläk, I. V. and I. J. Hirsh (1969). On the discrimination of frequency transitions. *The Journal of the Acoustical Society of America 45*, 1510–1519.

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

Nábeläk, I. V. and A. Ovchinnikov (1997). Perception of nonlinear and linear formant trajectories. *The Journal of the Acoustical Society of America 101*(1), 488–497.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America 85*(5), 2088–2113.

Nordmark, J. O. (1968). Mechanisms of frequency discrimination. *The Journal of the Acoustical Society of America 44*, 1533–1540.

Peterson, G. and I. Lehiste (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America 32*, 693–703.

Pierrehumbert, J. B. (1980). *The Phonology and Phonetics of English intonation*. Ph. D. thesis, MIT.

Pierrehumbert, J. B. (1993). Alignment and prosodic heads. In *Proceedings of ESCOL '93*, Cornell, New York, pp. 268–286.

Pierrehumbert, J. B., M. Beckman, and D. Ladd (in press). Conceptual foundations of phonology as a laboratory science. In N. Burton-Roberts, P. Carr, and G. Docherty (Eds.), *Phonological Knowledge: Its Nature and Status*. Cambridge: Cambridge University Press.

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

Pierrehumbert, J. B. and S. Steele (1987). How many rise-fall-rise contours? In *Proceedings of the XIth International Congress of Phonetic Sciences*, Tallin, Estonia.

Pierrehumbert, J. B. and S. Steele (1989). Categories of tonal alignment in English. *Phonetica 46*, 181–196.

Pisoni, D., R. Aslin, A. Perey, and H. B.L. (1982). Some Effects of Laboratory Ttraining on Identification and Discrimination of Voicing Contrasts in Stop Consonants. *Journal of Experimental Psychology: Human Perception and Performance 8*(2), 297–314.

Pompino-Marshall, B. (1989). On the psychoacoustic nature of the P-center phenomenon. *Journal of Phonetics 17*, 175–192.

Prieto, P., J. P. H. van Santen, and J. Hirschberg (1995). Tonal alignment patterns in Spanish. *Journal of Phonetics 23*, 429–451.

Prince, A. and P. Smolensky (1993). Optimality theory: Constraint interaction in generative grammar. Ms. Rutgers University and University of Colorado, Boulder.

Rietveld, T. and C. Gussenhoven (1995). Aligning pitch targets in speech synthesis: effects of syllable structure. *Journal of Phonetics 23*, 375–385.

Ritsma, R. J. (1965). Pitch discrimination and frequency discrimination. In *Proceedings of the fifth International Congress on Acoustics*, pp. B22.

Ritsma, R. J. (1967). Frequencies dominant in the perception of the pitch of complex sounds. *The Journal of the Acoustical Society of America 42*, 191–198.

Rossi, M. (1971). Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole. *Phonetica 23*, 1–33.

Rossi, M. (1978). La perception des glissandos descendants dans les contours prosodiques. *Phonetica 35*, 11–40.

Schouten, J. F. (1940). The residue and the mechanism of hearing. In *Proceedings of the Koninklijke Akademie van Wetenschappen 43*, pp. 991–999.

Schouten, M. E. H. and W. J. M. Peeters (1995). Matching frequency glides with steady tones. Technical Report OTS-WP-FON-95-002, Research Institute for Language and Speech, Utrecht University.

Schouten, M. E. H. and W. J. M. Peeters (1998). Searching for a time window for timbre: Dynamic tones and dinamyc spectral profiles. Technical report, Utrecht Institute of Linguistics.

Silverman, K. (1987). *The structure and processing of fundamental frequency contours*. Ph. D. thesis, University of Cambridge.

Silverman, K. and J. B. Pierrehumbert (1990). The timing of prenuclear high accents in English. See Kingston and Beckman (1990), pp. 71–106.

Sobrero, A. (Ed.) (1993). *Introduzione all'italiano contemporaneo – Le strutture*. Roma–Bari: Laterza.

Steele, S. (1986). Nuclear accent F0 peak location: Effects of rate, vowel and number of following syllables. *The Journal of the Acoustical Society of America 80*, s51.

Steriade, D. (1997). Phonetics in phonology: the case of laryngeal neutralization. Ms. UCLA, Los Angeles.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. D. Jr. and P. B. Denes (Eds.), *Human Communication: A Unified View*, pp. 51–66. New York: McGraw-Hill.

Strange, W. (1995). Cross-language studies of speech perception. A historical review. In W. Strange (Ed.), *Speech Perception and Linguistic Experience*, pp. 3–45. Baltimore, MD: York Press.

Studdert-Kennedy, M. and K. Hadding-Koch (1973). Auditory and linguistic processes in the perception of intonation contours. *Language and Speech 16*, 293–313.

't Hart, J. (1979). Relations between perceptual and physical aspects of intonation. *Annali della Scuola Normale Superiore di Pisa III–IX*, 367–379.

't Hart, J. (1981). Differential sensitivity to pitch distance. *The Journal of the Acoustical Society of America 69*, 811–821.

't Hart, J. (1991). F_0 stylization in speech: Straight lines versus parabolas. *The Journal of the Acoustical Society of America 90*(6), 3368–3370.

't Hart, J., R. Collier, and A. Cohen (1990). *A perceptual study of intonation: An experimental-phonetic approach*. Cambridge: Cambridge University Press.

Trubetzkoy, N. (1939). *Grundzüge der Phonologie*. Traveaux du Cercle Linguistique de Prague VII.

van Santen, J. P. H. and J. Hirschberg (1994). Segmental effects on timing and height of pitch contours. In *Proceedings of the International Conference on Spoken Language Processing*, Volume 2, Yokohama, Japan, pp. 719–722.

van Son, R. J. (1993). Vowel perception: a closer look at the literature. In *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam*, Volume 17, pp. 33–64.

Verhoeven, J. (1994). The discrimination of pitch movement alignment in Dutch. *Journal of Phonetics 22*, 65–85.

Ward, G. and J. Hirschberg (1985). Implicating uncertainty: The pragmatics of fall-rise intonation. *Language 61*, 747–776.

Ward, G. and J. Hirschberg (1988). Intonation and propositional attitude: the pragmatics of L*+H L H%. In *Proceedings of the Fifth Eastern States Conference on Linguistics*, University of Pennsylvania, Philadelphia, PA, pp. 512–522.

Woods, A., P. Fletcher, and A. Hughes (1986). *Statistics in Language Studies*. Cambridge: Cambridge University Press.