UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova
Dipartimento di Studi Linguistici e Letterari (DiSLL)

Scuola di Dottorato di Ricerca in Scienze Linguistiche, Filologiche e Letterarie
Indirizzo: Linguistica, Lingue Classiche e Moderne
XXVI Ciclo

The Phonetic Realization of Narrow Focus in English L1 and L2. Data from Production and Perception

Direttore della Scuola: Ch.ma Prof.ssa ROSANNA BENACCHIO
Coordinatore d'indirizzo: Ch.ma Prof.ssa CARMEN CASTILLO PEÑA
Supervisore: Ch.ma Prof.ssa MARIA GRAZIA BUSÀ

Dottorando: LUCA ROGNONI

Contents

Acknowledgements

First of all I want to express my deepest gratitude to Professor Maria Grazia Busà, my supervisor and guide through these challenging and rewarding path. She gave me a great opportunity to grow as a researcher and as a man, and I hope that I have at least partially repaid her constant support with my hard work and dedication.

I also want to thank the *Fondazione Cassa di Risparmio di Padova e Rovigo*, which fully funded my Ph.D. The generous scholarship awarded by the *Fondazione* also allowed me to spend a period of research abroad and to attend to international conferences, where I could present my research and be inspired by the works of the leading researchers in my field.

An important phase of my Ph.D. was represented by the period of research that I spent as a visiting student at the Phonetics Lab of the University of Leiden (Netherlands). In particular, I want to express my gratitude to Professor Vincent Van Heuven, who welcomed me to the lab and gave me valuable input for my research, and to Jos Pacilly, who has been always patiently available to discuss issues dealing with the technical aspects of phonetic research, from recording to scripting.

During these three years I had the chance to meet many fellow Ph.D. students. A few of them have also become friends, and I want to thank them individually for their help and sympathy. The first is Martina Urbani, fellow student at the Language and Communication Lab (LCL) of the University of Padua, who led the way as a big sister in the path towards the Ph.D.

The second is Rosario Signorello, pride of Italy and Sicily throughout the world, who taught me how to use *LimeSurvey* and inspired me with his keen enthusiasm. The third is Joaquín Atria, who helped me to recruit the English native speakers required for my research and who welcomed me to record them at UCL. Thanks a lot, my friends, I hope our paths cross again soon.

I left for the end the most important persons in my life: my family and friends. To name a few: Diletta, Franco, Nonna Ermanna, Roberto, Sabina, Michel, Lucia, Cisco, Giulio, Laura, Fed, Kat... Without you guys, I would be lost.

Finally, thanks to Carla, who makes my rainy days sunny and my sunny days flawless.

A tutti voi, grazie.

L

Abstract

The typological differences between the two languages are reflected in the strategies adopted to mark sentence-level prominence. While English mark focus by modulating prosodic parameters (namely, pitch, duration and intensity), Italian normally recurs to word order strategies, benefitting from the freer word order admitted by its syntax. This study is aimed to investigate the acquisition of the prosodic marking of narrow non-contrastive focus by Italian speakers of English L2.

This study was mainly aimed at: (a) determining and comparing the prosodic cues used by English native speakers and Italian speakers of English L2 when marking narrow focus; (b) verifying if the Italian speakers are able to acquire the English prosodic strategies in focus marking as a function of their competence in English, progressively avoiding the focus marking strategies that characterize their L1 in favor of more native-like solutions; (c) investigating the phenomenon not only at the production level, but also from the point of view of perception. Consequently, this work is composed by a production and a perception study.

The production study consisted in the acoustic analysis of native and non-native productions. The speech data were collected using a semi-spontaneous method, where speakers recorded a set of short sentences as replies to *wh*-questions, with the aim of eliciting sentences presenting narrow focus on subject or on verb. Three groups of speakers were recorded: English native speakers NS), Italian native speakers with a higher competence in English

L2 (NNS1), and Italian native speakers with a lower competence in English L2 (NNS2). A similar set of Italian L1 sentences was also elicited from the Italian speakers.

The acoustical analysis was performed at sentence and word level, and it was mainly based on the measurement of fundamental frequency and duration. The results confirmed that English native speakers mark narrow focus mainly by modulating pitch. NNS1 showed a progress towards the target model, by implementing an active use of pitch, although not perfectly matching with the native one. Finally, NNS2 were not able to mark focus with the use of prosodic parameters. The analysis of the Italian L1 data set suggested that in Italian narrow non-contrastive focus is not marked prosodically. Not even duration, which in Italian is the prosodic cue normally used to mark prominence at word level seems to play a role in signaling prominence at sentence level.

The perception study was designed to verify whether the differences shown by the acoustical measurements could also have an impact on the listeners' perception. Two perception tests were designed, based on a two-alternative forced-choice paradigm, where listeners were asked to identify narrow focus by guessing the *wh-* question that had triggered each sentence.

Experiment 1 presented natural sentences to two groups of listeners: 22 British native speakers and 22 Italian native listeners. The Italian native listeners were also presented with an extra set of stimuli, consisting of the Italian L1 data set. The results of Experiment 1 showed that English native listeners could correctly identify narrow focus even without extra contextual information. This happened for NS and NNS1, whereas the listeners could not recognize focus in the productions by NNS2. The Italian listeners could also detect focus well above chance level in the productions by NS. However, they failed to identify focus in the productions by NNS1 and NNS2. As for the Italian L1 data set, the Italian listeners failed to distinguish narrow focus, providing perceptual evidence to the hypothesis that Italians do not mark

narrow focus by prosody.

Experiment 2 was designed to investigate the effect of the differences in pitch modulation on the correct detection of narrow focus by English native listeners. In this case, the productions of the speakers were acoustically manipulated. The participants were 20 British English native speakers. In general, the results of Experiment 2 confirmed that pitch plays an important role in the recognition of narrow focus also from the perceptual point of view. This is particularly true for NS productions, while the listeners could not successfully identify focus in the modified non-native productions. The results of the production study and the perception study converged in showing that in English pitch plays an important role in the production and perception of narrow non-contrastive focus. As for non-native productions, NNS1 could approach the native model to a certain extent by modulating F_0. From the perceptual point of view, their productions were effective enough to be successfully understood by English native listeners. In contrast, NNS2 had not managed to adopt the strategies of English, showing a poor prosodic characterization of the constituent in focus. As a consequence, the listeners could not identify focus in the NNS2 productions.

These findings are particularly interesting not only for research in L2 phonetics, but also for their implications for language instruction, where prosody has only recently started to be studied and taught with renewed interest and momentum.

Sommario (Italian Abstract)

La differenza tipologica tra l'italiano e l'inglese si riflette nelle strategie adottate per segnalare il focus dal punto di vista fonetico. Mentre in inglese è possibile marcare il focus utilizzando solo indici prosodici (altezza tonale, durata e intensità), in italiano si ricorre più spesso a strategie sintattiche, traendo beneficio dal più libero ordine delle parole ammesso dalla grammatica. Questa tesi si propone di investigare la realizzazione fonetica del focus ristretto di tipo non-contrastivo da parte di parlanti inglese L1 e L2.

In particolare, il presente lavoro di ricerca si pone l'obiettivo di: (a) determinare e confrontare quali sono gli indici prosodici utilizzati da parlanti nativi anglofoni e da parlanti italiani di inglese L2 per segnalare la posizione del focus ristretto; (b) verificare se i parlanti italiani siano in grado di acquisire le strategie applicate dai parlanti nativi anglofoni in funzione della loro competenza in inglese L2, abbandonando progressivamente le strategie trasferite da L1 in favore di soluzioni più vicine a quelle adottate dai parlanti nativi anglofoni; (c) investigare il fenomeno non solo dal punto di vista della produzione, ma anche sul versante della percezione degli ascoltatori.

I primi tre capitoli della tesi sono dedicati all'introduzione del problema, alla sua inquadratura nel quadro teorico di riferimento (la fonetica acustica sperimentale) e alla rassegna critica della letteratura più rilevante. In questi capitoli introduttivi sono inoltre presentate le principali teorie dell'acquisizione della pronuncia in L2 e i principali problemi metodologici

connessi alla ricerca sperimentale su L2, con particolare attenzione all'ambito della prosodia. Il Capitolo 4 presenta le metodologie e i risultati di quattro studi pilota condotti dall'autore di questa tesi, con il duplice scopo di ottenere dati empirici sulla prosodia dell'inglese parlato dagli italiani e di verificare l'efficacia di diversi metodi di manipolazione del segnale per la preparazione di stimoli sperimentali.

La parte centrale della tesi è rappresentata da uno studio di produzione (Capitoli 5 e 6) e da uno studio di percezione (Capitoli 7 e 8). Lo studio di produzione consiste nell'analisi acustica di brevi frasi realizzate da parlanti inglese L1 e L2, raccolte in modo semi-spontaneo utilizzando un protocollo di registrazione in cui le frasi sono state elicitate come risposte a interrogative parziali (domande wh), in modo da stimolare la realizzazione di frasi con focus ristretto sul soggetto o sul predicato verbale. Sono stati registrati tre gruppi di parlanti: parlanti nativi anglofoni (NS), parlanti italiani con livello di inglese L2 avanzato (NNS1) parlanti italiani con livello di inglese L2 elementare (NNS2). I parlanti italiani hanno anche registrato un set di frasi in italiano dalla struttura simile a quella inglese.

Basandosi sui risultati riportati in studi precedenti (Cooper et al. 1985; Xu & Xu 2005; Breen et al. 2010), si è ipotizzato che i NS segnalassero il focus utilizzando indici prosodici, mediante significativi cambiamenti a livello di altezza tonale, durata e intensità. Nel caso dei parlanti inglese L2, si è ipotizzato che i parlanti NNS1 mostrino un significativo avvicinamento al modello dei parlanti nativi nel fare proprie le strategie prosodiche di segnalazione di focus. D'altro canto, si è ipotizzato che i parlanti NNS2 non riescano a usare la prosodia alla maniera dei nativi anglofoni, ricorrendo alle strategie proprie dell'italiano.

L'analisi acustica è stata effettuata a livello di frasi e parole, e si è focalizzata principalmente sulla misurazione della frequenza fondamentale (indice fonetico dell'altezza tonale) e della durata. I risultati confermano le ipotesi, mostrando che i parlanti NS segnalano la posizione del focus ristretto

principalmente con la modulazione dell'altezza tonale, mentre i parlanti NNS1 mostrano un avvicinamento al modello dei parlanti nativi, utilizzando in modo attivo l'altezza tonale come strumento per segnalare il focus, anche se in modo non del tutto consono al modello dei parlanti inglese L1. I parlanti NNS2, invece, non sembrano in grado di differenziare le loro produzioni sulla base degli indici fonetici analizzati. Per quanto riguarda l'analisi del set di frasi in italiano L1, l'analisi acustica ha mostrato che quando parlano la loro L1, gli italiani non marcano il focus con indici prosodici. La durata, che è l'indice acustico normalmente usato in italiano per marcare la prominenza a livello di parola, non sembra giocare un ruolo nel segnalare la prominenza a livello di frase.

I risultati dello studio di produzione hanno fornito le indicazioni per la creazione dello studio di percezione, con lo scopo di verificare se le differenze trovate nei risultati dell'analisi acustica trovassero un correlato nella percezione. Sono stati quindi creati due esperimenti percettivi, basati entrambi su un modello di risposta a scelta obbligata tra due alternative, in cui veniva chiesto agli ascoltatori di selezionare la domanda che aveva originato le singole frasi.

L'Esperimento 1 è stato presentato a due gruppi di ascoltatori: 22 nativi anglofoni e 22 italiani, parlanti inglese L2. I parlanti italiani hanno ascoltato un ulteriore set di stimoli, composto da frasi in italiano. I risultati dell'esperimento mostrano che gli ascoltatori nativi anglofoni possono distinguere la localizzazione del focus ristretto sulla base della prosodia anche senza la necessità di ulteriori informazioni legate al contesto della comunicazione. Ciò avviene sia quando ascoltano i parlanti NS che quando ascoltano i parlanti NNS1, mentre il riconoscimento delle produzioni dei parlanti NNS2 non supera il livello di casualità. Gli italiani invece sono anch'essi in grado di riconoscere il focus nelle produzioni dei parlanti nativi, ma non ottengono risultati significativi per le produzioni di entrambi i gruppi di parlanti inglese L2. Per quanto riguarda le frasi in italiano, nemmeno

in questo caso gli ascoltatori italiani non sono in grado di distinguere la localizzazione del focus, dimostrando che in italiano a livello percettivo gli indici prosodici in analisi (altezza tonale e durata) non sono abbastanza per riconoscere la posizione del focus.

L'Esperimento 2 è stato ideato per investigare l'effetto della differenza nella modulazione dell'altezza tonale nella corretta distinzione del focus ristretto da parte di ascoltatori nativi anglofoni, mediante la manipolazione del segnale acustico. In generale, i risultati dell'Esperimento 2 confermano che l'altezza tonale gioca un ruolo importante nel riconoscimento del focus ristretto anche dal punto di vista percettivo, almeno per quando riguarda le produzioni dei parlanti nativi anglofoni. Questo non è però generalizzabile per quanto riguarda le produzioni in inglese L2, dove i risultati degli ascoltatori non si allontanano significativamente dalla soglia della casualità, in nessuna delle condizioni sperimentali.

In conclusione, i risultati dello studio di produzione e dello studio di percezione convergono nel mostrare che in inglese l'altezza tonale gioca un ruolo fondamentale nella produzione e nella percezione del focus ristretto di tipo non-contrastivo. Per quanto riguarda le produzioni in inglese L2, i parlanti NNS1 sembrano in grado di avvicinarsi al modello nativo, almeno in una certa misura, con risultati apprezzabili sia dal punto di vista dell'analisi del segnale che della percezione acustica. I parlanti NNS2, invece, sembrano essere incapaci di adottare le strategie proprie dell'inglese, trasferendo in L2 le strategie tipiche dell'italiano, come si evince dal confronto con i risultati ottenuti nella produzione e percezione delle frasi in italiano L1.

I risultati riportati in questa tesi sono interessanti non solo per la ricerca fonetica, ma anche per la loro possibile applicazione nell'insegnamento e apprendimento delle lingue straniere, dove la prosodia sta iniziando a essere studiata e insegnata con rinnovato interesse e vigore come parte integrante dell'acquisizione di una corretta pronuncia in L2 (Busà 2012).

List of Figures

List of Tables

Part I

Background

Chapter 1

Introduction

1.1 The issue

It is well known that the role of prosody is crucial for effective communication. This is particularly true for communication in a second language (L2), where an incorrect use of prosodic features could lead to critical misunderstanding and, eventually, to communication breakdowns. The importance of the acquisition of L2 prosody has been remarked by Mennen, who wrote that "[j]ust as poor [segmental] pronunciation can make a foreign language learner very difficult to understand, poor prosodic and intonational skills can have an equally devastating effect on communication and can make conversation frustrating and unpleasant for both learners and their listeners" (Mennen, 2007: 54).

However, the acquisition of L2 prosody is not an easy task for a non-native speaker, not only with respect to phonetics and phonology, but also for the many levels of meaning that are conveyed through prosody. In this regard, Chun (2002) has grouped the functions of prosody into four different categories: grammatical, discourse, attitudinal and socio-linguistic. Along these categories, corresponding levels of meaning can be conveyed. For example, by uttering a sentence, a speaker can seamlessly convey grammatical

meaning by the use of an appropriate pitch contour (e.g., distinguishing between questions or statements) and highlight the most relevant pieces of information in the context of the on-going discourse (e.g., marking the new and the given information). At the same time, their production will also say something about the speaker's mood, or emotional attitude, and their socio-linguistic origin or status. If one considers this multifaceted nature of prosody, it is not surprising to conclude that "[suprasegmentals] seem to be extremely hard for second language learners to acquire" (Busà, 2007).

Another source of difficulty for non-native speakers of English is the lack of explicit instruction on prosody, as few curricula include explanations and activities specifically aimed to promote the acquisition of prosody (Grice & Baumann, 2007; Busà, 2007). In addition, it has been reported that language teachers often feel that they are inadequately prepared to teach prosody and prefer focusing on more familiar activities based on phonemic acquisition (Busà, 2010; Celce-Murcia et al., 2010). Since the learners normally acquire L1 prosody at vey early stages of their lives and are often not consciously aware of the mechanisms involved (Busà, 2008), the absence of methods that could promote a conscious awareness on prosody can seriously hinder the successful acquisition of L2 prosody. Fortunately, the importance of prosody has been generally acknowledged, and L2 prosody has become a thriving field in academic research. As a consequence, things are starting to change also in language instruction, with a renewed and deeper interest on the prosodic features of L2 (Trouvain & Gut, 2007; Busà, 2012).

This study aims to contribute to the study of L2 prosody. The topic of this dissertation is the phonetic realization of narrow focus by native and non-native speaker of English. Focus marking is what allows speakers to give prominence to words or larger constituents that are new or otherwise relevant in the context of an on-going conversation. The notion of focus is therefore closely connected to the 'discourse function' of prosody, as proposed by Chun (2002), since it involves the relation of the information presented in

a sentence to the whole surrounding discourse.

Although all languages have ways to signal prominence and to signal information structure, different languages have different ways to mark focus (Ladd, 1996). The focus marking strategies of the languages of the world can involve prosody, syntax and morphology. There can also be strategies based on the combinations of all these linguistic systems (Büring, 2009).

The two languages compared in this study, English and Italian, are very different in marking prominent information at sentence level. In English pitch accents (i.e., from the acoustical point of view, local F_0 peaks) play an important role in marking the most relevant information in the larger context of a conversation (Büring, 2007). For example, the appropriate response to the question 'Who ate the pies?' would be 'Paul ate the pies'. In contrast, the appropriate response to 'What did Paul eat?' would be 'Paul ate the pies'. In these sentences focus indicates that *Paul* and *pies* correspond to the most relevant, or new, information in the discourse and answers the preceding question. In Italian, instead, focus is normally marked with word order strategies, for example by moving the highlighted constituent to a fixed position in the right periphery of the sentence with a process of dislocation (Avesani & Vayra, 2000). More information on the differences between focus marking strategies in English and Italian will be provided in Section 2.6.

It is important for non-native speakers of English to learn how to correctly realize focus by the use of prosodic cues. Accenting the wrong word in a sentence can generate confusion in the listeners, since it provides them with distorted information on which constituents are new or old in the conversation or what the actual topic of a discussion is (Baker, 2010). As a result, a difficult identification of the prominent information in non-native speech "often obscures the intended pragmatic meaning and the understanding of the message" (Ramírez Verdugo, 2006: 9). From the perceptual point, of view, the ability to recognize prosodic focus marking in English allows a listener to benefit from a systematic mapping of new and given information

on accented an de-accented constituents respectively (see Section 2.5).

1.2 Research questions

This dissertation is aimed to study the phonetic realization of narrow focus by native and non-native speakers of English. In particular, attention will be directed to the non-native productions and to the possible progressive tuning that can be expected from L2 speakers with a higher competence in L2. The main research questions driving this study regard both sides of the communication process: production and perception. The production study is aimed to answer the following questions:

- Can Italian speakers of English L2 mark narrow focus by using prosodic cues, namely pitch and/or duration?
- Do Italian speakers with a higher competence in English L2 learn to mark narrow focus following L2 patterns?
- Do difficulties in acquiring prosodic focus marking depend on phenomena of prosodic transfer form L1?

As for the perception study, the questions to be answered are the following:

- Do fine-grained differences in prosodic cues have a discriminant effect in the perception of narrow focus?
- Can English native listeners successfully identify narrow focus only by prosody when listening to non-native productions? Does perceptual success depend on non-native speakers' competence in L2? Can Italian listeners recognize focus too in the English productions?

- Can Italian native listeners successfully identify narrow focus when listening to Italian sentences only by prosody, without any extra contextual information?

- Is there a relation between L2 proficiency and the successful perception of narrow focus?

It is expected that the results from production and perception will converge in showing that the acquisition of the prosodic marking of narrow focus is a difficult task for Italian speakers of English. However, it is also expected that the most experienced learners will be able to show a progressive tuning (Ueyama, 2012) to the native models. Their productions will show an active use of prosodic cues, mainly pitch, to mark focus. This progress will be reflected by better results in the listeners' perception.

1.3 Relevance and factors of innovation

Throughout this dissertation, the author will refer to 'narrow' focus intending 'narrow non-contrastive', or 'narrow informative' focus. This distinction is particularly important, not only for the difference between the two types of foci (see Section 2.3.3), but also for the general significance of this research. Much of the cross-linguistic research carried out on the acquisition of prosodic marking of focus has been based on narrow contrastive focus, sometimes abbreviated as NFC (cf., for Italian-accented English, Stella & Busà, in press; Busà & Stella, 2012; Gili Fivela, 2012). In contrast, to the author's knowledge, the realization of narrow informative (non-contrastive) focus by Italian speakers of English L2 has not yet been studied.

However, the acquisition of the prosodic marking of narrow focus seems a crucial point to study, since it represents a real difference between English and Italian. Italian has its own contrastive focus, which is used with the same pragmatic purposes of its English counterpart, while it is not clear whether

Italian can prosodically mark a non-contrastive narrow focus at all. As for English, instead, several experimental studies have shown that narrow (non-contrastive) focus is still acoustically characterized by a pitch accent on the word in focus (see Section 2.4.2).

Another factor of innovation of this study is the decision to work on British English, in particular on the so-called Standard Southern British English (SSBE), which is considered the standard variety for English spoken in the United Kingdom (Grabe et al., 2008). The experimental works based on this variety are few (e.g., Eady et al., 1985; Cooper et al., 1986), as most studies on prosodic focus marking in English are based on American varieties of English (e.g., Xu & Xu, 2005; Breen et al., 2010; Baker, 2010). The choice to work on British English was also motivated by the fact that the instruction of the Italian participants in this study is largely based on the British model and conducted by language instructors that are native from Britain. As for the variety of Italian, this dissertation is based on the Italian spoken in the Veneto region, in the North-East of Italy. This variety of spoken Italian was first studied in relation to English L2 pronunciation (Busà, 1995). Since then, Busà and colleagues have kept working on this variety, with a special interest on the acquisition of L2 (e.g., Busà, 2007; 2008; 2010; 2012; Busà & Urbani, 2011; Busà & Rognoni, 2012; Busà & Stella, 2012; Stella & Busà, in press).

Finally, the relevance of this dissertation can be seen also from the point of view of its implications for language instruction. It has been mentioned how effective teaching practice and materials can be inspired by the academic research in L2 prosody (Gut et al., 2007). The experimental nature of the research presented in this dissertation is meant to provide a good amount of empirical data that could also be used to make predictions on L2 learning.

1.4 Structure of the dissertation

This dissertation is structured in ten chapters, distributed in four parts. Part I includes the first four chapters of the dissertation, which present all the background information that led to the experimental research presented in this dissertation. In particular, the present chapter (Chapter 1) is dedicated to introduce the topic of this dissertation, presenting its relevance and outlining the main research questions driving the study. Chapter 2 will set the foundations of this study, starting from the definition of prominence and of concepts specifically dealing with focus marking, such as focus breath, focus location and focus type. The remainder of the chapter will present a review of the relevant literature on the phonetic realization of narrow focus in English and in Italian, with a discussion of the main theoretical frameworks that have been used in experimental studies of prosody, and, in particular, focus marking.

Chapter 3 will present the main features of the most influential L2 speech acquisition models, with a special attention on the compatibility of the acquisition of prosody within these theoretical frameworks. The chapter will also deal with the methodological issues in the study of foreign accent, reviewing relevant bibliography in the perception of non-native prosody. To conclude, the chapter will include a commented overview of the main methods used to manipulate the acoustic signal in order to study the relative importance of the single prosodic cues while limiting the influence of segmental information.

Chapter 4 will be aimed to bridge the gap between theory and practice in the structure of the dissertation. The chapter will discuss the methodology and the results of four pilot studies that were designed by the author in order to collect first-hand empirical data on the perception of prosody in Italian-accented English productions. These four experiments are mainly aimed to determine the relative importance of duration and pitch in the perception of Italian accent in English. At the same time, the four pilot studies are also used as a benchmark to test the viability of several manipulation methods

discussed in Chapter 3.

Part II corresponds to the production study. In particular, Chapter 5 will lie out the hypothesis driving the production study and the methodology adopted in selecting consistent groups of speakers of English L2 and in collecting the speech data. The chapter will also present the acoustic measurements that are used to analyze the phonetic realization of narrow focus at sentence and word level. Chapter 6 presents the results of the acoustic and statistical analysis for each of the mentioned three levels, with brief discussions that will anticipate the General Discussion (Chapter 9).

The perception study is presented in Part III, where Chapter 7 and 8 will be dedicated to the presentation of the first and second perception experiment, respectively. The two chapters will be organized with the same structure, presenting rationale and hypotheses, methodology and results of each experiment, followed by a brief discussion of the results. A full-scale discussion of the results will be found in Chapter 9.

Part IV is composed by the General Discussion (Chapter 9) and by the Conclusion (Chapter 10) of this dissertation. Chapter 9 will extensively discuss the experimental data, from both the production and the perception studies. The relation between the results from production and perception will also be discussed. Chapter 10 will close the dissertation by presenting the conclusions that can be drawn from the data. The implications of the results within the framework of the current L2 speech acquisition models and for language instruction will also be considered. The work will close with some reflections on the possible limitations of this study and with an outline of the future directions of research that could be started and expanded from the work presented here.

Chapter 2

Prominence and focus marking

2.1 Introduction

This chapter will begin by presenting the concepts of prominence (Section 2.2) and by proposing a three-level model of focus (Section 2.3), composed by location, breadth and type, with a mention to the connected phenomenon of deaccenting (Section 2.4).

Section 2.5 will discuss the two main approaches to the study of prominence, namely the Autometrical-segmental theory of intonational phonology and one based on the assumption of a direct relationship between the acoustic characteristics of the speech signal and prominence. When reviewing both approaches, particular attention will be paid to the relevant literature regarding the prosodic marking of sentence prominence in English and in Italian.

Section 2.6 will discuss focus marking from a cross-linguistic perspective, reviewing the most recent literature regarding the strategies adopted in English and Italian, while Section 2.7 will be focused on the review of production and perception studies on the acquisition of L2 prominence marking strategies.

Section 2.8 will conclude the chapter by presenting the reasons why the

direct-relationship approach was adopted to study the phenomenon presented in this dissertation.

2.2 Prominence

A widely quoted definition of prominence is the one given by Terken, who explains it as "the property by which linguistic units are perceived as standing out from their environment" (Terken, 1991: 1768). Similarly, Mertens states that "a syllable is prominent when it stands out from its context due to a local difference for some prosodic parameter"; the same author also argues that "[p]rominence is continuous (not categorical) and contributions of multiple parameters can interact" (Mertens, 1991: 218). Rump defines the prominence of a syllable as "its perceptual conspicuousness or salience relative to the neighbouring syllables" (Rump, 1996: 2), and in a recent study by Marotta and colleagues, prominence is similarly defined as "degree of perceived saliency assigned to some words or syllables within an utterance" due to a significant modification of the three main acoustic parameters, i.e., duration, intensity and frequency" (Marotta et al., 2012: 67, translation by the author).

These are only a few of the many definitions of prominence given in the literature, but they are all representative of three main characteristics of prominence: its relativity to the surrounding context; the fact that it is conveyed by an interaction of several acoustic cues; its perceptual nature. These main characteristics have motivated the majority of research on prominence, both within and across languages.

It is worthwhile to point out that, despite being a function of intonation, prominence needs to be clearly separated from the dimension of pitch. In this regard, Ladd distinguishes pitch and relative prominence as "two orthogonal and independently variable aspects" (Ladd, 2008: 6). Kohler also marks the separation of the two functions, writing that, although prominence shares

F_0 as a physical property with pitch, "it is not entirely determined by it, but also depends on syllable and segment duration, intensity, and possibly other features" (Kohler, 2003: 2930). In another work, the same author adds that "beside the accent category that is principally signaled by F_0 excursion and may be called pitch accent, another type of accent has to be recognized that is primarily related to non-pitch features, viz. acoustic energy, based on phonatory and articulatory force, and may therefore be called force accent" (Kohler, 2005: 99). The idea is that prominence is achieved through the interaction of pitch accents and force accents, in a dynamics of mutual interaction and reinforcement (Tamburini, 2009).

As will be shown in Section 2.5.2, many researchers have tried to find a direct connection between the realization of prominence and certain acoustic parameters, although the results of the studies are often conflicting. The contradictions in the results are motivated, on the one hand, by the intrinsic variability in the productions, even across speakers of the same language (see Vaissière, 2005); on the other hand, by the wide range of methodologies in data collection, which makes it difficult to compare results and to generalize them even within a single language (Breen et al., 2010). Many acoustic parameters have been proposed to account for prominence, from the direct observation of the acoustic cues traditionally associated to prosody (F_0, duration and intensity), to more complex parameters based on the distribution of energy across the acoustic spectrum, such as spectral tilt or spectral balance (Sluijter & Van Heuven, 1996; Heldner, 2003).

Another reason why the study of prominence is particularly complex resides in the fact that prosody is not the only way to mark prominent information. The languages of the world can recur to a variety of resources to mark prominence, such as word order movements, described by syntax and morphology (Ladd, 1996), or other pragmatic strategies (Büring, 2009). In this work we will consider the concept of sentence-level prominence and focus from the point of view of their realization through prosody. Pointers to wider

discussions in the literature about the concept of focus and its ramifications in syntax and pragmatics can be found in Ladd (1996) and Büring (2007).

2.3 Focus

The main function of prominence is to mark information structure, which can be defined as "the differential contributions of different sentence elements to the overall sentence meaning in relation to the preceding discourse" (Breen et al., 2010: 1044). The information status of the elements in an utterance is articulated in two levels: focus and givenness. From a functional perspective, focus has been defined as "an emphasis on some part of a sentence as motivated by a particular discourse situation" (Xu & Xu, 2005: 161), and it normally corresponds to the information that is introduced as new and/or is put on the foreground of the discourse. In contrast, given information is material that has already been made salient explicitly, that is, in the previous discourse, or implicitly, based on inferences drawn from world knowledge (Schwarzschild, 1999).

The present work will adopt a three-level model of focus marking, which is summarized in Tab. 2.1.

2.3.1 Focus location

The first level of focus is focus location. Location refers to where focus is placed, in particular, on which unit of a given utterance (Breen et al., 2010). As will be shown in detail in the next two subsections, focus can be located on virtually any element (subject, verb, object...) or constituent of a sentence, depending on the needs of the ongoing communication exchange.

Table 2.1: The three levels of focus marking

Focus Location	Where is the focus?	subject verb object ...
Focus Breadth	How wide is the focus?	*Narrow*: on a single constituent *Broad*: on a whole phrase
Focus Type	What kind of focus is it?	*Contrastive*: emphasis on a single constituent *Non-contrastive (informative)*: see *narrow* focus

2.3.2 Focus breadth

Focus can be marked with two different scopes, *broad* or *narrow*: this distinction is what has been called *focus breadth* (Selkirk, 1984, Gussenhoven, 1983) and it refers to the size of the set of the focused elements (Breen et al., 2010). Narrow focus applies to the cases where only a single constituent of a sentence is marked as prominent, while broad focus refers to wider strings of information, such as the entire event described in an utterance.

As an example, if the context preceding a sentence is a general question, the realization of the sentence will follow a neutral, or default, pattern. This neutral pattern represents broad focus. In English and in Italian broad focus is signaled by placing a pitch accent on the rightmost stressed element of the sentence. This is shown in the examples in (1), which show two pairs of questions and answers with the same meaning, the first in English and the second in Italian.

(1) (What's going on?) Bruno is eating the pear.
(Che cosa sta succedendo?) Bruno sta mangiando la pera.

However, communicative needs may also require a particular emphasis on

a single element. In this case, English speakers can highlight a constituent by moving the pitch accent on that particular element. Acoustically speaking, the emphasis can be conveyed with a peak in F_0, longer duration and higher intensity (cf. Eady et al., 1985; Xu & Xu, 2005; Breen et al., 2010, see Section 2.5.2). When a single constituent is highlighted, the utterance is said to present a narrow focus on that constituent. A typical example of narrow focus in English is what Büring (2007) calls *Question-Answer Congruence*: in replies to wh-questions, narrow "foci correspond to the wh-expression in a preceding constituent question" (Büring, 2007: 447). The example reported in (2) shows that in an answer to a wh-question, the prominence will be placed on the element of the utterance corresponding to the wh-element in the question, which will result narrowly focused.

(2) (Who's eating the pear?) Bruno is eating the pear.

Similarly, virtually any word of a sentence can be narrowly focused, depending on the preceding context. Further examples are provided in (3) and (4).

(3) (What's Bruno doing with the pear?) Bruno is eating the pear.
(4) (What's Bruno eating?) Bruno is eating the pear.

As for Italian, it is not clear whether the Question-Answer Congruence proposed by Büring (2007) can apply. As will be explained in section 2.6, in Italian focus is more often marked with word order strategies rather than with prosody (Ladd, 1996). In fact, it seems that in Italian focus is prosodically marked only when extra emphasis is needed, so it is possible that in Italian the prosodic marking is limited to the contrastive type of narrow focus (see Section 2.3.3). The results from production and perception presented in this dissertation (see Chapters 6-8) seem to confirm that in Italian the phonetic realization of narrow non-contrastive focus is non-prosodically marked (cf. Section 9.2.2 and 9.3.1 for the discussion of the relevant results).

Both in English and in Italian, there are cases where the opposition between broad and narrow focus is not so clearly defined, as can be seen by comparing the examples (1) and (4). When narrow focus is placed on the rightmost word in the sentence, which is the default location of broad focus in both languages, the resulting utterance becomes perceptually ambiguous (Ladd, 1996).

The difference between the realization of narrow (contrastive) focus and broad focus on the rightmost constituent of an utterance has been studied for regional varieties of Italian spoken in the central area of the country (e.g., Firenze: Avesani & Vayra, 2003; Pisa: Gili Fivela, 2002) and in the South (e.g., Naples: D'Imperio, 2002; Bari, Naples and Palermo: Grice et al, 2005; Lecce: Stella & Gili Fivela, 2009). Depending on the regional variety, the ambiguity between broad focus and narrow focus located in final position may or may not be solved by prosody alone. As for English, although there are studies aimed to find distinctions in the two realizations on the basis of the acoustic cues in the signal (e.g., Eady & Cooper, 1985; Xu & Xu, 2005), it seems that the realization of narrow focus and broad focus on the rightmost constituent of a sentence presents "an ambiguity that can only be resolved through contextual information" (Van Heuven, 1994: 17).

2.3.3 Focus type

Type represents the third level of focus. Within narrow focus, there can be two types: informative and contrastive. While the former type corresponds to what has already been said for narrow focus (for example, the marking of some new information in reply to a preceding wh-question), contrastive focus is typically used to highlight a concept or to correct a specific item that has already been mentioned in the preceding discourse (Ladd, 1996). Consider the examples in (5) and (6), where contrastive focus is used to correct a piece of information. Both examples show that even function words can be realized in contrastive focus, if this is required by the context (Wells, 2006).

(5) (Did Joe make a pizza with Meg?) No, he made a pizza <u>for</u> Meg.

(6) (Did you drink two beers?) No, I drank <u>one</u> beer!

In contrast, an Italian speaker would be likely to mark focus by moving the word to be highlighted to the right periphery of the sentence, which is the default position for focus, in a process known in syntax as dislocation. The resulting sentence would sound like the example reported in (7).

(7) (L'ha disegnato Mario?) No, l'ha disegnato <u>Gino</u>.
tr. (Did Mario draw it?) No, <u>Gino</u> drew it.

Theoretically, in Italian it might also be possible to mark narrow contrastive focus without recurring to dislocation, as it is shown in the example in (8).

(8) (L'ha disegnato Mario?) No, <u>Gino</u> l'ha disegnato.
tr. (Did Mario draw it?) No, <u>Gino</u> drew it.

However, an Italian listener would find the realization in (7) much more natural than the one in (8), as the latter results a marked case as compared to the more likely realization in (7). For both (7) and (8), it is interesting to point out that the translation in English would be the same.

In the literature there is no consensus on the relationship between informative and contrastive focus. While the two types of focus have been treated as different categories of information structure by some researchers (e.g., Chafe, 1976; Molnar, 2002), others have proposed that there is no systematic difference between the two (e.g., Bolinger, 1961; Rooth, 1992), being just instances of narrow focus. The researchers defending the latter position argue that every expression evokes an implicit set of alternatives even when they are not explicitly present in the discourse, considering therefore any narrow focus as contrastive. This is modeled in (9), where the constituent marked with a contrastive focus is seen as one of a set of virtual alternatives which may or may not be explicitly present in the previous discourse.

(9) Johhny | *plays* / *walks* / *jumps* / *runs* / ... | with the green frog

The existence of a contrastive focus has also been debated in more strictly phonetic terms as contrastive (pitch) accent. The different positions are well presented in Krahmer & Swerts (2001), where the authors review the main contributions in the discussion on the titular "alleged existence of contrastive accent". Among the works cited, the positions of Couper-Kuhlen (1984) and Chafe (1976) are worth noting, who found that contrastive accents are followed by a sudden drop in pitch, while pitch tends to descend more gradually after their non-contrastive counterparts. The idea that contrastive accents are more emphatic than the informative ones (Ladd, 1996) was experimentally confirmed in Bartels & Kingston (1994), where it was shown that contrastive accents are characterized by higher F_0 peaks.

Within the theoretical framework of intonational phonology (see section 2.5.1), Pierrehumbert and Hirschberg (1990) suggested that contrastive accents follow an L+H* pattern (a steep rising movement in pitch from a low to a high tonal target), whereas informative accents have an H* configuration (a gradual rising movement towards a high target). Although this difference was demonstrated by Ito et al. (2004) and found in other languages analyzed within the same framework (e.g., Grice et al., 2005, and Avesani & Vayra, 2003 for regional varieties of Italian), researchers following a more direct approach to the analysis of the speech signal have pointed out the difficulty of reliably distinguishing H* and L+H*, suggesting a more quantitative approach for the analysis of focus type (see Xu, 2011a; Breen et al., 2012).

In absence of clear evidence that might conclusively exclude the existence of a difference between the contrastive and non-contrastive (or informative)

types of narrow focus, this study will maintain the traditional distinction between the two types of foci and pitch accents.

2.4 Deaccenting

An inevitable by-product of the prosodic marking of narrow focus on specific words is a phenomenon known as deaccenting (Ladd, 1980). Deaccenting has been defined as 'the absence of an accent on a word that might otherwise be expected to be accented" (Swerts et al., 2002: 630) or as 'the removal of phonological accent on a constituent" (Tancredi, 1992: 2). While accenting is normally used as a pointer to new or contrastive information, deaccenting is used to counterbalance this by signalling that a word or a constituent is to be considered as given information (Avesani & Vayra, 2005). English and Italian adopt different focus marking strategies (see Section 2.6); as a consequence, the two languages also differ in the way they deaccent information. While English insists on deaccenting given material, Italian "quite strongly" resists it (Ladd, 2008: 232). For example, in English it is possible to deaccent single words, while in Italian only longer constituents can be deaccented (Swerts et al., 2002). This difference can be seen in the examples (10) and (11), adapted from Ladd (1996). The example in (10) shows what normally happens in a production by a native English speaker.

(10) Running is like walking in haste, only you have to go much more in haste.

The example reported in (11) represents a hypothetical version of (10) in Italian, maintaining the same balance between accenting and deaccenting found in English.

(11) *Correre è come camminare in fretta, soltanto che si deve andare più in fretta.

An Italian listener would be very likely to reject this realization, because the adverbial phrase is only partially deaccented. A more realistic realization would be the one reported in (12).

(12) Correre è come camminare in fretta, soltanto che si deve andare più in fretta.

These examples are consistent with recent works published by Bocci & Avesani (2008; 2010), where it is argued that deaccenting in Italian works as a placeholder for post-focal information in the rightmost position and not as a specific marker of given information as in English. The systematic differences in accenting and deaccenting the elements that are relevant or irrelevant, respectively, facilitate English speakers and listeners in consistently mapping new and given material, while in Italian the link between givenness and deaccenting is only partial or occasional (Avesani & Vayra, 2005). It is very likely that this difference in mapping the information status in the two languages can cause serious problems to Italian learners of English L2. As mentioned in Section 1.1, incorrectly marked prominence can generate confusion in the listeners in determining the actual topic of a discussion or the information structure of a sentence intended by the non-native speaker.

2.5 Approaches to the study of L2 prosody

Empirical research on the prosodic realization of prominence has mainly followed two different theoretical frameworks. The first is represented by the autosegmental-metrical (AM) theory of intonational phonology (Ladd, 1996), an approach that is based on the assumption that the relationship between signal and meaning is mediated by phonological categories. The second framework has been called *direct-relationship approach*(Breen et al., 2010), and it is based on the acoustic analysis of the signal, with the aim of finding the possible direct correlates of the functions played by prosody.

This section will present the two perspectives, exploring advantages and disadvantages of both approaches in relation to prominence and focus marking. Particular attention will be paid to studies describing English and Italian.

2.5.1 The AM theory of intonational phonology

The auto-segmental metrical (AM) theory of intonational phonology is one of the leading theoretical frameworks in the study of intonation. Inspired by the American autosegmental and metrical phonology of the, 1970s, the theory of intonational phonology has its foundation stone in Pierrehumbert (1980). The approach, initially based on the description of American English, was then adopted and applied to the study of a great number of languages, soon becoming one of the main research paradigms in the study of intonation.

In his book *Intonational Phonology*, Ladd (1996) states the *four tenets* of the approach, which will be summarized here. The first is the sequential tonal structure: the intonation structure consists of a sequential series of local events that are associated with specific points in the segmental string. The second is the distinction between pitch accent and stress: while pitch accents are considered the building blocks of intonation in the AM framework, (word) stress is considered a specifically phonetic phenomenon, the study of which belongs to the field of acoustic phonetics. The third principle of intonation phonology is the analysis of pitch accents in terms of level tones, in contrast with models based on continuous pitch movements. The last of the *four tenets* is the local sources of global trends: global pitch movements are generated by the sum and combination of a series of locally implemented events. These four concepts are the theoretical bases for the elaboration of one of the most notable contributions of intonational phonology to the study of intonation: the Tone and Break Index (ToBI) transcription system for intonation.

One of the early purposes of ToBI (Silverman et al., 1992) was to offer a basis to synthesize intonation by rule, and this practical orientation was reflected by the structure of the transcription system. In contrast with the

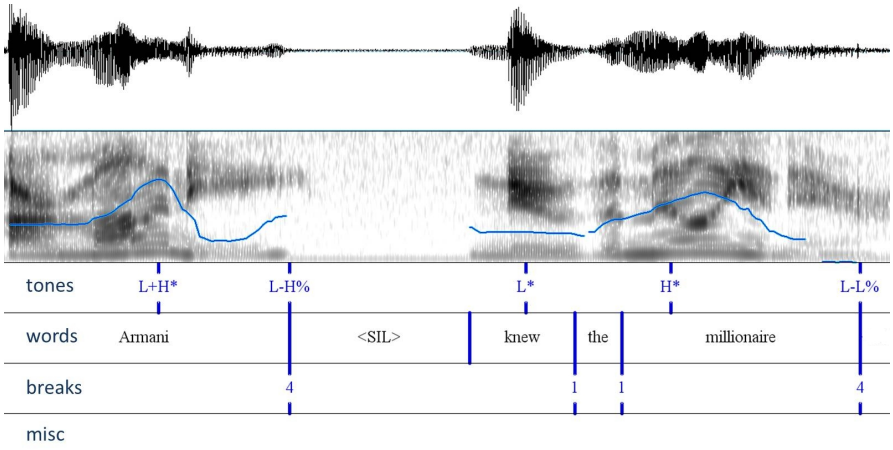

Figure 2.1: A sample transcription with ToBI (from http://anita.simmons.edu/ tobi/tutorial.html).

previous notation systems, based on the visual reproduction of pitch movements (see the *British* school, e.g., Cruttenden, 1997; Wells, 2006), the assumption behind ToBI is that the continuous realization of pitch movements can be described as a succession of discrete, categorical, tone levels. Therefore, ToBI presents a limited inventory based on a binary scheme consisting of two tone levels, low (L) and high (H). These may correspond to pitch accents (marked with a star, e.g., L* and H*) or boundary tones (marked with a - or %, e.g. L% or H%). The two tone levels can also be combined together in bitonal accents (e.g., L+H*). ToBI is also used to describe the hierarchical organization of intonation, or phrasing, marking the strength of prosodic boundaries with a series of break indexes. A complete ToBI transcription includes a series of tiers accompanying the visual representation of the F_0 contour: one for the orthographic or phonetic transcription, a second for the tone levels, a third for break indexes, and an optional fourth one for miscellaneous annotations and comments (see Fig. 2.1).

ToBI-based annotations have been widely used to describe pitch contours and the associated syntactic functions (e.g., declarative vs. interrogative intonation), or the relationship between pitch contours and phrasing. The

annotations are normally assigned by hand by expert researchers, who base their judgments on the visual and auditory analysis of the signal. However, a few automatic methods have been recently proposed (e.g., Rosenberg, 2010, Mertens, 2013).

With its elegance and richness in information, the ToBI-based annotation has soon become a widely accepted standard, not only for the study of the varieties of English, but also for many other languages (cf. Jun, 2005). However, not all phoneticians are satisfied with this annotation system, and have criticized it on several grounds. From the point of view of the theoretical assumptions behind ToBI, there have been criticisms against its sequential and categorical nature: decomposing the continuity of pitch contours in smaller sequential events leads to treat intonation more as a segmental rather than as a suprasegmental phenomenon (Albano Leoni, 2009). There have also been criticisms on the alleged poverty of the system for accounting for the great variety of intonation patterns and for capturing the sizable differences among regional varieties within the same language (Marotta, 2008). A solution to this problem could be adopting expanded versions of ToBI, with the risk of drifting away from the elegance and from the shared conventions that were considered the foundations of the original model.

Wightman (2002), who was one of the creators of the original ToBI system (cf. Silverman et al., 1992), presents a series of more practical issues. A first practical problem is the inter-transcriber agreement: while the agreement is normally very high when labeling boundaries, it is much lower when it comes to assigning intonational labels, even among highly and uniformly trained labelers working in ideal laboratory conditions. This issue has also been recently studied by Breen et al. (2012), who found confusion in labeling contrastive focus as H+L vs. H. Another practical issue reported by Wightman is the slowness of the labeling procedure, taking "typically [...] 100 to, 200 times real time" (Wightman, 2002: 27). Wightman concludes that the recent reductions in costs and time for hardware and software tools needed

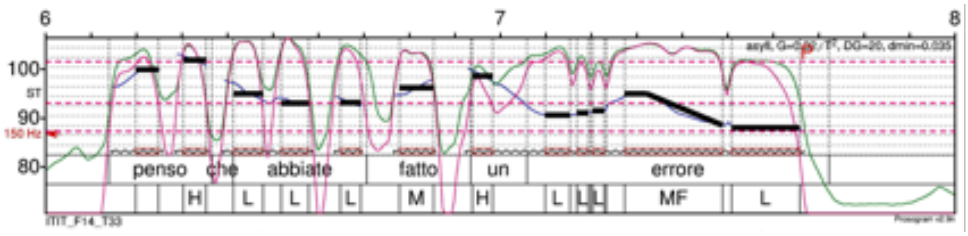

Figure 2.2: An example of annotation output using *Prosogram* (from Mertens, 2013).

to annotate prosody have obviated the need for the descriptive labeling offered by ToBI, since "virtually anybody can now get time-aligned waveform, pitch track and spectrogram displays" (Wightman, 2002: 28). This is what motivated the development of new software meant to create multi-layered transcriptions of intonation, based on the holistic visual inspection rather than recurring to a fixed system of labels. Among these alternative solutions one can quote *WinPitch* (Martin, 2004), *Prosogram* (Mertens, 2013, see Fig. 2.2) and *Prosomarker* (Origlia & Alfano, 2012).

Prominence and focus marking have been studied extensively within the intonational phonology framework, mainly in terms of their manifestation as pitch accents. Büring (2007) states that "[t]he main correlate of perceived prominence in English is a pitch accent, acoustically a local maximum or minimum of the fundamental frequency" (Büring, 2007: 445). Moreover, the author points out that within an utterance the "final pitch accent is invariably perceived as the most prominent one" and is referred to as the nuclear pitch accent" (Büring, 2007: 446).

The studies on focus within the intonational phonology framework are mainly centered on the categorical distinction between narrow and broad focus. The view expressed by Pierrehumbert & Hirschberg (1990) that contrastive accents have a peculiar manifestation as L+H* patterns, mentioned in Section 2.3.3, has been maintained by many followers of the AM phonological theory and tested on other languages. In particular, narrow contrastive

focus has been often used in studies comparing the production and perception of narrowly vs. broadly focused constituents (e.g., Avesani & Vayra, 2003, Busà & Stella, 2012). This choice is particularly motivated when researchers deal with languages that normally recur to strategies other than prosody alone (e.g., the Romance languages). Since narrow contrastive focus is supposed to be realized with particular emphasis (Ladd, 1996), it is normally preferred to its less prosodically characterized informative counterpart.

In the case of English, the main contributions to the study of prominence and focus within the intonational phonology framework have been reviewed in Ladd (2008). Within the intonational phonology framework, focus has been typically studied in terms of its relationship with pitch accents, as reported in the already mentioned passage by Büring (2007). A fair amount of work within this framework has been more oriented towards the study of the relationship between syntax, phonology and semantics rather than towards the phonetic realization of focus. This is the view expressed by the *Focus-to-Accent* (FTA) approach (see Ladd, 1980; Gussenhoven, 1983; Ladd, 1996).

In recent years Ladd and Mennen have also promoted a more empirical approach to the study of intonational phonology, in order to explain how tones are implemented phonetically. In particular, two phonetic measurements have been proposed: tonal alignment and scaling.

Tonal alignment can be defined as the temporal relation of pitch accents with the segmental string, and it has been shown to present language- and dialect-specific characteristic. These differences have been related to the differences in voice onset time (VOT) found in cross-linguistic studies on L2 phoneme acquisition (Mennen, 2007). An example of how alignment is used cross-linguistically is shown in Fig. 2.3, which compares the realization of the Italian proparoxytonic word *Mantova* (the name of an Italian city) by a non-native and a native speaker of Italian.

Fig. 2.3 shows that the L2 speaker correctly places prominence on the

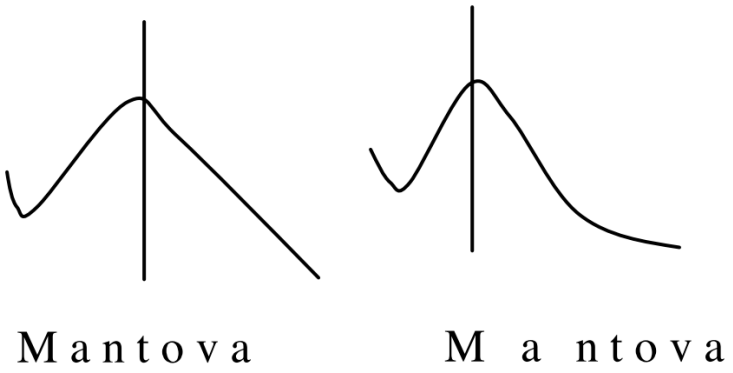

Figure 2.3: A schematic representation of the difference in alignment between a native (left) and a non-native (right) realization of the Italian word *Mantova*. The non-native production presents a delayed peak as compared to the native one (from Mennen, 2007: 59, based on an example provided in Ladd, 1996: 128).

first syllable as done by the L1 speakers, but s/he delays the moment when pitch and segments are aligned. As a result, L1 listeners may interpret this delay in alignment as a mistake in the placement of word stress, when in fact it is only a mistake in the phonetic implementation of tonal alignment (Ladd, 1996; Mennen, 2007).

The second phonetic measure is scaling. Scaling refers to the analysis of pitch range, which for Ladd and Mennen must be seen in terms of two different measures: level and span. Pitch level has been defined as "a reference line calculated over the rises and falls within each contour" (Urbani, 2013: 52), and can be equated to the average F_0 value in a pitch contour. In contrast, pitch span is a measure of the distance between the maximum and minimum values of F_0 in a pitch contour. The two dimensions of pitch range are visualized in Fig. 2.4.

Mennen et al. (2012) and Urbani (2013) have recently shown that in cross-linguistic studies pitch span seems to be more informative then pitch level. For this reason, pitch span will be one of the acoustic measures calculated in the production study presented in this dissertation (see Chapters

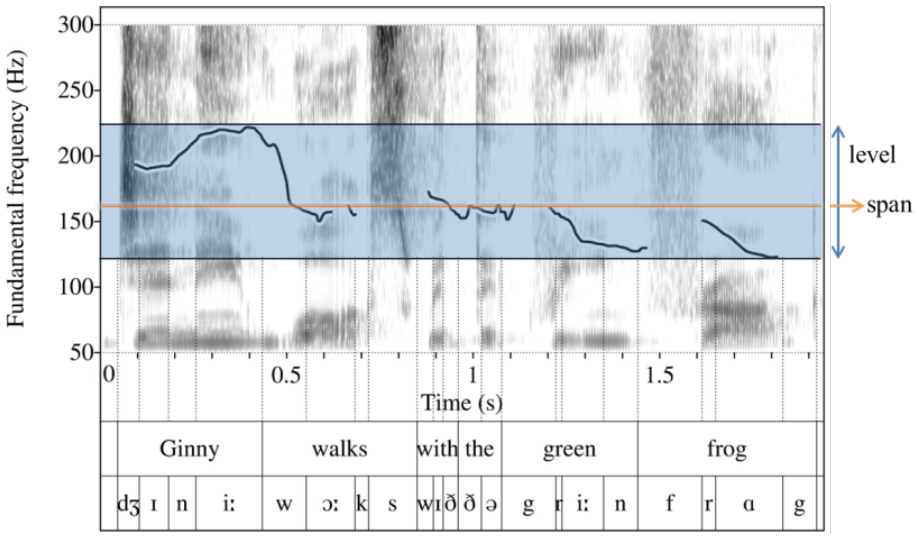

Figure 2.4: Pitch range measurements: pitch span (light blue area) and pitch level (orange line).

5 and 6). As for Italian, with no agreement about the concept of Standard Italian accent (Lepschy & Lepschy, 1977), the study of intonation is a particularly complex issue because of the great socio-linguistic differences between regional varieties. The creation of a unified model to describe Italian intonation is the purpose of the *Atlas of the Italian Intonation* (AItI) project (Gili Fivela et al., under revision), which is comparable to the IvIE project for British English (Grabe, 2004). The AItI project is based on empirical data and aims to apply a shared methodological approach to describe the many intonational varieties of Italian. However, the project is currently being developed and it will take time to see its completion.

In these days, most of the research on Italian intonation, and prominence in particular, is performed within the intonational phonology framework. As mentioned before, studies on Italian varieties have often been based on the opposition between narrow contrastive focus and broad focus, especially in production studies. The results of these studies show differences from a regional variety to the other, although common patterns can be found. As

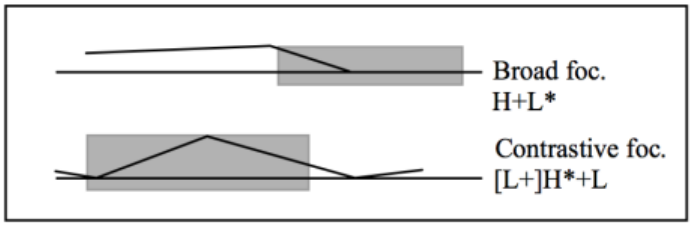

Figure 2.5: Schematic representation of the pitch accent corresponding to broad and contrastive focus in Pisa Italian (from Gili Fivela, 2002).

in English, most studies aim to describe the different realizations of focus in terms of pitch accents and to find the most suitable tone labels to account for them. A few studies have followed the example proposed by Ladd and Mennen, moving from a perspective mainly based only on tone annotation and phonological distinctions to an approach encompassing the analysis of the phonetic detail. This approach has been useful to find differences in the realization of focus: Gili Fivela (2002) and Frascarelli (2004), for example, have shown that broadly focused information is characterized by a more compressed pitch span as compared to narrowly focused information, in Pisa and Roma Italian, respectively, as shown in Fig. 2.5.

In sum, the present section has discussed the theoretical framework known as the AM theory of intonational phonology. This is the main theoretical framework followed in the study of Italian varieties, and one of the most widely adopted to describe the intonation of any language. The next section will present a different, and to a certain extent, complementary approach to the study of prosody, based on the direct analysis of the acoustic signal.

2.5.2 The *direct-relationship* approach

The so-called *direct-relationship* approach (Breen et al., 2010) studies prominence by adopting the research paradigms and methodologies of acoustic phonetics. In this theoretical framework, the study of prominence is based

on the assumption that the functions of speech, and, to a certain extent, meaning, can be directly mapped on acoustic parameters, without the need for the mediation of phonological categories. When studying prosody, acoustic parameters (generally F_0, duration and intensity), are extracted from the signal and analyzed with quantitative statistical methods to describe the speaker's productions and to generate predictions to be tested in perception tests on human listeners.

Many followers of the AM theory of intonational phonology (Ladd, 1996) criticize the *direct-relationship* approach for lacking consideration of the phonological level of intonational meaning. The wide adoption of the intonational phonology framework marked a paradigm shift in the research on prominence and focus marking in favor of studies based on annotation and introspection. However, recent years have witnessed a return to instrumental acoustical studies based on the *direct-relationship* approach. Dissatisfaction with the ToBI-based descriptions and with the confidence on the impressionistic definition of pitch levels rather than on the instrumental clarity of numbers (Breen et al., 2012) was one of the causes behind this revival, together with an easier availability of computation tools that could simplify complex mathematical analyses (e.g., *Praat*).

Early studies on the phonetic realization of prominence in English started to appear in the literature since the, 1950s, with research on the acoustic correlates of word stress in British English (Fry, 1955) and American English (Lieberman, 1960). The results of these studies, based on production and perception, show that the intensity and the duration of the vowel in the stressed syllable have the strongest contribution in the perception of prominence. Conversely, stress perception did not require big F_0 differences (Fry, 1955).

As for Italian, the *direct-relationship* approach was followed in several acoustic studies on word-level prominence carried out in the 1970s and in the 1980s. Magno Caldognetto et al. (1983), Bertinetto (1981) and Marotta

(1985) carried out acoustic studies aimed to the investigation of the realization of prominence in word stress. These studies, based on the measurements of the fundamental acoustic cues of F_0, duration and intensity, agree on the fact that the main acoustic correlate of word stress is duration for all the regional varieties of Italian that were examined. As for experimental research on prominence at sentence level and on the phonetic realization of focus, the two main studies were Magno Caldognetto & Fava (1972) and Kori & Farnetani (1983). These two pioneering studies are both based on the North-East variety of Italian studied in this dissertation and they agree in reporting that narrow contrastive focus is expressed by an F_0 peak.

For English, the first notable contributions in the research on prominence at sentence level are the articles published by Cooper and associates in the 1980s (Cooper et al., 1985; Eady et al., 1985; Cooper et al., 1986). These works are aimed to find the acoustic correlates of different breadths and types of focus in the speakers' productions. From the methodological point of view, these studies were particularly important because they set an example of data elicitation protocol that would be used and adapted in many following studies on the phonetic realization of focus. The speakers were asked to answer wh-questions that could recreate a context in order to trigger a controlled realization of focus on particular keywords corresponding to the wh-elements in the questions. The results of these studies offer empirical evidence to the impressionistic intuition that the element in focus is characterized by a concentration of acoustic cues, all contributing to focus marking. In particular, it is shown that focused words presented peaks in F_0, that they are longer than their unfocused counterparts and that they are realized with higher intensity. Rump & Collier (1996) integrates the results of these production studies with perceptual evidence. The authors demonstrate the relative nature of prominence, showing how the perception of focus is not to be sought in the acoustical analysis of the focused units, but by looking at the big picture, represented by the whole sentence. The main finding is that

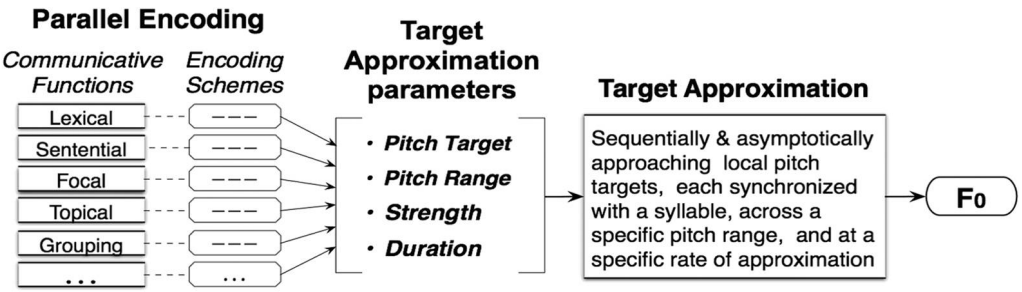

Figure 2.6: Scheme of the PEnTA model (from Xu, 2005).

post-focus pitch range suppression is crucial for focus perception in Dutch: focus can be perceived only if it is final and not followed by any other focused information. Considering the structural similarities between Dutch and English in prosodic focus marking (Büring, 2009), similar results are very likely to be replicated for English.

Applying a methodology that had already been adopted in the study of prominence in Mandarin Chinese (Xu, 1999), Xu proposed a functional approach to the study of English intonation, in contrast with the formal approaches adopted in the studies based on the models of intonational phonology. Xu's contributions can still be considered representative of the direct-relationship approach, although the same author claimed that his model was meant to go beyond a plain direct relationship between acoustics and functional meaning (Xu, 2004). Xu's *Parallel Encoding and Target Approximation* (PEnTA) model offers a multi-faceted analysis of intonation, which accounts for many contemporary functions and events at play (Xu, 2004, 2005). As suggested by its name, the model is based on the two tenets of parallel encoding and target approximation, and is summarized in the scheme reproduced in Fig. 2.6.

In this model, a variety of information streams are encoded in parallel and conveyed through intonation. Pitch is calculated and visualized as a complex set of functions, and its movements are described in terms of dynamic approximation to specific targets, rather than being decomposed in tone levels corresponding to pitch accents (as in the ToBI-based annotation systems). In the PEnTA model, the wide set of annotations include pitch range and a division of focus in pre-focus, focused and post-focus material. All the annotations correspond to a series of complex computations based on the acoustic values extracted from the signal. A detailed explanation of the model can be found in Xu (2004) and Xu (2005).

As for the analysis of prominence and focus, these are specifically addressed in Xu & Xu (2005). In this study, the authors find that the post-focus pitch range suppression mentioned in Rump & Collier (1996) is confirmed for American English, and it is renamed post-focus compression (PFC). In further studies, Xu reports that PFC is a key feature in conveying prominence, being consistently present as marker of focus in many languages of the world (Xu, 2011b). Moreover, Xu & Xu (2005) present evidence of a three-zone pitch range adjustment around focus: expansion under focus, compression after focus (PFC), and limited or no change before focus. For the authors "[t]his three-zone pitch range adjustment is [...] what is unique about focus" (Xu & Xu, 2005: 186). A direct consequence of the three-zone pitch range implementation is that focus is followed by a sharp F_0 drop: this result is compatible with the findings of studies within the AM theoretical framework (see Section 2.5.2) and with the results of the production study in this dissertation (presented in Section 6.3.1 and discussed in Section 9.2.2).

From the point of view of the representation of intonation, Xu adopts the use of time-normalized visualizations of pitch contours for the impressionistic analysis of intonation, rather than ToBI-based annotations. This solution is particularly informative when comparing different realizations under different types of focus, as in the opposition between broad and narrow focus (see Fig.

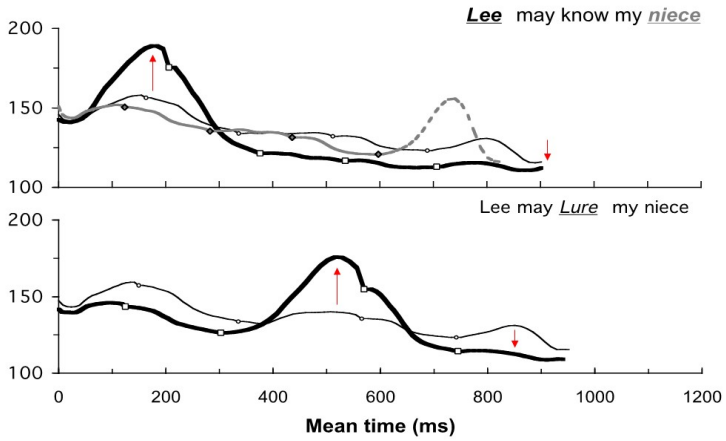

Figure 2.7: Comparison between narrowly focused vs. broadly focused (from Xu & Xu, 2005).

2.7).

The work by Xu is solid and well motivated, firmly based on the acoustic analysis of the signal and on a non-trivial relationship between acoustics and intonational meaning. Nevertheless, the complexity of his mathematical model makes it less accessible, requiring specifically designed speech data sets to be measured with the full range of potentialities.

In another paper, Xu points out that the studies on the prosodic realization of prominence are typically oriented on production or perception, rarely encompassing both (Xu, 2011a), and from this point of view, Xu & Xu (2005) is no exception. A notable change is represented by Breen et al. (2010) who not only present a production study, but also test the results in a perception experiment on human listeners. In order to find the acoustic correlates of information structure in American English, the authors seek to determine if listeners could distinguish focus on the three levels of location, breadth and type (see Section 2.3) only by hearing differences in prosody. The production study presented by the authors is based on speech data collected with an elicitation procedure similar to the ones adopted in Cooper et al. (1985; 1986) and Xu & Xu (2005), consisting of a question-and-answer paradigm to

collect data in controlled contextual situations.

Although the acoustic analysis does not reach the complexity of Xu's model, Breen et al. (2010) explore the signal with a wide set of acoustic measurements extracted from words as focus-bearing units. Among these, the acoustic features which result the best in discriminating the different focus conditions were duration (of a word) plus silence (following the word), mean F_0, maximum F_0 and maximum intensity. The pre- and post-focus pitch range values were not measured. The results show that speakers systematically provide acoustic cues to disambiguate focus location, namely increased duration, higher mean F_0, higher maximum F_0, and higher intensity. Similarly, speakers consistently mark focus breadth with prosody, presenting subtle but noticeable differences in intensity and mean F_0 on the final narrowly focused constituent (the object) when compared to the broadly focused counterpart. As for focus type, speakers were able to differentiate between contrastive and non-contrastive focus only when they were made aware of an explicit ambiguity to solve.

As for the two perception experiments presented in Breen et al. (2010), the results only partially reflect the ones of the production studies. Listeners were successful in distinguishing among focus locations, but failed to discriminate between focus types and between focus breadths. The outcome suggests that listeners cannot directly use the acoustic cues used by the speakers to disambiguate these two levels of focus.

The perception of focus is also assessed in Bishop (2011), presenting a study of a prominence-rating experiment where listeners where asked to distinguish between realizations of the same sentences under broad or narrow-contrastive focus. The results showed that listeners do have knowledge regarding how different focus breadths relate to different patterns of prosodic prominence, as narrowly focused constituents were rated as more prominent than their counterparts under broad focus. However, the author warns the reader against the possibility of "an auditory illusion" (Bishop, 2011: 315):

pre-focal prominence could have been heard as lower, and focused information as more prominent not because of the intrinsic acoustic information, but because of the listeners' expectations for recognizable patterns found in the productions. This is in line with what is reported by Wagner (2005) as top-down interpreting strategy, which can enhance or interfere with the detection of focus (see Section 2.7).

After the advent of intonational phonology, most studies on focus in Italian have been carried out within this theoretical framework. An exception is represented by the research recently carried out by Marotta and associates (e.g., Marotta & Sardelli, 2004; Marotta et al., 2007; Marotta et al., 2012). In particular, Marotta et al. (2012) includes a production and a perception study, where the acoustic realization of prominence is studied across three varieties of Italian. The authors use vowels as prominence-bearing units, first exploring the differences between duration and F_0, and then testing the relative importance of the same acoustic cues in the perception of prominence with resynthesized stimuli. From the point of view of production, duration was confirmed as the most robust acoustic value for prominence in all the three varieties of Italian. However, the interpretation of the results of the perception study was not so straightforward, suggesting that listeners tend to rely more on pitch variations rather than on duration. Nevertheless, these results might have been originated from a bias in the discrimination task, where the original stimuli were paired to stimuli containing vowels with an inverted F_0 contour. This manipulation probably generated unnatural or at least perceptually odd realizations that were easy to discriminate as different from the original.

2.6 The cross-linguistic perspective

The study of prominence and focus marking is particularly interesting when set in a cross-linguistic perspective, since the strategies in marking infor-

mation status vary a great deal across languages, both at structural level (phonology and syntax) and at the level of phonetic implementation (Ladd, 1996, Büring, 2009).

It has been mentioned that prominence-marking strategies in Italian differ significantly from the native English ones. Traditionally, literature has opposed the two languages: while English would consistently mark focus by using prosody, Italian would mainly, if not exclusively, rely on word order strategies. This is the view expressed by Vallduvì (1991) and embraced by Ladd (1996). In particular, Vallduvì (1991) presents a clear-cut division between what he called *plastic* and *non-plastic* languages. Plastic languages are those that can use prosody to differentiate between information status, while non-plastic languages are the ones that rely mostly on word order modification strategies and morphology. Examples of the former group are English and Dutch, while the latter group includes most Romance languages, in particular Spanish and Italian. Two experimental studies carried out by Swerts and colleagues (Swerts et al., 2002, Krahmer & Swerts, 2004), comparing the perception of contrastive and non-contrastive focus by Dutch and Italian listeners seem to confirm this divide between plastic and non-plastic languages: while contrastiveness can be successfully detected by the Dutch listeners only via prosody, the Italian listeners cannot retrieve contrastiveness without the aid of contextual information. This happens both when the listeners were presented with audio stimuli (Swerts et al., 2002), and when they were presented with audio-visual stimuli (Krahmer & Swerts, 2004).

However, recent experimental studies have provided empirical evidence showing that such a sharp distinction between plastic and non-plastic languages is unjustified (see Face & D'Imperio, 2005 for a review). Based on empirical data, Face & D'Imperio (2005) showed that Italian and Spanish use prosody as well as word order modification to mark prominence, although more rarely than in English or Dutch. This finding led the authors to propose a revised version of the traditional model, to be considered more as

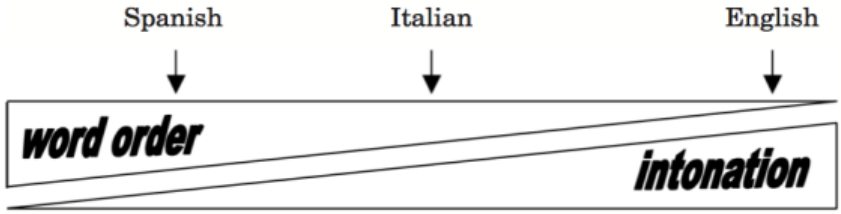

Figure 2.8: Placement of Spanish, Italian and English on the typological continuum (from Face & D'Imperio, 2005).

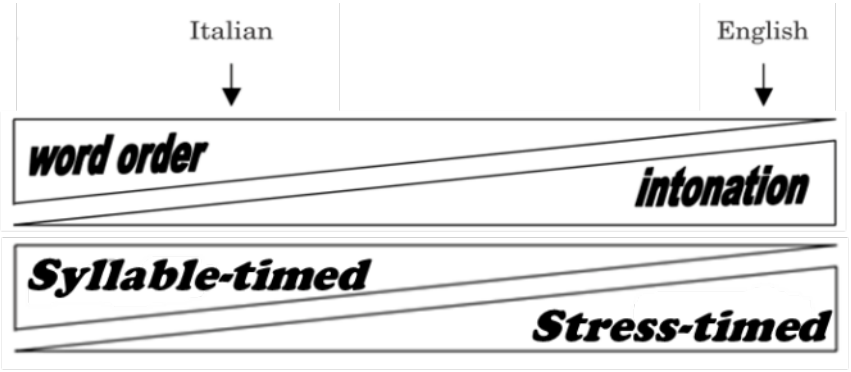

Figure 2.9: Place of Italian and English on the combined continua (from Dauer, 1983 and Face & D'Imperio, 2005).

a *continuum*, rather than a binary opposition, between languages that use word order and languages that use prosody to mark focus. The placement of English and Italian in this *continuum* is represented in Fig. 2.8.

It is interesting to note that this revised model mirrors the evolution of the opposition between stressed-timed and syllable-timed languages based on empirical studies, which was initiated by Dauer (1983) and further supported by studies based on rhythm metrics (cf. Mairano, 2011). A visual combination of the stressed-timed vs. syllable-timed *continuum* and the one proposed by Face & D'Imperio (2005) is proposed in Fig. 2.9.

To the author's knowledge, research on the relationship between the two *continua* has not yet been carried out; this topic deserves further attention in the future.

2.7 Studies on L2 prominence marking

Non-native prosody is a thriving field of research: recent years have witnessed a paradigm shift from the study of segmental phenomena and segmental L2 acquisition to research based on suprasegmental aspects and prosody (Chun, 1998; Busà, 2012). Moreover, research on prosodic transfer (cf. Raisier & Hiligsmann, 2007; Ueyama, 2012) has been growing steadily, especially for L2 English.

In a review of the main results published in the literature, Mennen (2007) reports a list of the most frequently reported errors in the production of L2 English intonation: among these typical errors, at least two are directly connected with the phonetic realization of prominence. Mennen argues that L2 learners have "problems in the correct placement of prominence" and that their productions may present "incorrect pitch on unstressed syllables" (Mennen, 2007: 55), which is typically too high. In the same article, Mennen claims that "[j]ust as a language can have phonemic contrasts [...], the prominence system within a language is also a system of contrasts. [...] Just as phonemes serve to distinguish one word from another word, a system of prominence allows a speaker to contrast the relative importance of words" (Mennen, 2007: 62). In addition to the errors presented by Mennen, it is also shown that there can be errors originated by the cross-linguistic differences in the acoustic cues used to signal prominence between L1 and L2 (Adams & Munro, 1978).

An important contribution to the study of L2 focus marking is Raisier & Hiligsmann (2007). This study is particularly interesting because it is based on the bidirectional L1-L2 combination between a plastic language (Dutch) and a non-plastic one (French). It can therefore be suggested that the results could be replicated in similar studies comparing speakers of English and Italian. As for the methodology, the authors follow an experimental setup similar to the one adopted by Swerts et al. (2002) in their cross-linguistic studies on the perception of contrastive accents in Dutch and Italian. Speakers are pre-

sented with a series of colored geometric figures. Situational contrasts with various combinations between focus and given are created with appropriate question prompts. The results of this production study confirm that learners transfer their prominence-marking strategies from L1 to L2, resulting in overuse of pitch accents, incorrect placement of prominence and incorrect choice of accent type. These results confirmed the initial hypothesis that the fine-detailed phonetics of prosody is more difficult to be learned than its phonology, which is normally acquired later (see Mennen, 2007; Ueyama, 2012).

As for the English-Italian combination, Busà & Stella (2012) and Stella & Busà (2013) have recently carried out research on the intonational variations in focus marking in English L2 spoken by Italians. In their studies, based on the comparative analysis of the production of narrow-contrastive vs. broad focus in Italian and English L2, the authors show that the Italian productions present "a complete transfer of the use of prosodic cues to mark the different pragmatic function" (Busà & Stella, 2012: 35), showing that the values of alignment and scaling are systematically transferred from L1 to L2.

As for perception, studies on the perception of prominence by native vs. non-native speakers of a given language are rare. A notable example is a perception study by Wagner (2005), aimed to test whether the impact of acoustic vs. top-down expectations is different in the disambiguation of focus types for native and non-native speakers of German. The author hypothesizes that native speakers and proficient non-native speakers would rely more on top-down expectations based on their knowledge of the language rather than on the different acoustic cues corresponding to different types of focus. The results confirm the hypotheses, showing once more the contemporary and complex interaction between acoustic factors and other aspects connected with context and discourse.

2.8 Conclusion

This chapter reviewed the main approaches in the study of prominence and prosodic marking of focus, namely the AM theory of intonational phonology and the *direct-relationship* approach. While the former is based on a phonological and categorical vision of the phenomena of intonation and prominence marking (See Section 2.5.2), the latter is aimed to the definition of the acoustic correlates of prosodic functions, based on the quantitative methods and paradigms of acoustic phonetics. It is important to remark that both approaches can coexist, and that the strictly instrumental approach of the *direct-relationship* approach can still be a preliminary foundation for more formal studies within the framework of intonational phonology.

In this study, it was decided to follow the *direct-relationship* approach, because it was deemed more suitable to tackle the problem of the phonetic realization of narrow focus. As mentioned in Section 2.5, the studies on the phonetic realization of narrow focus by Italian speakers of English L2 are very limited (cf. Busà & Stella, 2012 and Stella & Busà, 2013), and it is not even clear whether Italian speakers prosodically mark narrow non-contrastive focus in their L1 (cf. Section 2.3.2). The limited amount of empirical evidence on the topic of this study suggested the adoption of a more parsimonious approach (Breen et al., 2010), which could provide experimental evidence to start studying the problem at its roots, that is, at the acoustic level. This dissertation will therefore tackle the problem of the phonetic realization of narrow focus in English L1 and L2 (and in Italian L1) with the acoustical analysis of speech data and with perception experiments, seeking to define which are the acoustic correlates (if any) that are used to produce and perceive narrow focus.

Chapter 3

Theoretical and methodological issues in the study of L2 prosody

3.1 Introduction

This chapter discusses a series of issues in the study of L2 speech in general and L2 prosody in particular, both in theory and practice.

Section 3.2 will review the main models of L2 speech acquisition, namely the Speech Learning Model (SLM, 3.2.1), the Native Language Magnet (NLM, 3.2.2) and the Perceptual Assimilation Model (PAM, 3.2.3). Section 3.4 will discuss the issues faced by the researchers when attempting to frame the study of the prosody acquisition within the existing models, paying particular attention to the acquisition of the prosodic marking of focus in English L2.

Section 3.5 will move to the discussion of more practical issues involved in the experimental study of L2 speech and foreign accent. The section will discuss the main factors that are involved when carrying out experimental work on the perception of non-native speech, with particular attention to the study of L2 prosody.

Section 3.6 will review the main methods of signal manipulation adopted

in the experimental study of L2 prosody, in particular L2 intonation. The final part of the chapter will review the main methods used to manipulate the acoustic signal in order to study the relevance of the different prosodic aspects in the perception of non-native speech.

Finally, Section 3.7 will conclude the chapter, leading the reader to Chapter 4, which will test several of the methods reviewed here in a series of pilot studies.

3.2 Models of L2 speech acquisition

The acquisition of L2 speech has been studied with increasing interest in the last three decades. The results of extensive experimental studies have been used to formulate several models of L2 speech acquisition (Flege, 1995; Best & Tyler, 1995; Kuhl & Iverson, 1995; Major, 2001; Escudero, 2008; Darcy at al., 2012). These theoretical models were mainly designed to describe and predict the production and perception processes involved in the acquisition of L2 phonemes. The next subsections will review the most widely accepted models used as frameworks of reference for the research on L2 speech acquisition, namely: the Speech Learning Model (SLM, Flege, 1995), the Native Language Magnet (NLM, Kuhl, 1995) and the Perceptual Assimilation Model (PAM, Best & Iverson, 1995).

3.2.1 Speech Learning Model (SLM)

Flege's Speech Learning Model (SLM) was the first organic model of second language phonology learning. The model was built on the basic assumption that many segmental production errors in L2 are likely to have a perceptual basis (Flege, 1995, Flege et al. 1999), and was tested in an extensive series of experimental studies. The SLM is rigorously presented as a set of four postulates and seven hypotheses meant to be "a heuristic for planning research" and for generating "testable predictions" (Flege, 1995: 238). The

four postulates can be summarized as follows: (i) the mechanisms and processes involved in L1 learning remain intact over time and can be used in L2 learning; (ii) language-specific characteristics of speech sounds are stored in phonetic categories, which are long-term memory representations of sounds; (iii) the phonetic categories generated for L1 in childhood evolve over the life span and account for the characteristics of all L1 or L2 speech sounds identified as examples of each category; (iv) the speakers of two or more languages strive to keep the contrasts between L1 and L2 phonetic categories from overlapping in the same phonological space. Seven hypotheses are derived from the postulates to structure the model in more practical terms, all stemming from the central idea that an L2 sound will be easier to learn if it is different enough from the ones in the L1 inventory.

According to the SLM, new phonetic categories will be easier to establish when an L2 sound is perceived as clearly different from L1 phonemes. Conversely, if the perceived phonetic differences are too small, the acquisition of similar sounds will undergo the risk of being prevented by the mechanism of equivalence classification, which was defined by Flege as "a basic cognitive mechanism that permits human to perceive constant categories in the face of the inherent sensory variability found in the many physical exemplars which may instantiate a category" (Flege, 1987: 49). In more practical terms, in the SLM two sounds are considered similar if they have the same IPA symbol in the source and in the target language, and if they differ only at the subphonemic level. For example, /t/ and /d/ are similar sounds in the English-Italian combination: both phonemes are represented with the same IPA symbols in English and Italian, although the place of articulation is different, being alveolar in English, and dental in Italian. Flege argues that a non-native speaker may perceive such speech sounds as perfect substitutes, even though the two sounds deviate measurably from the target norm. As a consequence, the non-native speaker would articulate these sounds following the norms of L1. Their productions may therefore be perceived as

inadequate, or foreign-accented, by L1 listeners.

Another claim is that "cross-linguistic phonetic interference is bidirectional in nature" (Flege, 1995: 241). The consequence of this hypothesis, together with the mentioned filtering effect of equivalence classification, is that an L2 sound "might not be produced exactly as it is produced by native speakers" (Flege, 1995: 243), resulting in a merger of the two concurring sounds.

The SLM was further tested and refined over the years on a vast amount of data, with a variety of language combinations. What remained through the years is the exclusive focus on phoneme acquisition, which makes the model not readily adaptable to account for the acquisition of L2 suprasegmentals.

3.2.2 Native Language Magnet (NLM)

The SLM is mainly aimed at the prediction and explanation of the outcomes of L2 speech perception and acquisition, taking into account speakers' language background and the effect of age on L2 speech acquisition. Kuhl's Native Language Magnet (NLM) model, instead, is devised to go beyond the empirical results and to explore causes at a cognitive level: the thesis driving Kuhl and associates' model is that "language experience alters the mechanisms underlying speech perception, and thus, the mind of the listener" (Kuhl & Iverson, 1995: 121). In this regard, Kuhl had previously claimed that infants are born with a wide and indiscriminate sensitivity to speech sounds, while the culture-bound adults show a much more limited perceptual range for foreign sounds (Kuhl 1993). The reason why phonetic perception changes as a function of the exposure to a language is to be found in a phenomenon called *perceptual magnet* effect.

According to Kuhl & Iverson (1995), the exposure to a certain language causes a distortion of the perceived distance between speech stimuli, so that that language experience *warps* the listener's perceptual space. When acquiring the L1, listeners establish phonetic categories based on phonetic pro-

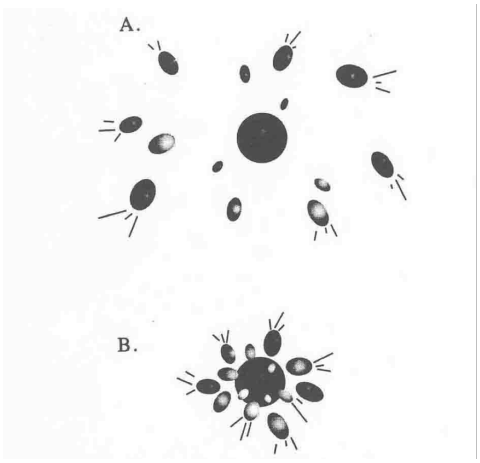

Figure 3.1: The perceptual magnet effect. Stimuli surrounding the phonetic prototype A are perceptually attracted toward the prototype B, warping the perceived distance between prototype and other members of the category (from Kuhl & Iverson, 1995).

totypes, that is, particularly good instances of categories. These prototypes work as *perceptual magnets* for other sounds in the category, which are recognized as exemplars of the category by being *attracted* by the good instances stored in the listener's memory (see Fig. 3.1).

The application of the perceptual magnet model to L2 speech perception studies led to the formulation of the NLM, which is based on the assumption that L2 language perception and acquisition are affected by the L1 perceptual magnets. Experimental data showed that the exposure to language in early life produces a change in the perceived distances between speech sounds: the perceptual magnet effect can be seen already in 6-month-old infants, and it gets stronger in adult age. The model was tested in adult listeners both for vowels and consonants, and in a variety of language combinations, showing that certain categorical distinctions are maximized near the boundaries between two phonetic categories (or magnets), while others are minimized when near the center of the category, resulting in the assimilation of similar sounds to the perceptual magnets. In other words, the L2 sounds that adult

listeners perceive as being similar to their L1 phonetic categories are more difficult to discriminate from the native-language counterpart, while different sounds will be easier to identify. This is in line with the SLM (Flege, 1995, see Section 3.2.1) and the PAM (Best, 1995, see Section 3.2.3).

It is interesting to point out that the experimental data suggest that the perceptual space can be reconfigured even in the adult age: the sensory ability to discriminate contrasts is still present, but instead of being immediate, as in infants, it needs to be trained. This finding is also compatible with the first postulate of the SLM (see Section 3.2.1), which claims that L2 speech acquisition is possible throughout the life span of an individual and is not limited to a critical period.

3.2.3 Perceptual Assimilation Model (PAM)

The third model presented here is the Perceptual Assimilation Model (PAM) (Best, 1995; Best & Tyler, 2001). Like the previous two models, the PAM is based on the concepts of phonetic category separation and similarity between L1 and L2 sounds. However, the PAM differs from the other two proposals in defining similarity in terms of gestural configurations rather than in terms of acoustic cues in the signal. The PAM is based on the direct realist theory, which considers the epistemological process as a direct, not mediated, acquisition of perceptual objects rather than through their representation (Best, 1995). As in the motor theory (cf. Perkell et al., 2000), speech perceptual primitives are considered as gestures, and not as acoustic information decoded by the auditory system. From the point of view of L2 perception and learning, the simple gestures that are not present in the native space need to be assimilated. Non-native segments tend to be perceived according to their similarities to, and differences from, the gestural constellations characterizing the L1 phonological space.

The PAM also differs from the SLM because it is mainly thought to account for patterns of L2 segmental perception by naïve listeners with limited

or no experience with the L2, while the SLM is focused on the acquisition achieved by L2 advanced learners. In fact, the PAM was only recently extended to the prediction of the behavior of more advanced L2 learners with the label PAM-L2 (Best & Tyler, 2007).

According to the PAM, perceptual objects can be assimilated to a native category in three ways: as a categorized exemplar of a native phone (on a 1-7 goodness scale) (C); as an uncategorized phone that falls in between two native categories (i.e., similar to more than 2 native phones) (U); as a non-assimilable speech sounds that bears no resemblance to any phone in the L1 system (N). Phonological contrasts between two non-native speech sounds can be assimilated to L1 categories following six pairwise assimilation types depending on how each member of the contrast is assimilated: TC (two-category assimilation), when both members of the contrast can be assimilated to a different category in L; SC (single-category assimilation), when both target sounds are assimilated to a single L1 sound; CG (category goodness difference), similar to SC, but here one sound fits an L1 category better than the other; UC (uncategorized-categorized), when only one member fits an L1 category; UU (both uncategorized): when neither sound fits an L1 category; NA (non-assimilable): when both L2 sounds are perceived as non-speech. The PAM predicts that discrimination between two target sounds is very good if they are perceived as the same as an L1 contrast (TC); slightly lower but still good if the two sounds are perceived phonetically as good versus poor samples of the same L1 phoneme (CG); much lower if both sounds are perceived as equally good or equally poor tokens of one L1 phoneme (SC). Even if the theoretical assumptions are different from the SLM, one can see how the models agree when predicting that the phonetic difference between L1 and L2 sounds facilitates the assimilation of new sounds, while similarity hinders it.

As for the compatibility with NLM findings, results from experimental studies based on the PAM seem to disprove the existence of a perceptual

magnet effect, showing that very good discrimination of L2 contrasts is still possible even when they are close to L1 prototypes, although with lower success than with native contrasts.

3.3 L2 speech models and the acquisition of prosody

All the current models of L2 speech acquisition are based on the study of the perception and acquisition of L2 phonemic inventories. It is not clear whether the models could be adapted to generate predictions and provide explanations for the processes characterizing L2 prosody acquisition. Certainly, such adaptation is not a trivial task, because of the great differences in the nature of the suprasegmental aspects of speech as compared to the segmental aspects.

First of all, most of the experimental studies based on the current L2 acquisition models consist of perception tests where subjects are asked to identify or discriminate single phones, presented without any contextual information (Strange, 1995). This approach cannot be directly applied to the study of the prosodic dimension for a variety of reasons.

First of all, prosodic information is coded in bundles of acoustic cues (F_0, duration, intensity, spectral structure). These acoustic cues interact with each other and with the segmental information at the same time, so that "all the parameters of speech melody, local and global, are perceived in an integrated way" (Vaissière, 2005: 239). As a consequence, prosodic features are perceived in relation to their surrounding context. This context can be seen in strictly phonetic terms, that is, as the information that surrounds a sound, but also as the wider context of communication. As for the phonetic context, the relative nature of prominence implies that a prominent constituent can only be perceived as such when the constituent is judged in relation to the neighboring information (see Section 2.2). For example, a prominent word

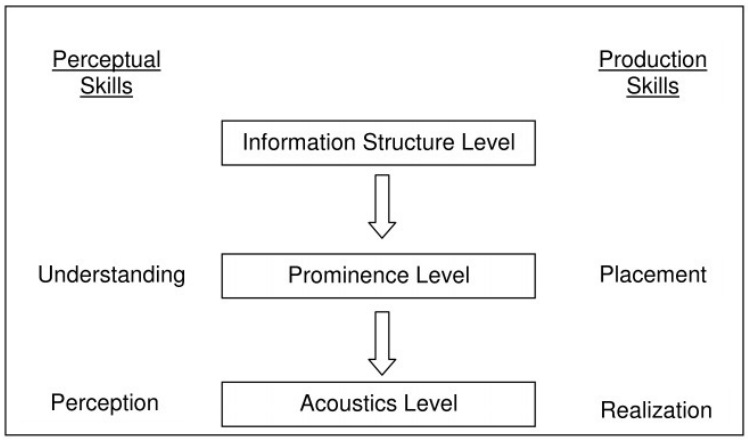

Figure 3.2: Chart showing the three levels of prosodic focus marking and the relationships between them (from Baker, 2010).

or syllable cannot be perceived as such if it is not presented within a wider contrast where it would stand out against the background of given material. As for the communication context, it has been mentioned that prosody can have many functions and many levels of meaning (see Section 1.1).

Baker (2010) proposed a model where prosodic focus marking is conveyed on three levels of meaning, which is represented in Fig. 3.2.

First, at the *information structure level*, the speakers determine which words are in focus and which words represent the background material. At this level, prosody interacts with the syntactic and pragmatic systems. Second, at the *prominence level*, the speakers determine how both information in focus and background information should be realized within the syntactic, morphological, and prosodic structures of a language. In English this is done by selecting a word or words to be marked with pitch accents, and by selecting the type of pitch accents (e.g., contrastive or non-contrastive) that will be used to mark focus. Third, at the *acoustics level*, speakers manipulate certain acoustic cues to realize the prosodic structures that were selected at the prominence level. Baker's model clearly shows how prominence marking cannot be studied while ignoring the interaction of the many domains

involved in the process.

However, researchers have recently claimed that one of the shared basic assumptions of the current L2 speech acquisition models can also be applied to the study of intonation. This assumption is the process of categorical distinction and category formation that shapes the perception of non-native contrasts. In this regard, it has been claimed that the AM intonational models (see Section 2.5.1) and transcription systems like ToBI (Silverman et al., 1992) allow for "a category-based interpretation of intonation that is compatible with the leading theories of second language acquisition [...], which are segment-based" (Jilka, 2007: 82). Moreover, by adopting the intonational phonology framework, which separates the phonological and the phonetic domains, one can identify non-native deviations from the norm both in terms of transfers of different tonal categories, but also at the level of deviating phonetic implementations (Ladd, 1996). The adaptation of the L2 speech acquisition models can be potentially achieved not only for the implementation of intonation, but also for prominence marking. In this regard, Mennen (2007) claimed that the prominence system of a language could be seen as a system of contrasts comparable to the set of phonemic contrasts within a language (see Section 2.7). Mennen (1999) also found some compatibility of the study of intonation with the SLM, showing that Dutch learners of Greek L2 were more successful at producing new pitch contours when there was no counterpart in L1. These results would confirm that similarity is more problematic than difference also for the acquisition of new pitch contours, in accordance with the SLM.

In recent works, Gili Fivela (2012), has also suggested that the predictions of the L2-PAM (Best & Tyler, 2007) could be adapted to the study of phonetic aspects of prosody, like alignment and scaling, which are supposed to be identified categorically by listeners. The results of the first experiments in this regard, where Italian native listeners are asked to judge Italian sentences with native versus non-native (English) prosody and where con-

textual information was provided, seem to confirm this compatibility (Gili Fivela, 2012).

To conclude, more research is needed to find a consistent way to fit the description of prosody acquisition within the framework of the existing L2 speech acquisition models. More research is needed to formulate new models, or accommodate the current ones so that they can predict and explain the mechanisms involved in L2 prosody acquisition.

3.4 Practical issues in the study of L2 speech and foreign accent

The SLM, NLM and PAM agree in showing that L2 speech acquisition is difficult to be achieved completely, resulting in differences in pronunciation between native and non-native speakers. A direct consequence of these differences in pronunciation is the production and perception of foreign accent. Foreign accent (FA) has been defined as "a set of pronunciation patterns, at both segmental and suprasegmental levels, which differ from pronunciation patterns found in the speech of native speakers" (Volín & Skarnitzl, 2010: 1010), or as "speech which differs acoustically from the native phonetic norm, and is auditorily detectable by native speakers" (Wayland, 1997: 346). The notion of FA is therefore based on a systematic contraposition between non-native speakers' speech, which can diverge to a certain extent from the native norm, and the native speakers' speech, which is considered as the standard of reference. Consequently, research on FA is often particularly oriented to perception, and a crucial role is played by native listeners' judgments of foreign-accented speech (Derwing & Munro, 2009), so that listeners' judgments are required at some level of the analysis, even when a study is not specifically aimed to the perceptual domain (McCullogh, 2013). The results of listeners' judgments are normally correlated with a series of linguistic and cognitive factors (see Section 3.4) and generalizations are drawn.

As for the nature of the judgments, listeners can be asked to rate a variety of aspects of L2 speech along a variety of dimensions. In this regard, Munro & Derwing (1995; Derwing & Munro 1997; 2009) have established three specific constructs to assess non-native speech: accentedness, intelligibility and comprehensibility.

Accentedness is understood as "how different a pattern of speech sounds compared to the local variety" of the target language (Derwing & Munro 2009: 478), and it basically corresponds to a narrow definition of FA as speech characterized by perceivable deviations from a native phonological norm. The rating of accentedness is normally based on the listener's global judgment of stimuli.

Intelligibility is "the degree of a listener's actual comprehension of an utterance" (Derwing & Munro, 2009: 479), that is, the extent to which a native listener understands the meaning as intended by the speaker. Being based on the correspondence between speech and meaning, intelligibility is mainly carried by segmental information (Wang et al., 2011). Typical methods to test intelligibility are dictation tasks where native listeners are asked to transcribe what they hear, and the resulting transcriptions are then compared to the original texts to verify how much of the message intended by the speaker is successfully understood by the listener.

Finally, comprehensibility is defined as "the listener's perception of how easy or difficult it is to understand a given speech sample" (Munro & Derwing, 1995: 478) or the "perception of intelligibility" (Derwing & Munro, 1997: 2). This dimension is also tested with the listeners' global judgments.

Munro and Derwing have based many of their studies on the comparison and correlation of listeners' judgments along the three dimensions, finding that the relation between the three constructs is not always direct and, for example, that "the presence of a strong foreign accent does not necessarily result in reduced intelligibility or comprehensibility" (Munro & Derwing, 1995: 90). When studying L2 speech and FA, the amount of variability normally

characterizing any empirical study in phonetics is amplified (Munro, 2008). In particular, factors of variation can depend on the speakers, the listeners, the experimental procedure, and the speech materials used in the studies. The next subsections will review the practical issues connected to each one of these aspects.

3.4.1 Speakers

The learners' pronunciation depends on a wide range of linguistic and cognitive factors. These factors include the age of learning, the length of residence in the country where L2 is spoken, and the frequency of use of L1 (see Bohn, 1995; Munro, 2008). All these factors need to be adequately controlled in order to obtain homogeneous groups of L2 speakers.

Empirical studies on FA normally require the presence of at least two groups: one group of L2 speakers, representing the experimental group, and one group of native speakers, working as the control group. The inclusion of a control group of native speakers serves the purpose of providing reference data for the native-speaker norms. The data are collected within the same experimental paradigm used to elicit data from the L2 speakers. The resulting data set is promptly comparable to the productions of the group, or groups, of L2 speakers. Furthermore, native groups may also serve the practical purpose of testing the reliability of native judges during perception tasks: those who are not able to identify native speech are normally considered outliers and therefore discarded before any statistical analysis of the results is carried out.

One of the main sources of inter-speaker variation is represented by the regional varieties of the languages studied. For example, considering the two languages that will be the object of the present investigation, i.e., Italian and English, the amount of variation depending on the speakers' geographical origin is very wide. For Italian, the socio-cultural variation between regional varieties is enormous, especially at the level of prosody (Sorianello, 2006;

Marotta, 2008). For British English, the recent empirical studies published within the IViE Project (Grabe, 2004) have found a great deal of variation in intonation not only between different regional varieties, but also within the Southern Standard British English (SSBE). This is the reference variety for the English spoken in Great Britain and it is also the variety that will be studied in this dissertation. Therefore, when dealing with the study of prosody, control must be particularly tight.

The definition of level groups is particularly important and it is normally achieved by combining a variety of instruments. Surveys and questionnaires can be used to collect metadata regarding age, age of learning, length of residence and self-evaluation of L2 competence. Another strategy to define levels is collecting FA rating scores from a panel of native judges. These are asked to globally evaluate the accentedness of the L2 speakers' productions in perception tests (e.g., Busà, 1995). Finally, the determination of level groups might also include vocabulary tests (Darcy et al., 2013) or oral competence tests (Baker, 2010) as diagnostic indexes of non-native speakers' competence in L2. Obviously, the best results in determining a homogeneous group are achieved by combining as many of these methods as it is possible. In this dissertation, the definition of level groups will be based on a vocabulary size test and on a perception test where native listeners were asked to rate the accentedness of the L2 speakers' productions (see Section 5.2.1.3).

Beside these issues, the researcher has to pay attention to the levels of variation present in any experimental study in phonetics. It will be therefore necessary to build groups of speakers directly comparable in terms of variables such as age, gender, level of instruction, and health conditions, depending on the purpose of the study.

Another question concerns the number of speakers to consider. This might range from as few as one to 240 (Jesney, 2004). However, the researcher has to keep in mind that while a big set of speakers provides a higher potential for generalization, it also increases variation and therefore

the risk of obtaining spurious results.

3.4.2 Listeners

"One dimension that listeners are amazingly sensitive to is the presence or absence of a foreign accent" (Derwing & Munro, 2009: 477). This assertion explains why the pièces of resistance of most experimental studies on FA are perception tests involving the presentation of audio stimuli to listeners. These are typically native speakers of the target language, who are asked to identify or rate non-native speech for intelligibility, accentedness, or comprehensibility. Native listeners' fine-grained sensitivity to foreign-accented speech is well known (cf. Flege, 1984), and it is thought to be the key to understanding the relative importance of the many acoustic cues contributing to creating FA (Derwing & Munro, 2009).

A first listener-based factor to be taken into account, and controlled for, is native listeners' potential familiarity with non-native speakers' source language and with the characteristics of their FA in L2. It has been demonstrated that such familiarity can affect native listeners' judgments (Gass & Varonis, 1984), so listeners with no formal knowledge or regular contact with speakers' L1 should be selected. Building on the idea that familiarity with a linguistic background helps FA detection, it has been argued that L2 speech produced by speakers sharing the same L1 background could result more intelligible to non-native speakers. In this regard, Bent & Bradlow (2003) proposed the Interlanguage Speech Intelligibility Benefit (ISIB) hypothesis. The core of the ISIB is that non-native listeners would find that L2 speech produced by other non-native speakers is more intelligible than the speech produced by native speakers of the target language (matched ISIB). In addition, non-native speech in a target language would be more intelligible to L2 listeners, no matter the L1 background (mismatched ISIB). The central idea is that, regardless of native language background, "certain features of non-native speech will make non-native talkers more intelligible to all non-native

listeners" (Bent & Bradlow, 2003: 1602), such as the absence of connected speech phenomena (e.g. vowel reduction, assimilation) or slower speech rate.

However, several studies designed to replicate the ISIB effect shown in Bent & Bradlow (2003) found contradictory evidence, doubting the validity of the ISIB hypothesis (see Munro et al., 2006). In addition, a more sophisticated statistical approach suggested by Hongyan & Van Heuven (2007) to the data set presented in Bent & Bradlow (2003) showed that even in the original results it is questionable whether the fact that non-native speakers and listeners have different native languages is a benefit or a hindrance. So far, ISIB is an interesting possibility, but it needs more evidence not to be rejected. Another frequently debated issue regards the choice to use phonetically trained judges, such as language instructors or phoneticians, or naïve native listeners. While there are studies showing more inter-rater reliability for expert listeners, it may as well be argued that phonetic expertise could also represent a bias (Derwing & Munro, 2009; McCullogh, 2013). Moreover, the use of naïve listeners may be more representative of the processes involved in natural communication context and can be considered more generalizable.

Besides the issues presented here, it is always advisable to control for homogeneity in listeners too, even though one can be more lenient than when dealing with speakers. For example, listeners using a different variety of the same target language can still identify native productions in the most prestigious varieties of their L1, to which they have been normally exposed in school and through the media, as asserted by Grabe et al. (2008) with respect to the perception of SSBE by speakers from the North of England.

3.4.3 Experimental tasks

As mentioned in Section 3.4, in order to test hypotheses based on production, studies on FA often include perception tests, where native listeners are asked to give a behavioral response to the stimuli they are presented. Gili Fivela (2012) divides the types of perception tasks in *metalinguistic judgments* and

response and action taking tasks. The former include all kinds of tasks where a listener is asked to judge stimuli after being explicitly instructed to focus on particular aspects of the speech samples. Accent-rating or language identification tasks fall under this label. The latter type of perception task is based on tasks where subjects are asked to react without reflecting on the type of response by performing some kind of immediate action. The required actions can range from imitation tasks and delayed repetition tasks (see Piske et al., 2001), where subjects are asked to repeat stimuli, to tasks where subjects are asked to select pictures matching auditory stimuli. When performing these actions, reaction times are collected, being representative of the cognitive load required in processing the different stimuli: the longer the reaction time, the more difficult the task.

The experimental paradigms of *gating* and *shadowing* are among the action-taking tasks that could be required from subjects. Gating consists in presenting the subject with progressively longer couples of segments (*gates*) cut from base stimuli that are representative of two categories, in order to check how much information is needed to identify a category from the other (Grosjean, 1980). Face (2007) and Petrone (2008) recently applied this paradigm to the study of L1 prosody with interesting results. The shadowing task requires listeners to repeat a stimulus once they have recognized it; listeners' reaction time is measured (Slowiaczek, 1994). This procedure was recently used in a study on stress placement and vowel reduction in English L2 spoken by native speakers of French and Italian (Le Page & Busà, in press).

Other ways to assess the cognitive load in processing L2 speech is through the use of eye-tracking, a technique that records data on gaze direction and fixation duration, or neuroimaging techniques, such as Event Related Potentials (ERPs) or functional Magnetic Resonance Imaging (fMRI). These methods are mainly used in studies on L2 lexical representation (e.g., Mitterer, 2011).

Another important issue when dealing with FA perception is the choice of the right instrument to rate accentedness or comprehensibility. Most studies have been based on the use of Likert scales, ranging from three to ten points, with a marked preference for nine-point scales (cf. Piske et al., 2001, Jesney, 2004). However, other studies have adopted the use of sliding scales (Major, 1987; Flege & Fletcher, 1992; Jilka, 2000; Rognoni & Busà, in press), where raters are asked to adjust the position of a lever, or a handle, along a *continuum* where only the extremes are marked. The position marked by the rater is then converted in numeric values by a program. With this approach, even finer distinctions can be obtained (up to 0-100, or even 0-256 ranges). At the same time, judges need to be specifically instructed and trained on how to use sliding scales, as they are not fully aware of the individual gradients (Jilka, 2000).

3.4.4 Speech material

The range of stimuli presented in studies of non-native speech perception is vast and it depends on the purpose and the theoretical models adopted by the researcher. Typically, experiments aimed to the perception of L2 phonemes are based on the identification and/or discrimination of phones, providing subjects with little or no contextual information. In contrast, studies on L2 prosodic aspects typically focus on longer stretches of speech, normally aiming for global judgments or ratings of non-native speech at word or sentence level.

When collecting data for production and perception studies, one important issue is their ecological validity, or their naturalness. Theoretically, recurring to spontaneous speech would be the best choice to explain what really happens in face-to-face interactions, but uncontrolled speech would also bring in a great deal of variation, not only at the inter-speaker level (see Section 3.4.1), but also along other dimensions such as communication context, style (diaphasic or inter-style dimension, see Marotta, 2008) and

attention (Flege, 1987; Hincks, 2005). On the other side, the so-called lab speech may lack the naturalness of real-life speech but it has the advantage of being highly controlled, resulting in highly comparable speech samples presenting a reduced amount of variation.

As for the collection of speech samples, the literature offers a plethora of data elicitation tasks that can be organized in a *continuum* (Face, 2003), ranging from reading carrier sentences or longer bits of a written text, to freer tasks, including direct or delayed repetition of items (see Piske et al., 2001; Trofimovich & Baker, 2006), map-tasks (see Anderson et al., 1991) card games (Rasier & Hiligsmann, 2007), the retelling of a story or a cartoon (Derwing & Munro, 2012), extemporaneous speech (Elliott, 1995; Thompson, 1991). All these tasks can be prompted by written instructions or by other kinds of audio-visual prompts. However, it has been demonstrated that highly controlled speech, such as read speech, is not acoustically different from less controlled conditions of speech and that controlled speech can be still considered a useful starting point for generalizing findings to real-life speech (Face, 2003; Zipp & Dellwo, 2011).

Another issue connected with the naturalness and ecological validity of the speech materials is the use of natural versus synthetic or acoustically manipulated stimuli in perception tests. This is a particularly important issue in the study of prosody. Given its relevance to the topic of this dissertation, this issue will be discussed in detail in Section 3.5, which will also review the main manipulation techniques that are applied in non-native prosody studies..

3.5 Signal manipulation techniques: resynthesis of stimuli

Synthetic speech made its first entry in the field of L2 speech perception with parametric speech synthesis (Strange, 1995). This type of speech syn-

thesis is based on the creation of speech sounds starting from the numeric expression of the acoustic phenomena involved (cf. Klatt, 1980). This technique produces stimuli where virtually any acoustic parameter (e.g., formant structure, F_0, frication noise...) can be manipulated. While this method is good for identification and discrimination tests based on speech without context, parametric synthesis cannot be used with sentence-length stimuli as it generates highly unnatural stimuli.

In the last thirty years, technological advances have redefined the range of possibilities in the manipulation of the acoustic signal. User-friendly and multi-platform signal analysis packages have often been developed as open-source or freeware software for research purposes. It is the case of *Praat* (Boersma & Weenink, 2013), *Wavesurfer* (Medina & Solorio, 2006) and *Tandem-Straight* (Kawahara, 2008). Parametric speech synthesis has been replaced by the acoustical manipulation of speech and the resynthesis of the recorded speech signal. In particular, the development of speech processing algorithms such as the *PSOLA* (Moulines & Charpentier, 1990) has allowed selective control over one or more acoustic factors in the speech samples recorded by actual speakers.

The main problem when testing the impact of the single prosodic aspects is that, in natural speech, prosody cannot be separated from the segmental dimension. One way to separate the concurring streams of information in natural speech is recurring to acoustically manipulated, or resynthesized, speech. The speech signal can be digitally manipulated to degrade or remove some parts of the information while preserving others. As a result, the resynthesized stimuli allow researchers "to systematically change one parameter at a time, such as F_0, which represents a clear advantage over natural speech production for evaluating the contribution of each individual parameter" (Vaissière, 2005: 241).

The tradeoff of the application of resynthesis techniques is the difficulty to obtain fine-grained judgments from the listeners. The judgment of nat-

ural speech enables rating along global dimensions, such as intelligibility, accentedness and comprehensibility, counting on the native listeners' high sensitivity to foreign-accented speech (see Section 3.4). This is because, in natural speech, the process of global listening and rating is facilitated by the redundancy of many contemporary acoustic cues, both at segmental and at suprasegmental level. In contrast, when listening to severely manipulated speech sample, the listeners can rely on a smaller amount of information, and, as a result, their sensitivity is limited to more general tasks, such as language identification or FA detection, rather than FA rating (Munro, 1995; Munro et al., 2010). In addition, it is important to mention that there is always a chance that the results of perception tests based on heavily manipulated stimuli might not exactly reflect the impression that a listener could have when listening to the kind of speech that naturally occurs in face-to-face conversation.

The next subsections will review the main resynthesis techniques adopted in the study of L2 prosody perception. Section 3.5.1 will discuss delexicalization techniques, which are meant to neutralize, or limit, the effects of segmental information, and are among the most frequently used manipulation methods (see Munro et al. 2010). Section 3.5.2 will present the method of monotonization, which is used to neutralize the effects of F_0, resulting in monotone stimuli characterized by a flat pitch contour. Section 3.5.3 will discuss the lack of a standardized method to neutralize the effects of segmental duration and rhythmic patterns, presenting some possible solutions to test the impact of these cues on FA perception and rating. Finally, Section 3.5.4 will present the prosody transplantation method, which has been recently used with success in various studies of L2 prosody perception.

3.5.1 Delexicalization

A quite extensive set of signal manipulation methods used in the study of L2 prosody has been labeled delexicalization, or content-masking techniques.

These techniques are based on the application of various technological tools to remove or degrade part of the segmental information that is present in the speech signal, making it unintelligible. As a result, speech is stripped from the lexical meaning normally conveyed by the segmental information, while the residual prosodic information remains untouched. One of the first studies using delexicalized stimuli in a cross-language identification task was Ohala & Gilbert (1981), where it was shown that the residual prosodic information was enough for the speakers to identify languages well above chance level in a forced-choice task based on 'hummed' stimuli presenting no segmental information.

One of the most frequently adopted delexicalization techniques is low-pass filtering. With this method, the frequencies composing the speech signal are band-filtered at a fixed cut-off frequency. In the resulting speech signal all the information regarding the fundamental frequencies and the first harmonics is retained, while the highest bands of frequencies are eliminated. From the auditory point of view, low-passed filtered stimuli sound like muffled speech, similar to the sound of speech through a thin wall or a door. Fig. 3.3 shows a visual representation of how low-pass filtering affects a speech sample, where only the lower frequencies are preserved and the higher frequencies are cut off.

Other delexicalization methods include reverse speech and cross-splicing (Munro et al., 2010) or the application of methods comparable to low-pass filtering (e.g., Portele & Sonntag, 1997). However, as already mentioned that delexicalized stimuli have the strong disadvantage of severely reducing the sensitiveness of listeners to FA. Since fine-grained distinctions are obviously difficult to make when judging degraded speech, forced-choice tasks are usually preferred to FA rating tasks. Content-masked stimuli therefore result more suitable for language identification tasks (Ohala & Gilbert, 1981; Ramus & Mehler, 1999), native/non-native status detection (Rognoni, 2012) or attitude judgments (Signorello et al., 2012).

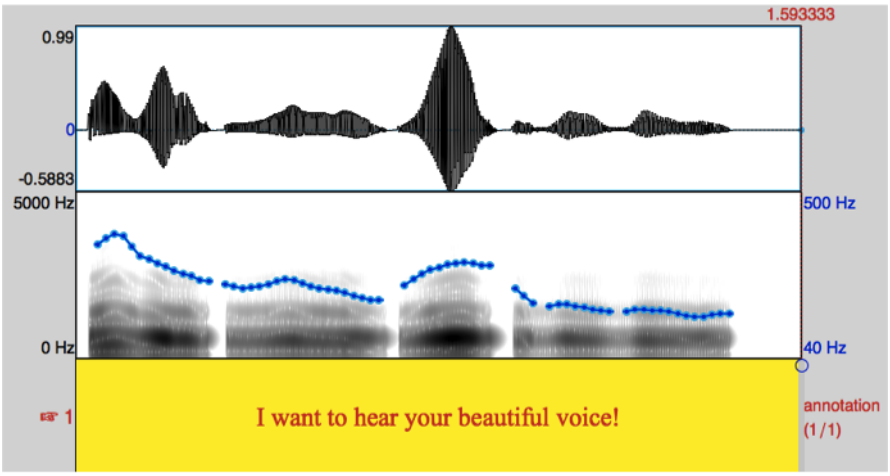

Figure 3.3: Example of a low-pass filtered speech sample. The frequencies that are higher than the cut-off value are eliminated from the signal, while the lower frequencies remain intact.

Another serious drawback of delexicalization techniques is represented by what is left in the residual information. Even if intelligibility is lost, the residue can still include a variety of different clues for accentedness (Munro, 1995). First of all, traces of the segmental information (e.g., the succession of voiced and devoiced sounds, and, to a certain extent, vowels and consonants) may still be present and guide the listeners' judgment. Moreover, the prosodic cues that are left in the signal are multiple and still entangled one with another: not only is it impossible to tell the relative importance of duration, intensity and F_0, but it is also difficult to rank the importance of intonational (e.g., events connected with the F_0 contour, such as pitch range) versus temporal aspects of prosody (e.g., rhythmic structure and speech rate).

3.5.2 Monotonization

Another way to separate the segmental and suprasegmental levels of information is approaching the problem from the opposite direction, that is, by removing or strongly limiting the influence of prosodic aspects. Pitch mono-

tonization has been often used to neutralize the influence of pitch in the signal (Van Els & de Bot, 1987; Jilka, 2000; Rognoni, 2012). With this method, the F_0 contour is resynthesized at a fixed frequency value set by the researcher (e.g., 220 Hz, Jilka 2000), resulting in monotone speech samples where the rises and falls of melody are completely neutralized. Fig. 3.4 shows how the resynthesized pitch contour in a monotonized stimulus results in a flat line at a fixed value.

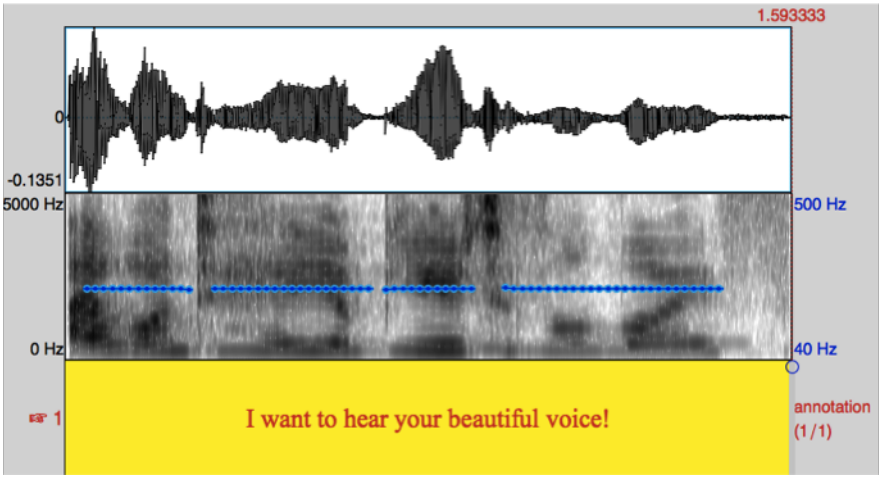

Figure 3.4: Example of a monotonized speech sample. The pitch contour is flattened to a fixed value.

Like low-pass filtering, this technique presents strong limitations. First, the residual segmental information is usually enough to betray the non-native background of L2 speakers (Van Els & de Bot, 1987, Rognoni, 2012). Second, the manipulation only involves the F_0 contour, factoring out the prosodic aspects involved in the melody (e.g. pitch patterns and pitch range), but not the ones involved in the temporal dimension (e.g. rhythm and speech rate). Third, from the perceptive point of view, a flattened pitch contour results particularly unnatural because it lacks the progressive physiological fall in F_0 and intensity known as declination (t'Hart et al., 1990).

3.5.3 Neutralized duration

Differently from delexicalization and monotonization, there is no standardized signal manipulation method specifically aimed to systematically neutralize the differences in duration between the segments in a speech sample. Ideally, it should be possible to manipulate duration similarly to what can be done for the segmental information and F_0 with delexicalization and monotonization, respectively. The resulting stimuli would present all the phones, or a selected set of them, with a fixed length that can be set by the researcher. Such a method would be particularly useful to neutralize the effect of vowel length, which is one of the main phonetic cues to betray Italian accent in English (cf. Busà, 1995; Flege et al., 1999; Azzaro, 2006), or geminate consonants.

The manipulation of duration can be straightforwardly executed with programs like Praat with *PSOLA* or *LPC* synthesis. For example, Tajima et al. (1997) and Magen (1998) studied the effect of segmental duration in FA perception by using resynthesized stimuli where the duration values of vowels produced by native speakers had been superposed to non-native speakers' productions and vice versa. However, the results of such an application can be limited to minimal pairs of vowels that only differ in length. When dealing with vowels that also differ in their spectral structures, the results would be very unnatural and would present artifacts. For example, just stretching the schwa in a word like to [tə] in connected speech would not result in the full vowel that is pronounced when uttering the word to [tu:] in isolated or careful speech. Conversely, the effects of centralization could not be replicated by simply compressing the length of a full vowel. To sum up, it would be necessary to use a method where reduction could be accounted for on both dimensions. A possible solution to this problem is to combine the manipulation of duration with speech synthesis, where vowel sounds can be generated by rule, following the input of the researcher in terms of duration and spectral structure.

In a study on Dutch synthetic speech, Drullman & Collier (1991) used a semi-automatic TTS (text-to-speech) speech synthesis module to create stimuli where the parameters of duration and quality of the vowels could be set in advance. In the resulting synthetic stimuli, syllable duration was neutralized and vowel quality preserved. However, to the author's knowledge no attempt has been made to adapt such a method to cross-linguistic studies. An implementation of a similar method to generate duration-neutralized stimuli was attempted by the author in a pilot study presented in Chapter 4 with inconclusive results (see Section 4.3 and subsections).

Recent cross-linguistic studies have attempted to determine the impact of segmental duration indirectly, that is by using stimuli that were modified with a combination of delexicalization and monotonization techniques. Fig. 3.5 shows the result of the application of the two methods on a speech sample, where only temporal information is available. The scores obtained with stimuli generated in this way were then compared to the ones modified by applying only one of the two manipulations (delexicalization or monotonization) in order to determine the effect of the residual temporal cues in the signal. With this approach, the impact of temporal aspects is therefore not calculated directly, so the results must be considered with caution.

Another method that has been used in cross-linguistic studies in the perception of rhythm and segmental duration is the generation of *SASASA* stimuli (e.g., Mairano, 2011; Gut, 2012), where all the consonants are replaced with a synthetized [s] and all vowels with a synthesized [a], following Ramus & Mehler (1999). The peculiarity of this method is that it preserves some of the information regarding the syllable structure of the original speech samples, while masking the content like the delexicalization methods presented in Section 3.5.1.

A possible solution to the limitation in the listeners' sensitivity to FA caused by the manipulation techniques presented so far is the adoption of the prosody transplantation paradigm, which will presented in the next section.

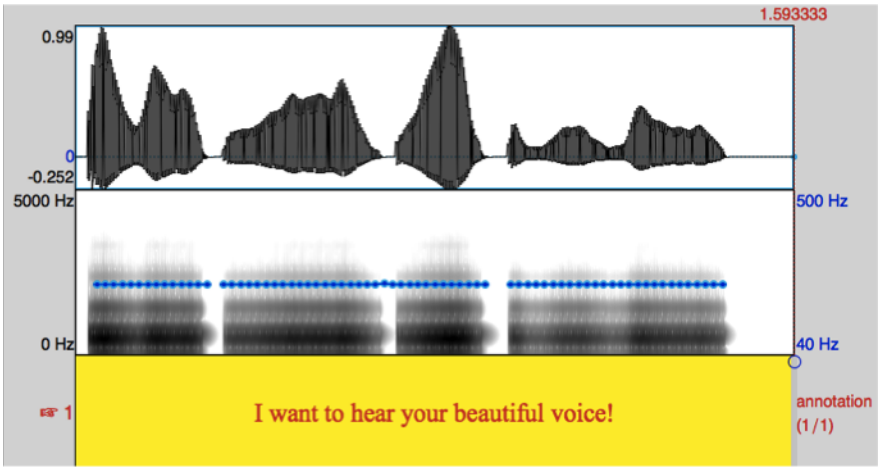

Figure 3.5: Example of a speech sample resynthesized by combining low-pass filtering and monotonization. The frequencies that are higher than the cut-off value are eliminated from the signal, and the pitch contour is flattened to a fixed value.

3.5.4 Prosody transplantation

The basic principle of prosody transplantation is that the prosodic aspects of a native speaker's production can be imposed on non-native segments, and vice versa. This makes it possible to maintain perfectly intelligible stimuli while selectively manipulating prosodic cues. The resulting stimuli can still present artifacts, but they sound more natural than the delexicalized or monotonized ones, and they allow listeners to resort to their fine-grained sensitivity in rating foreign-accented speech.

Prosody transplantation, also referred to as *prosody cloning* (Yoon, 2007) or *prosodic transplantation* (Gili Fivela, 2012), has been recently applied in a many experimental studies on L2 prosody and FA (cf. Rognoni & Busà, in press, for a review). The method has been applied for a variety of purposes, ranging from the determination of the relative importance of prosodic cues in FA rating and detection (Boula de Mareüil & Vieru-Dimulescu, 2006; Rognoni & Busà, in press) to the categorization of English pitch contours (Gili Fivela 2012). Pettorino, De Meo and associates have used prosody

transplantation in a variety of studies based on the perception of credibility in foreign-accented speech (e.g., Pettorino et al. 2012; De Meo, 2012; De Meo et al. 2011). The same group of researchers has also succeeded in applying the method as a language-learning aid (De Meo et al., 2013). An in-depth description of the architecture of the prosody transplantation method is provided in Pettorino & Vitale (2012).

The method of prosody transplantation requires at least two sentences, one produced by a native speaker and one by a non-native speaker. The number of native and non-native segments must match perfectly; it is therefore advisable to use highly controlled speech samples, such as read speech (Yoon, 2007). After a careful segmentation of the two sets, paying particular attention to the possible presence of silent pauses (Pettorino & Vitale, 2012), the transplantation of prosody can be applied using signal manipulation software, such as *Praat* (Boersma & Weenink, 2014) or *Tandem-Straight* (Kawahara, 2008). Through the application of the *PSOLA* algorithm as implemented in the software, it is then possible to automatically superimpose the duration and F_0 of one sentence (the *donor*) on the segments of the other (the *recipient*). The segments of the recipient sentence are first stretched or shrunk in order to match the duration of the donor sentence, and then the F_0 contour of the donor sentence is superimposed on the recipient segments. Selective transplants are also possible: the process can be stopped after the first step (duration transplant) and the F_0 contour can be adapted to the original duration of the recipient segments (F_0 transplant).

The main drawback of the prosody transplantation method is that the transplants are uniformly applied segment by segment, leaving the subphonemic level untouched (Yoon, 2007), as observed in Section 3.5.3 for the superimposition of duration. This could still affect the stimuli leaving artifacts, resulting in a somewhat limited naturalness.

3.6 Conclusion

The main purpose of this chapter was to outline the main issues in the study of non-native prosody, both in theory and in practice. One of the main theoretical issues in studying L2 prosody is the partial compatibility with the existing L2 acquisition models, which were specifically designed to predict and explain phonemic acquisition, rather than the acquisition of the suprasegmental aspects of L2. Although researchers have been recently attempting to frame the study of certain aspects of L2 prosody within the existing acquisition models (Mennen, 1999; Gili Fivela, 2012), the peculiar nature of suprasegmentals makes it difficult to apply traditional experimental paradigms, as they often result inadequate for the study of prosody (Vaissière, 2005). The chapter also discussed the practical dimensions of L2 prosody research, regarding the many sources of variation based on speakers, listeners, experimental procedures and speech materials. The picture that emerges from this review of theoretical and practical issues in the experimental study of L2 prosody is the need for standardized methods to limit the enormous variation that characterize prosody at many levels (Vaissière, 2005).

The final section of this chapter discussed the main resynthesis procedures adopted in the study of L2 prosody. This section is directly connected with Chapter 4, where all the methods reviewed will be evaluated in a series of pilot studies carried out by the author. Both the considerations reported in this chapter and the results of the pilot studies in Chapter 4 were functional to the development of the experimental procedures that were used in the production study (Part II) and the perception study (Part III).

Chapter 4

Italian-accented prosody in English L2: four pilot studies

4.1 Introduction

In the previous chapters, it was mentioned that the empirical studies focusing on the perception of L2 prosody are still limited, as compared to the research carried on the production and perception of L2 segments. In particular, Chapter 3 discussed the need for a suitable method for testing the single prosodic aspects (e.g., pitch and duration) and limiting the influence of segmental information in foreign accent detection tasks and accent rating tasks. Moreover, Italian-accented English has only recently started to be studied from the point of view of prosody (Busà, 2012), and the studies published so far have been focused more often on production rather than on perception (see Chapter 2).

For these reasons the author carried out a series of pilot studies, which were mainly aimed to determine the relative importance of pitch and duration in the perception of Italian accent in English. These exploratory studies were also used as a benchmark to evaluate the effectiveness of some of the signal manipulation methods presented in Chapter 3.

The first experiment (Pilot Study 1) was aimed to define a possible hierarchy between pitch and duration in the perception of Italian accent in English, presenting the listeners with stimuli where the influence of segments was neutralized by using a combination of signal manipulation methods, namely delexicalization and monotonization. The results showed that native English listeners could detect foreign accent above chance level not only when the segmental information had been degraded, but also when the pitch was reduced to a fixed value, showing the importance of temporal information (duration and speech rate) in the perception of Italian accent in English.

A second experiment (Pilot Study 2) was aimed to directly test the relative importance of pitch and duration by using another delexicalization technique meant to neutralize the effects of segmental duration. This study, investigating both Italian-accented English L2 and English-accented Italian L2, showed that both groups of native listeners were able to recognize the stimuli containing pitch and segmental duration characterizing L1 productions, while none of the other experimental conditions presented values above chance level.

In the third experiment (Pilot Study 3) the segmental information was reintroduced to exploit the listeners' fine-grained sensitivity in an accent rating task rather than adopting the forced-choice paradigm of the previous pilot studies. The method adopted in this study was prosody transplantation. The main purpose was comparing the effects of segmental information, pitch and duration on the degree of perceived foreign accent. The results of this study clearly confirmed that segmental information has the strongest effect. As for the relative importance of pitch and duration, the results did not show which cue was the most importance between duration and pitch.

The fourth experiment (Pilot Study 4) was based on the data collected for this thesis, and was aimed to test the influence of pitch span on the degree of perceived foreignness of Italian-accented productions. In this case, prosody transplantation was paired to text-to-speech (TTS) synthesis. With

this combination of methods it was possible to avoid the influence of segmental information, while at the same allowing for the manipulation of single prosodic aspects. However, the listeners' ratings were affected by the high degree of unnaturalness of the stimuli, which yielded data that were biased towards the equation 'more unnatural = more foreign'.

The following sections will briefly present each pilot study, outlining their methodology and results. The discussion of the results will focus on the effectiveness of the methods adopted and tested.

4.2 Pilot Study 1

4.2.1 Rationale and hypotheses

This pilot study (previously presented in Rognoni, 2012) was aimed to investigate the relative contribution of prosodic aspects in the perception of Italian accent in English L2 using a combination of signal manipulation techniques. In particular, read speech samples uttered by Italian speakers of English L2 were treated with monotonization and delexicalization (see Chapter 3), in order to verify if non-native speech could be recognized as such without the influence of segmental information. The following two hypotheses were formulated:

- H_1: Native English listeners can detect foreign accent when most of the segmental information is degraded, but pitch and duration have been left untouched;
- H_2: Native listeners can still detect foreign accent when segmental information is degraded and pitch patterns have been monotonized, basing their judgment on the remaining temporal aspects (i.e., duration and rhythm).

4.2.2 Methodology and experimental procedure

Speech samples were elicited from 5 Italian native speakers from the North-East Veneto area and from 5 British English native speakers from Southeastern counties of England by asking them to read a version of Aesop's fable *The Fox and the Crow* adapted by the author. Four sentences were selected from each speaker, presenting a variety of intonation patterns and syntactic structures; the resulting set of speech samples consisted in 40 utterances (4 sentences x 10 speakers). The British English speakers were all exchange students at the University of Padua.

A set of 40 delexicalized stimuli was then created by modifying the original utterances with the PURR (Prosody Unveiling through Restricted Representation) method developed by Sonntag & Portele (1998). The PURR method, originally meant for the evaluation of prosody in text-to-speech software, was chosen because of the smoothness of the resulting filtered speech, which sounded easy and not tedious to be evaluated in a perception test.

A second set of 40 utterances was generated by monotonizing the F_0 contours of the delexicalized sentences. The resulting stimuli presented degraded segmental information and a flat line replacing the F_0 contour. As a consequence, the main cues available to the listener were the temporal aspects of prosody (rhythm and speaking rate). Both techniques were applied using *Praat* scripts adapted or written by the author.

As for the experimental procedure, 10 English native speakers participated in the perception test, which was conducted using the *OpenSesame* stimuli presentation program (Mathôt et al., 2012). After a brief training session, the subjects were asked to give their responses by choosing an option in a forced-choice between 'English native speaker' and 'Italian native speaker'. The sentences were presented in two blocks corresponding to the two experimental conditions of the stimuli: delexicalized only, or delexicalized and monotonized. The order of presentation of the two conditions was randomized, as was the presentation of the stimuli within the two blocks.

Each stimulus was presented three times: as a result, the total number of tokens to be evaluated was 120 per condition. The stimuli were presented with the orthographic transcription of each sentence on screen to make the task less demanding, since the interest was not in the actual intelligibility of the sentences but in their global accentedness (see Munro et al., 2010; van Els & De Bot, 1987).

4.2.3 Results and discussion

The results of Pilot Study 1 are summarized by condition in Tab. 4.1 and visualized in Fig. 4.1.

Table 4.1: Total number of responses, mean number and standard deviation of correct responses given by the English native listeners in Pilot Study 1, presented by condition.

Condition	N	Mean	SD
Delexicalized	120	78.90	12.59
Delexicalized and monotonized	120	68.20	3.26

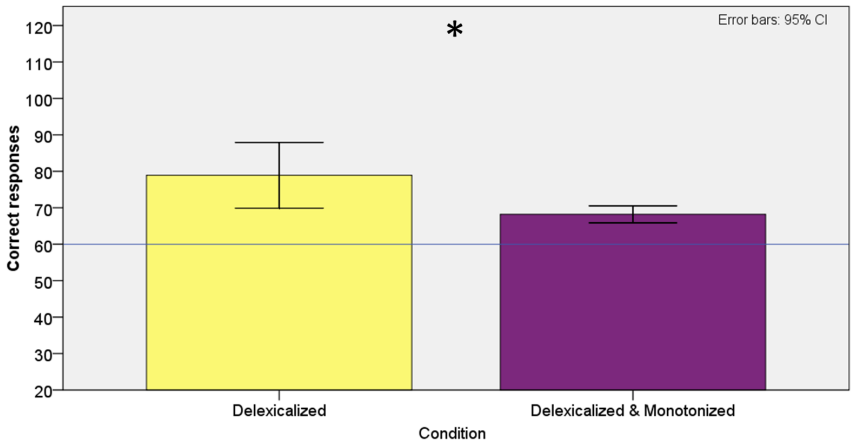

Figure 4.1: Bar chart showing the mean number of correct responses given by the English native listeners in Pilot 1, presented by condition. The asterisk indicates statistical significance.

The numbers if correct responses were well above chance level for both the delexicalized and the delexicalized and monotonized stimuli, showing that listeners were able to detect foreign accent in both conditions. A Mann-Whitney U test comparing the mean number of correct answers in the delexicalized and in the delexicalized and monotonized conditions showed that there is a significant difference between the numbers of correct answerss obtained in the two conditions (z=-2.198, p=0.03). This finding showed that pitch is a stronger cue to detect foreign accent when compared to the temporal prosodic aspects.

The results confirmed the hypotheses. The numbers of correct answers obtained when judging the delexicalized stimuli and the delexicalized and monotonized stimuli were both well above chance level, showing that prosodic cues indeed plays a crucial role in the detection of foreign accent even without intelligible segmental information.

Among the prosodic cues, pitch seems to have the greatest impact: the significant difference between the numbers of correct answers obtained when judging the delexicalized and the delexicalized and monotonized stimuli shows that the presence of discernible pitch patterns significantly improves foreign accent detection, as found out by Jilka (2000) for German-accented English. However, this pilot study tested the importance of temporal aspects only indirectly, that is, by comparing the results obtained in the two conditions, with or without the influence of pitch. Further tests on specifically modified stimuli where also duration could be directly manipulated were needed in order to have a clearer insight on the impact of duration in the detection of Italian accent.

Although the results of this study confirmed the hypotheses, a word of caution in interpreting the results is in order. Besides the limited number of subjects that were tested, one cannot completely rule out the possibility that subjects' relative familiarity with Italian could have been played a role as a facilitating factor in accent detection (Gass & Varonis, 1984).

4.3 Pilot Study 2

4.3.1 Rationale and hypotheses

The results of Pilot Study 1 showed that prosodic cues, namely pitch and duration, are both important in the detection of Italian accent in English. Pilot Study 2 was aimed to define the relative importance of pitch and duration in foreign accent detection both in English L2 and in Italian L2. Furthermore, the delexicalization method was changed in favor of a technique that could retain information on syllable structure and rhythm, and a method was designed in order to neutralize the differences in segmental duration. Hence, two perception tests were prepared, one where native English listeners were presented with Italian-accented stimuli in English L2, and one where Italian native listeners were presented with English-accented productions in Italian L2. The hypotheses to be tested were the following:

- H_1: Both groups of listeners can detect foreign accent when the segmental information is reduced, but pitch and duration are left untouched;
- H_2: Both groups of listeners can still detect foreign accent when segmental information is reduced and pitch patterns are monotonized;
- H_3: Both groups of listeners will also be able to detect foreign accent when duration is neutralized.

4.3.2 Methodology and procedure

This study was again based on read speech. The samples in English partially corresponded to the ones used in Pilot Study 1; they consisted of sentences extracted from the recording of a fable read by 4 Italian native speakers from the North-East Veneto area and 4 British English native speakers. For each speaker, four sentences were selected; the resulting set of productions consisted in 32 utterances (4 sentences x 8 speakers). As for the Italian data

set, similar speech samples were elicited from 4 Italian native speakers and 4 British speakers, based on the reading of a translation of the same passage in Italian. The sentences selected for each speaker were again 4, resulting in 32 utterances (4 sentences x 8 speakers).

For each language group, a set of 32 *SASASA* files (Ramus & Mehler, 1999) was created. These are 'sound files in which an [s] sound replaces all consonantal intervals of the original file, whereas an [a] sound replaces all vocalic intervals of the original file' (Mairano, 2011: 91). The resulting sounds are chains of [s] and [a] segments, which still maintain the original prosodic aspects (pitch, duration and intensity), thus reminding stimuli produced with the reiterant speech (RS) paradigm (Tajima et al, 1996; Ueyama, 2012). The main difference between *SASASA* and RS is that *SASASA* files are resynthesized with a computer program (in the case of this study, *Praat*), while for reiterant speech speakers are specifically instructed to produce utterances where "every syllable of a phrase is replaced with a standard syllable such as [ma], but most of the rhythmic and melodic features of the phrase are maintained" (Tajima et al., 1996: 2493). Since one of the main aims of this study was to collect evidence in respect to the relative importance of segmental duration, *SASASA* seemed the right choice.

The *SASASA* files were then further manipulated by monotonizing the F_0 contours of the delexicalized sentences, similarly to Pilot Study 1. As a result two sets of 32 so-called flat *SASASA* stimuli (Ramus & Mehler, 1999) were generated, one for each language data set.

The final step of the stimuli preparation involved a procedure that could neutralize the effects of duration in a way similar to what monotonization and delexicalization did in neutralizing the effects of pitch and segmental information, respectively. Since such a technique was not readily available (see Section 3.5.3), the author created a method based on a *Praat* script. The script would replace the duration of the vowels with a fixed value represented by the average value of vowels in English and Italian, based on the literature

(230 ms for British English, based on Wells, 1962; 320 ms for Italian, based on Giordano, 2006). As a result, the vowels resulted stretched or compressed to match the fixed value, neutralizing any difference between stressed and unstressed, or full and reduced, vowels. The application of this technique to natural speech would result in highly artificial stimuli, but the unnaturalness was counterbalanced by the use of the chains of synthetic *SASASA* phones as a segmental base. Being synthesized ad hoc, the duration of the single vowel segments [a] could be set without causing any distortions or artifacts in the final stimuli. As a result, two more sets of 32 sentences were generated, one for each language data set.

The six experimental conditions are summarized in Tab. 4.2, listed by their coding name and accompanied by a summary of the status of duration and F_0, which could be native, non-native or neutralized. The number of stimuli for each condition is also provided.

Table 4.2: The six experimental conditions of Pilot Study 2, with the number of stimuli for each condition.

Condition	**Duration**	**F_0**	**Number of stimuli**
all_NS	native	native	16
all_NNS	non-native	non-native	16
flat_NS	native	monotonized	16
flat_NNS	non-native	monotonized	16
timefixed_NS	neutralized	native	16
timefixed_NNS	neutralized	non-native	16

As for the experimental procedure, 10 British English native listeners and 11 Italian native listeners took the perception tests based on English and Italian, respectively. Both tests were conducted using the *LimeSurvey* survey presentation software (Schmitz, 2012). The task was similar to the one in Pilot Study 1: after a brief training session, the subjects were presented with the stimuli one by one and they were asked to judge them by choosing one of the two options in the forced-choice between 'Native speaker'

and 'Non-native speaker'. The stimuli were pooled in the same block and presented in randomized order to each listener. The number of tokens to be evaluated was 16 per condition, resulting in a total of 96 tokens. As in Pilot Study 1, the stimuli were presented along with the corresponding orthographic transcription.

4.3.3 Results and discussion

The results of Pilot Study 2 are summarized in Tab. 4.3 and Fig. 4.2, showing the mean and standard deviation for the six experimental conditions, again listed by their coding name and accompanied by a summary of the status of the two acoustic cues analyzed (duration and F_0), which can be native, non-native or neutralized.

Table 4.3: Total number of responses, mean number and standard deviation of correct responses given by English native listeners and Italian native listeners in the respective perception tests, presented by experimental condition

Condition	**English listeners**			**Italian listeners**		
	N	**Mean**	**SD**	**N**	**Mean**	**SD**
all_NS	16	12	2.31	16	12.18	2.96
all_NNS	16	6.90	3.84	16	7.09	2.47
flat_NS	16	4.60	5.21	16	7.09	4.89
flat_NNS	16	10.20	5.25	16	10	4.90
timefixed_NS	16	8.20	3.85	16	6.82	3.87
timefixed_NNS	16	10.60	3.92	16	10.73	3.85

Fig. 4.2 shows that the mean number of correct responses given by the English listeners when evaluating English productions were significantly above chance level only for the 'all_NS' condition. This was confirmed by the results of a One-Sample t-test against chance (=8): t(N=10, M=120)=5.477, $p<0.01$. For all other conditions, the difference against chance level was not significant.

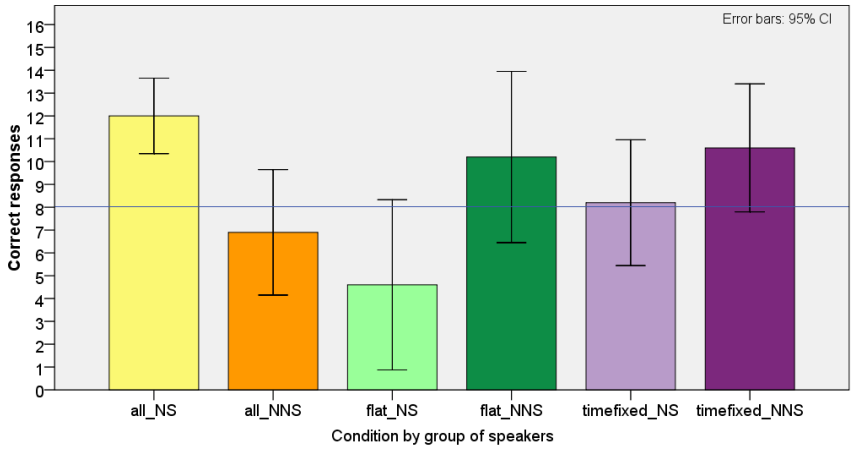

Figure 4.2: Mean number of correct responses given by English native listeners in the perception test based on Italian-accented English productions, presented by experimental condition.

Fig. 4.3 shows that the results observed for the Italian listeners were very similar. In particular, the correct answers given by the Italian listeners when judging Italian productions were significantly above chance level only for the 'all_NS' and the 'timefixed_NNS' conditions.

The statistical significance of the differences was confirmed by the results of a One-Sample t-test against chance: t(N=11, M=12.18)=4.685, p=0.01 ('all_NS) and t(N=11, M=10.73)=2.350, p=0.04 ('timefixed_NNS'). As for the other conditions, the difference against chance level was not significant.

The results of both perception tests shows that the only condition where the listeners could successfully identify the stimuli was when the stimuli presented native values of F_0 and duration. In all the other cases the mean values were never significantly above chance level. The fact that this trend was virtually the same for both groups casted doubts on the validity of the experimental setup and was useful to better understand the risk and the consequences of heavy signal manipulation. In particular, the analysis of the results showed that in the 'flat' and 'timefixed' conditions there is a bias in the listeners' judgment towards foreignness: it seems that the odder a stim-

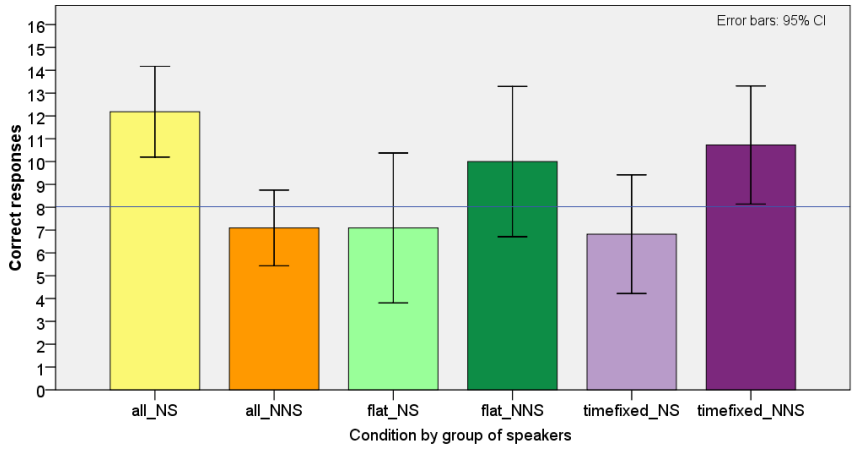

Figure 4.3: Mean number of correct answers given by Italian native listeners in the perception test based on English-accented Italian productions, presented by experimental condition.

ulus sounds, the more it is likely to be considered foreign. This equivalence between odd and foreign seems to be a byproduct of the application of manipulation techniques that resulted particularly invasive, resulting in highly unnatural stimuli. Considering this effect, the statistical significance of the 'timefixed_NNS' condition observed in the results of the Italian perception test must be seen as an artifact originated by the mentioned bias, rather than as an effective preference for the non-native productions with neutralized duration. This bias effect caused by heavy signal manipulation was also observed in the results of Pilot Study 4 (see Section 4.5.3).

4.4 Pilot Study 3

4.4.1 Rationale and hypotheses

This pilot study, previously published in Rognoni & Busà (in press), was designed to investigate the relative importance of segmental and suprasegmental cues in the perception of Italian accent in English, and to determine

whether it is duration or pitch that is a more important prosodic cue in this perception process. In this case, the manipulation method adopted was prosody transplantation (see Chapter 3). This solution allowed for the selective manipulation of duration and pitch, while at the same time maintaining the segmental information intact. Moreover, with prosody transplantation it was possible to present the listeners with a fine-grained accent-rating task, rather than with a forced-choice task limited to two options. The experiment was set up to test the following two hypotheses:

- H_1. Segmental information is the strongest cue for foreign accent perception;
- H_2. Segmental duration is a stronger cue as compared to pitch.

4.4.2 Methodology and procedure

All sentences were first manually segmented and annotated using *Praat*. The same program was used to transplant prosody on the segments running the 'prosody cloning' script written by Yoon (2007, see Section 3.5.4 for an extensive explanation of method). Native and non-native duration and F_0 values were transplanted both together and selectively, resulting in 8 different experimental conditions, summarized in Tab. 4.4.

21 native British English listeners participated in the perception test; all of them claimed to have no knowledge or familiarity with Italian. The stimuli were presented to the listeners using the survey presentation platform *LimeSurvey* (Schmitz, 2012). The listeners were asked to listen to the stimuli at their own pace, and to rate them using the full length of a slider scale, where they could rate both the degree of foreign accent in a continuum from *no foreign accent* to *very heavy foreign accent*, and the native vs. non-native status of the speakers (Fig. 4.4).

The values in the sliding scale ranged from 0 to 100, but they were not visible to the listeners, who were asked to move the handle of the slider from

Table 4.4: Summary of the eight experimental conditions generated with prosody transplantation for Pilot Study 3.

Condition	Segments	Duration	F_0	Number of stimuli
1	native	native	native	16
2	native	non-native	non-native	16
3	native	native	non-native	16
4	native	non-native	native	16
5	non-native	native	non-native	16
6	non-native	non-native	native	16
7	non-native	native	native	16
8	non-native	non-native	non-native	16

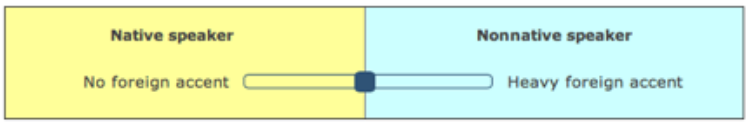

Figure 4.4: Sliding scale used by the English native listeners in the perception test to rate foreign accent.

the default central position (50) towards one of the two extremes of the scale as a function of the degree of perceived foreignness. All 128 stimuli were played to each listener in a single block in randomized order. The overall running time of the experiment was approximately 20 minutes.

4.4.3 Results and discussion

The results of the statistical analysis are visually summarized in Tab. 4.5.

In addition, Fig. 4.5 shows that the greatest difference in accentedness is between native and non-native segments. The hierarchy of the suprasegmentals is the same for native and non-native segments, suggesting that segmental duration has a slightly higher effect than F_0 on accentedness.

Accentedness was analyzed by a repeated measure Analysis of Variance (RM-ANOVA) with condition (8 levels) as within-subjects factor.

Table 4.5: Summary of the eight experimental conditions generated with prosody transplantation for Pilot Study 3.

Condition	N	Mean	SD
1	16	62.26	12.78
2	16	41.47	11.05
3	16	70.99	10.85
4	16	25.94	8.77
5	16	72.13	9.50
6	16	20.99	11.19
7	16	78.67	9.00
8	16	15.53	10.43

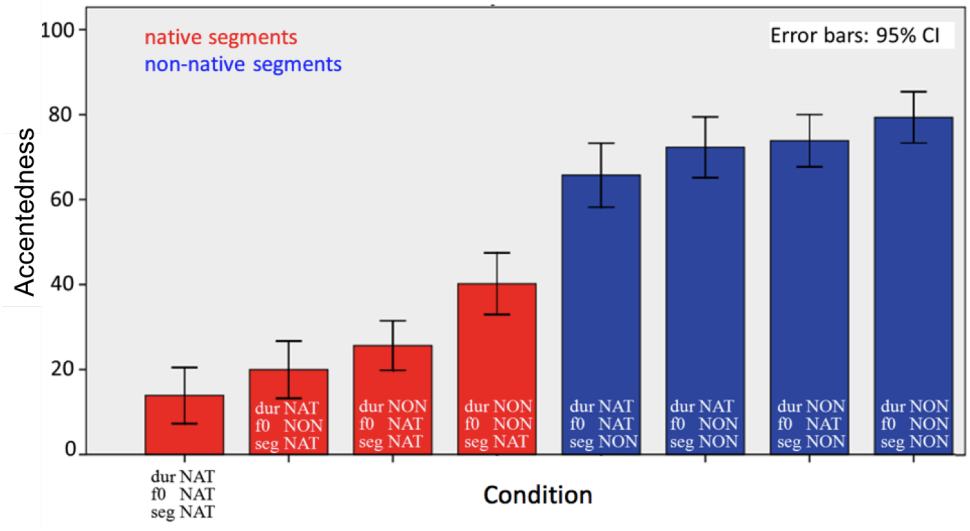

Figure 4.5: Bar chart showing accentedness (0-100) by condition in Pilot Study 3, where 0 corresponds to *no foreign accent* and 100 to *heavy foreign accent* (from Rognoni & Busà, in press).

The RM-ANOVA showed a significant effect for condition on accentedness ($F(1,20)=203.62$, $p<0.01$). Pairwise comparisons (with Bonferroni adjustment) between the eight different conditions showed significant differences in all cases except the ones between transplanted duration and transplanted pitch, both on native and non-native segments.

To sum up, the results of the perception test show that segments have the greatest effect in foreign accent rating, confirming the first hypothesis tested in this study, that is, that segments provide the strongest cue for accent perception. The second hypothesis, that segmental duration is a stronger cue in accent rating as compared to pitch, was not confirmed by the experimental data: the results showed a tendency for segmental duration to be a stronger cue, the difference in accentedness between a stimuli with selective transplant of duration and stimuli with selective transplant of pitch was not statistically significant. This was probably due to the intrinsic limits of the prosody transplantation method, through which duration can only be manipulated by stretching or shrinking the borders of the segments, without touching the subphonemic level and the spectral structure of the phones (see Chapter 3). Differences in duration between Italian and English are connected with the phenomenon of vowel reduction (see Busà, 1995), which affects both the temporal and the spectral levels. The lack of differentiation in the formant structure of vowels has probably limited the listeners' sensitivity to vowel duration as a relevant phonetic cue to foreignness.

To conclude, the prosody transplantation paradigm proved to be a suitable methodological tool to test the relative effects of segmental and suprasegmental information in accent rating, confirming that segmental information has a stronger effect on the perception of foreignness. However, prosody transplantation did not provide definite answers to the question involving the relative importance of pitch and duration in accent detection. The experiment did show that they are both important enough to change significantly the perception of foreignness when compared to all-native or all-non-native stimuli, encouraging the author to further test the influence of the two prosodic cues in further experimental studies.

4.5 Pilot Study 4

4.5.1 Rationale and hypotheses

This fourth and last pilot study was based on the speech material collected for this thesis and on the results of the production study, suggesting that the productions of non-native speakers of English present a significantly wider pitch span as compared to native productions (see Section 6.2.3). The main research question driving this pilot study was to determine whether differences in pitch span could be enough to betray foreign accent. In particular, the listeners were asked to perform a double task: an accent detection task and an accent rating task. The hypotheses that were formulated were the following:

- H_1: English native listeners can distinguish between native and non-native productions only by listening to a correct or incorrect implementation of pitch span;
- H_2: English native listeners will perceive a higher degree of foreign accent when sentences present non-native pitch span values as compared to the ones where pitch span is characterized by the native values.

4.5.2 Methodology and procedure

The synthetic stimuli created for this experiment were based on a subset of the sentences analyzed in the production study (see Section 5.2.1). The productions of two groups were considered: English native speakers (NS) and non-native speakers with a high competence in English L2 (NNS1). The resulting number of stimuli was 80 (40 sentences x 2 groups).

In order to test these hypotheses it was necessary to adopt speech resynthesis techniques that could disentangle pitch from the influence of duration on the one side, and segmental information on the other (see Chapter 3).

Even the productions by NNS1 presented an easily recognizable foreign accent and this required a technique that could reduce the influence of segmental errors in the judgment of non-native productions. The manipulation method used to overcome these issues consisted in a combination of speech synthesis and prosody transplantation.

The first step was to use a text-to-speech (TTS) program to generate a set of synthetic sentences. The software used was the *Mary* (Modular Architecture for Research on speech sYnthesis) *TTS* system, developed by the DFKI institute (Schröder & Trouvain, 2003). The orthographic transcriptions of the sentences required were inserted in the interface of *Mary TTS*, and 80 audio files in .wav format were generated, consisting of the sentences of the two groups (NS and NNS1) pronounced by two synthetic voices based on SSBE pronunciation, *Poppy* (female) and *Spike* (male).

The second step was to apply prosody transplantation. This method was used to extract F_0 values from the productions of NS and NNS1. The F_0 values were then time-aligned, and superimposed onto the synthetic utterances previously generated with *Mary TTS*. These operations were all performed by running the 'prosody cloning' *Praat* script written by Yoon (2007), already used in Pilot Study 3. As a result, 80 stimuli were created. These were divided in two sets:

- 40 sentences with synthetic British English segments and duration, with pitch values transplanted from the productions by NS;
- 40 sentences with synthetic British English segments and duration, with pitch values transplanted from the productions by NNS1.

In both groups the speakers' genders were matched with the gender of the synthetic voice. In order to control for memory effects, a series of 20 distractors was also included. The distractors consisted of an extra set of sentences generated using *Mary TTS*, uttered by the same two voices, but

with a completely different content as compared to the one of the target sentences.

The subjects participating in the experiment were 12 British English native speakers. Again, the stimuli were presented to the listeners by using the *LimeSurvey* platform (Schmitz, 2012). Experience with perception tests based on heavily manipulated synthesized stimuli (see Pilot Studies 1 and 2) led the author to create an experimental task with a motivating presentation, in order to limit the tediousness and disorientation which had often been pointed out by participants in similar experimental tasks. Therefore, it was decided to present the task as a role playing game. The instruction page told the subjects that they were going to listen to utterances produced by robots (i.e., the synthetic voices) that were programmed to speak with a British English (i.e., SSBE) pronunciation. However, a hacker had modified their productions by transplanting non-native intonation (i.e., pitch) into the robots' productions. The task was then presented as an attempt to discover if the utterances were produced with native or non-native intonation in order to restore the robots to normality. At the end of the task, the participants were presented with their results so that they would know if they had succeeded or not in restoring the order.

Since the judgment required from the listeners was based on the immediate and global impression they could get from listening to each stimulus, the subjects were invited to listen to each stimulus only once before giving their responses. The listeners were asked to respond to the stimuli by performing two different actions. The first was judging if the intonation of the utterance was native or non-native by clicking on the appropriate option in binary forced choice (native vs. non-native speaker). The second was to rate the degree of foreign accent (if any) that they had perceived in the utterance. Rating was possible by using the full length of a 7-point Likert scale, where 1 was labeled 'no foreign accent' and 7 'very heavy foreign accent'.

The 80 experimental stimuli were pooled together in a single block and

Table 4.6: Total number of stimuli, mean and standard deviation of the correct responses given by English native listeners in the accent-detection and accent-rating tasks of Pilot Study 4.

Condition	Accent detection			Accent rating		
	N	Mean	SD	N	Mean	SD
Native	40	16.92	6.20	40	2.43	0.71
Non-native	40	24.54	8.12	40	2.63	0.87

presented in a different randomized order for each participant. The experiment was preceded by a short training session, where the subjects could familiarize with the manipulated stimuli and with the interface. The average running time of the experiment was approximately 20 minutes.

4.5.3 Results and discussion

The results of the statistical analysis are visually summarized in Tab. 4.6. Figures 4.6 and 4.7 show the results of the accent detection and accent rating tasks, respectively.

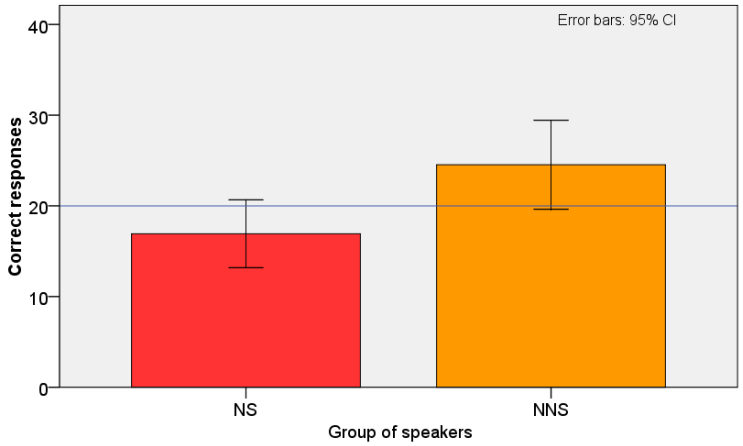

Figure 4.6: Bar chart showing the mean number of correct responses given by English native listeners in the accent detection task of Pilot Study 4, presented by group of speakers.

Fig. 4.6 shows that the results of the accent detection test did not reach significance above chance level for either group of speakers. Moreover, there was no statistical significance between the numbers of correct responses obtained when judging stimuli with native or non-native pitch span values.

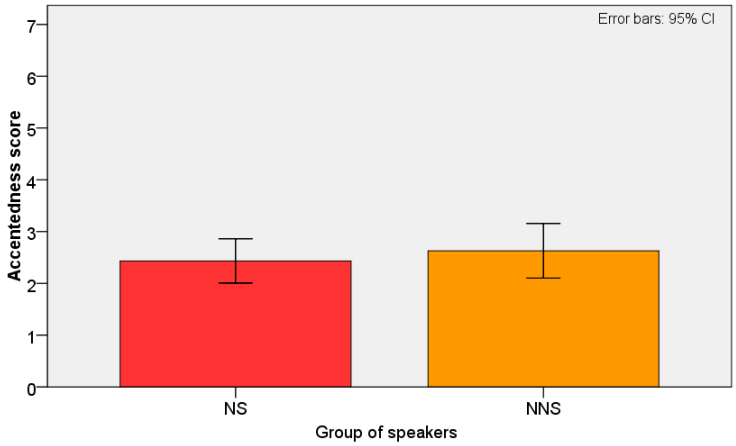

Figure 4.7: Bar chart showing the mean number of correct responses given by English native listeners in the accent rating task of Pilot Study 4, presented by group of speakers.

As for the results of the accent-rating task, Fig. 4.7 shows that there was no sizable difference between the results obtained when rating stimuli presenting native pitch span and the ones presenting non-native pitch span.

In general, the results of Pilot Study 4 did not confirm the hypotheses that native listeners could identify native and non-native speakers on the basis of pitch span alone. As a consequence, the research question regarding the importance of an incorrect implementation of pitch span in the detection and rating of foreign accent remained unanswered.

However, these results must be considered with a grain of salt. It is very likely that the manipulation method adopted in the experiment was one of the main causes of its inconclusive results. This impression was corroborated by the feedback given after the experiment by several participants, who commented on the difficulty of the task. Furthermore, it seems that the

combination of methods used in this pilot study yielded the same kind of bias found in Pilot Study 2. The sentences generated with speech synthesis were supposed to neutralize differences in pronunciation between native and non-native productions to allow listeners to focus on differences in the realization of suprasegmental features. However, the results showed that this solution ended up hindering the listener's sensitivity to foreign accent rather than facilitating it.

Positive comments reported by the participants regarded the motivational aspects and the framework of the experiment. The fact that participants enjoyed this setting shows that the aim of creating a less tedious experience and to arouse interest and to keep up the participants' attention was achieved. This could be interesting in the view of applications of similar experimental tasks to L2 language instruction. While the task will probably result as demanding as it was for the participants in the experiment and needs to be modified, the motivating setting could be maintained and implemented in similar computer-based activities to improve awareness and pronunciation on English prosody.

4.5.4 Conclusion

The results of the four pilot studies were useful to collect empirical evidence on the general perception of the prosody of Italian-accented English L2, and they provided empirical evidence that was used to formulate the research questions and hypotheses to be tested in this thesis.

The results of Pilot Study 3 seem to confirm the overriding importance of segmental information in accent perception and rating tasks when compared to prosodic information. As a result of this strong effect of segmental information on FA perception, the same study did not achieve conclusive results as for the relative importance of segmental duration vs. pitch. The results obtained in Pilot Study 1 suggested that pitch has a stronger effect as compared to temporal aspects. However, the combination between delex-

icalization and monotonization used in the experiment made it impossible to specifically test the influence of pitch vs. the single temporal aspects, such as speaking rate and overall duration. Pilot Study 2 and 4 did not achieve conclusive results, mainly because of the high level of unnaturalness of the stimuli, which resulted very difficult to be judged by the listeners.

Besides collecting first-hand data on the perception of the prosody of Italian-accented English, these pilot studies were also necessary to choose suitable methods to use in the perception study created for the present work. Since the manipulation techniques heavily influenced the results at least two cases (Pilot Studies 2 and 4), it was decided to base one perception test on natural stimuli (see Chapter 7) and the other on slightly manipulated stimuli (see Chapter 8).

Part II

Production Study

Chapter 5

Methods

5.1 Rationale and hypotheses

Chapter 2 has shown that English and Italian present different strategies for focus marking. In English focus is marked prosodically, that is, by sizable changes in pitch, duration and intensity. Since word order is relatively fixed (e.g., SVO structures for declarative sentences), prosodic cues are used to convey emphasis on the pieces of information that are particularly relevant in discourse. Previous studies have shown that the phonetic realization of narrow focus is conveyed by a combination of higher F_0 and longer duration on the focused constituent when compared to the rest of the utterance (Eady et al., 1985; Xu & Xu, 2005; Breen et al., 2010). In particular, it has been suggested that "[t]he main correlate of perceived prominence in English is [...] a local maximum or minimum of the fundamental frequency" (Büring, 2007: 447).

In Italian, instead, emphasis is more often achieved with the dislocation of the information in focus to marked positions in the right periphery of the sentence, thanks to the freer word order allowed by the Italian grammar. As a result, the use of prosodic cues in focus marking becomes redundant, and it is normally reserved to cases where extra emphasis is needed, for example

when contrasting or correcting information that has been previously given in the context of conversation.

When considering the differences in the phonetic realization of narrow focus in the two languages, it can be hypothesized that the progressive tuning towards the target language by Italian speakers with a higher competence in L2 will involve the activation of the phonetic cues that are used by native speakers to mark focus, especially F_0. In contrast, the speakers with a lower L2 competence will still rely heavily on L1 strategies, confirming that the impact of prosodic transfer from L1 to L2 is higher for less competent non-native speakers (Ueyama, 2012).

This production experiment was designed to test the following hypotheses:

- H_1: Native British English speakers (NS) can mark narrow non-contrastive focus by prosody, in particular by modulating pitch;
- H_2: Italian speakers with a high competence of L2 (NNS1) can activate pitch modulation as a focus marking strategy, at least to a certain extent;
- H_3: Italian speakers with a low competence in L2 (NNS2) fail to activate pitch modulation and present undifferentiated productions for focus marking;
- H_4: When speaking their L2, Italians do not mark narrow focus marking by prosodic means.

The hypotheses will be tested by analyzing speech samples in English L1 and L2 using the methods described in the following sections of this chapter. The results of the acoustic and the statistical analyses will be presented and discussed in Chapter 6.

5.2 Methodology

5.2.1 Speakers

Three groups of speakers were recorded: two groups of Italian speakers of English L2, divided on the basis of their competence level in English L2 (see section 2.1.2.1) and consisting of 4 speakers each, and a control group of 4 English native speakers. Before the recordings, all speakers were asked to fill in a consent form and to complete a brief questionnaire to collect information regarding their geographical origin, age, profession and languages spoken. The Italian speakers were also asked to tell at what age they had started learning English and to specify whether they had spent more than six months in an English-speaking country. The models of the consent forms and questionnaires that were submitted to both groups are reported in Appendix A.

5.2.1.1 Native speakers (NS)

The 4 English native speakers (NS) were undergraduate students and staff at the Division of Psychology and Language Sciences at the University College of London (UCL). They were all original from Southern counties of the United Kingdom, and they were all speakers of the Southern Standard British English (SSBE) variety. Two speakers were female, and two were male. At the time of the recordings, the average age of the speakers was 32.7.

5.2.1.2 Non-native speakers

The non-native speakers were undergraduate and graduate students enrolled at the University of Padua. They were born and living in Italy, and were all original from the Veneto region, in the North-East area of the country. At the time of the recordings, the average age of the Italian speakers was 24.4. . All the Italian speakers confirmed that they had begun to learn English at

school at the age of 11. Initially 12 non-native speakers were recorded. From this group 8 speakers were selected and assigned to two different groups, consisting of 4 speakers each and based on the level of their competence in English. In this study, particular attention was paid to the criteria used to assign the Italian speakers to two homogeneous groups. The definition of the two groups is therefore presented in detail in the next subsection.

5.2.1.3 Definition of groups based on L2 competence

In order to select the speakers and to objectively assign them to two groups based on their L2 competence, two methods were used: a vocabulary size test performed by the speakers and a perception test based on the judgments given by a panel of English native listeners.

It is known that the results of lexical tests can offer an effective way to quickly diagnose the general competence in a language (see Darcy et al, 2013). It was therefore decided to include a vocabulary size test in order to define the participants' level of competence in English L2. The chosen test was the "Vocabulary Size Placement Test' included in the *Dialang* project (Council or Europe, 2001: 226-230). This test was chosen for the balance between brief duration and diagnostic power and for the quick readability of the final scores. In this test the participants are presented with a total of 75 words, some of which are real and some are nonsense; the task is to identify the real words (e.g., *to settle*) and the nonce words (e.g., *to markle*). The score attributed by the test ranges from 1 to 1000, and it is distributed in six ranges corresponding to the six levels of the *Common European Framework of Reference for Languages* (CEFR) (Council of Europe, 2001). From the six corresponding descriptors the participants can have an immediate idea of their lexical competence (see Tab. 1).

The results of lexical tests can be seen as a reliable diagnostic tool for the overall competence in L2, but for this study it was necessary to assess the productions from the point of view of pronunciation. For this purpose,

Table 5.1: The six ranges of the *Dialang* 'Vocabulary Size Placement Test', with the corresponding CEFR levels and descriptors (from Council of Europe, 2001: 226-230).

Range	**CEFR level**	**Descriptor**
0-100	A1	This level indicates a person who knows a few words, but lacks any systematic knowledge of the basic vocabulary of the language.
101-200	A2	This level indicates a very basic knowledge of the language, probably good enough for tourist purposes or "getting by", but not for managing easily in many situations.
201-400	B1	People who score at this level have a limited vocabulary which may be sufficient for ordinary day-to-day purposes, but probably doesn't extend to more specialist knowledge of the language.
401-600	B2	People who score at this level typically have a good basic vocabulary, but may have difficulty handling material that is intended for native speakers.
601-900	C1	People who score at this level are typically advanced learners, with a very substantial vocabulary. Learners at this level are usually fully functional, and have little difficulty with reading, though they may be less good at listening.
901-1000	C2	A very high score, typical of a native speaker, or a person with near-native proficiency.

a set of 24 sentences (2 sentences for 12 speakers) was presented to a panel of native listeners in a brief perception test, where 20 native speakers of British English were asked to judge the degree of global foreign accent of the non-native speakers' productions.

The test was presented using the *LimeSurvey* platform (Schmitz, 2012): the participants were asked to listen to the sentences at their own pace, and to rate them using the full length of a 9-point Likert scale, where they

could globally rate the degree of foreign accent by moving a handle along a continuum ranging from *no foreign accent* to *very heavy foreign accent*. All 24 sentences were played to each listener in a single block in randomized order. At the moment of taking the test none of the participants declared to know Italian nor was living or had lived in Italy. The running time of this brief evaluation session was approximately 2 minutes.

The foreignness score for each speaker was calculated by considering the mean value of the evaluations given by the native listeners for each speaker. Inter-rater agreement was also calculated, showing that the 20 raters were very consistent in their judgments (Cronbach $\alpha = .96$). A Pearson product-moment correlation coefficient was computed to assess the consistency between the vocabulary size test scores and the accent-rating test scores. There was a positive correlation between the two variables ($r = -0.922$, $n = 8$, $p = 0.001$).

Based on the results of two tests, the non-native speakers were divided in two groups, according to their level of competence in English L2:

1. one group of 4 non-native speakers with a higher competence in English;

2. one group of 4 non-native speakers with a lower competence in English.

Throughout this dissertation, the two groups will be respectively referred to as NNS1 and NNS2 respectively. Four female speakers composed the NNS1 group, while the NNS2 group was composed by two females and two males. The four Italian speakers who had obtained intermediate scores in both tests were excluded from the production study.

The background information collected in the questionnaire, the scores achieved by each speaker of the two groups in the vocabulary size test and the average ratings assigned by native listeners are summarized in Tab. 5.2.

The speakers GD and EP of group NNS1 were the only ones who had lived more than one year in English speaking countries (in both cases, Great Britain and Ireland).

Table 5.2: Background information and scores of NNS1 and NNS2. The speakers are referred to with the initials of their names.

Speaker	**Age**	**Gender**	**Foreign languages spoken**	**Score in Dialang test (0-1000)**	**Mean score in accent-rating test (1-9)**
			NNS1		
GD	29	female	English Portuguese, Spanish	1000	2.9
EP	30	female	English, Spanish	1000	3.6
EM	21	female	English, German	829	3.7
MG	24	female	English, Russian, German	805	5.25
			NNS1		
FV	22	male	English, Portuguese, French, German	143	6.7
SZ	23	male	English	403	6.8
FZ	21	female	English	102	7.1
CC	25	female	English	266	8

As for the control group of English native speakers, the background information obtained in the questionnaire is provided in Tab. 5.3.

5.3 Speech material

The speech material was designed to present clear instances of narrow focus marking. It consists of a set of short declarative sentences with fixed syntactic

Table 5.3: Background information and scores of NS. The speakers are referred to with the initials of their names.

Speaker	Age	Gender	Foreign languages spoken
FM	27	female	None
MW	36	female	None
NN	25	male	Spanish
SN	43	male	French

structure and number of syllable (7), in the following form:

1	**2**		**3**	**4**
subject	verb	"with the"	attribute	complement.

The four numbered words are referred to as *keywords* (see Xu & Xu, 2005; Breen et al., 2010); they are the words that were initially designed to test the phonetic realization of narrow focus. For each of the four keywords, five sentences were produced by each speaker, resulting in a corpus that was initially composed of 240 tokens (5 sentences x 4 keywords x 12 speakers = 240 sentences).

The sentences consisted of a fixed string of words, where only the keyword was changed, while the rest of the sentence remained unaltered. The whole set of sentences, divided in four blocks corresponding to the keywords is presented in Appendix B, along with the prompt questions used in the elicitation protocol.

5.3.1 Elicitation protocol

An original elicitation protocol was designed based on a combination of written and visual prompts. This procedure was designed in order to obtain an ecologically valid balance between controlled productions and samples that were more spontaneous than read speech. The speakers were presented with

a series of *PowerPoint* slides, where each slide corresponded to one target sentence. Each slide presented three prompts (see Fig. 5.1 for an example):

1. a written question on the top of the slide, consisting of a wh-question, designed to trigger the location of narrow focus on a specific keyword;
2. a visual representation in the central part of the slide showing a visual representation of the keyword;
3. a written prompt at the bottom of the slide, reproducing the target sentence with a gap where the keyword was expected.

The subjects' task consisted of uttering one sentence for each slide by using the information provided in the written and visual prompts.

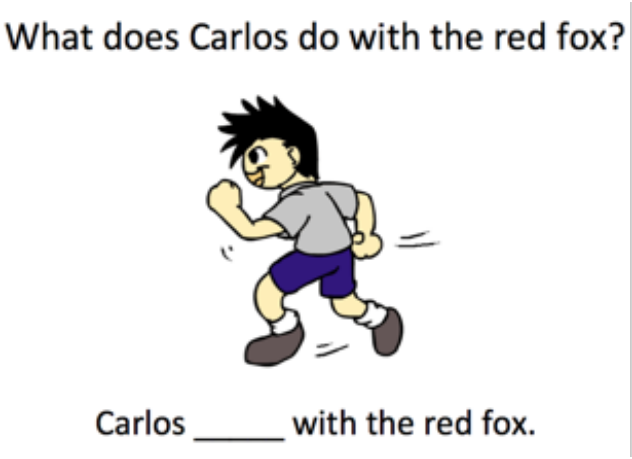

Figure 5.1: Example of one of the *Powerpoint* slides presented to the speakers to elicit narrowly focused sentences. In this case, the speaker is expected to mark a narrow focus on the verb *runs*, which corresponds to the picture and to the wh-word in the question.

The recording session was preceded by a short training phase. After being presented with the instructions on how to perform the task, the participants had the chance to familiarize with the picture on screen. In this phase, the

author went through the illustration with each participant by naming the pictures one by one, so that the participants would know how to name the keywords without doubts or hesitations. The subjects were then asked to practice with the aid of small set of images, which were not included in the study. Once the speakers were familiar with the task and ready to begin, they could start the actual recording session.

Speakers were instructed to repeat each sentence once. However, they were invited to repeat the sentences in case of any disfluencies or hesitations. They could move forward the presentation of the slides at their own pace. The order of the slides was randomized.

The non-native speakers were also asked to repeat the same task with a similar set of sentences in Italian, resulting in an extra set of 20 sentences per speaker (20x8=160), in the following form:

1	**2**		**3**	**4**
subject	verb	"con il/la'	attribute	complement.

In this set of sentences, the syntactic structure and the number of syllables (9) were controlled in the same way as they were for the English set. This second set of sentences was recorded to check for prosodic transfer effects from Italian L1. The transcriptions of the full Italian data set can be found in Appendix B.

All Italian speakers were recorded using a *Shure* SM58 microphone connected to a *TASCAM* DR-05 digital audio recorder, in a silent room at the Language and Communication Lab (LCL) at the University of Padua. The frequency rate was 48 kHz (16-bit). The English native speakers were recorded with the same equipment and the same frequency rate in a sound-treated booth at the University College of London (UCL), Division of Psychology and Language Sciences.

5.3.2 Acoustic analysis

5.3.2.1 Segmentation and annotation

After a first screening, it was decided to study only the productions with focus on sentence subjects and verbs, which will be hence referred as S and V, respectively. The reason for this choice was that the keywords corresponding to the constituents of the prepositional phrases (e.g., "with the green frog") presented a sizably longer duration, lower intensity and lower F_0 values for all groups of speakers. These values were not determined by the focus condition of the keywords, but they were rather the result of the combined action of the physiological phenomena of final lengthening and declination (t'Hart & Collier, 1990; Grice & Baumann, 2007). The impossibility of directly comparing such values with the ones of the other constituents in focus led to the decision of excluding the analysis of the last two keywords (i.e., attribute and complement). The analysis was therefore limited to the first two keywords, namely S and V, resulting in a subset of 120 tokens. However, the presence of the final prepositional phrase still played an important role in controlling for possible final lengthening of the verbs at the end of an intonational phrase, as noted by Breen et al. (2010), who included similar prepositional phrases in their target sentences to avoid final lengthening effects.

The 120 sentences were then segmented and labeled using *Praat* (Boersma & Weenink, 2014). The transcription procedure was semi-automatic: a first phonetic annotation was generated using the automatic tool *SPPAS* (Bigi & Hirst, 2012), then the author manually reviewed the transcriptions one by one. This manual check was performed in order to guarantee a fine-graded alignment between the boundaries in the annotation tiers and the events shown in the oscillogram and spectrogram views provided in the *Praat* Editor Window. The resulting data set was a total of 120 couplets of audio and *TextGrid* files. The latter were organized in five different annotation tiers, which were used to obtain a variety of acoustic

values for every marked interval. The intervals contained in the five tiers included the following information:

1. whole sentence;
2. single words;
3. syllables;
4. phonetic transcription (following the I.P.A. conventions);
5. focused and non-focused material (pre- and post-focus).

5.3.2.2 Acoustic measurements and data processing

Following the example of previous studies on focus marking (Eady et al., 1985; Cooper et al., 1986; Xu & Xu, 2005; Breen et al., 2010), it was decided to use words as the main units of reference to measure the acoustic correlates of focus. In addition, the acoustic measurements were also run over sentences. While the measurements at sentence level were useful for the comparison between groups, the values of words were used for a more detailed within-group analysis. The acoustic measures that were applied are listed with a brief description in Tab. 5.3.

The measurement called *normalized* F_0 was calculated in order to determine the local values of F_0 in correspondence with the selected intervals. Besides, this measurement made it possible to normalize F_0 values across speakers of different genders (cf. Xu & Xu, 2005). The first step in computing normalized F_0 was to calculate the minimum value of F_0 for each speaker and each sentence. This value could be used as an individual baseline for each utterance. Then this baseline value was subtracted from the mean F_0 value in each keyword, yielding a value that was representative of the local pitch movements on the selected interval.

As for the analysis of sentences, the measurement of normalized F_0 was replaced by *pitch span* (Ladd, 1996; Mennen, 2007 and Mennen et al., 2012,

Table 5.4: Summary of the acoustic measurements applied to the data set, with the respective units of measure and a brief description.

Acoustic measurement	**Unit**	**Description**
Duration	ms	Duration of a selected interval
Mean F_0	Hz	Mean F_0 in a selected interval
Minimum F_0	Hz	Minimum F_0 value found in the sentence (baseline)
Maximum F_0	Hz	Maximum F_0 value found in the sentence
Normalized F_0	Hz	Normalized F_0: difference between Mean F_0 and Minimum F_0
Pitch span	Hz	Difference between Maximum F_0 and Minimum F_0
Speaking rate	syllables/s	Total number of syllables divided by total duration of the utterance

see Section 2.5.1), calculated as the difference between maximum and minimum F_0 values across each sentence. This is because a measurement of the mean F_0 value along the whole sentences would have yielded low values, which would not have been representative of the speakers' actual pitch range.

As for speaking rate, this was calculated by dividing the fixed number of syllables in the sentences (7) for the total length of each sentence, following Trofimovich & Baker (2006) and Hincks (2010).

All acoustic measurements were performed automatically using a set of *Praat* scripts that were adapted from preexisting ones or written *ex novo* by the author. The results were saved in comma-separated text files, which were used as *SPSS* data sets for statistical analysis. Similarly to what was done in the annotation phase, the results were manually verified with a visual inspection of every couplet of audio and *TextGrid* files in the *Praat* Editor Window. This procedure was performed in order to detect and control for any visible error that might have been caused by microprosodic events with the risk of altering the results of the acoustic measurements based on F_0 (see

Ladd, 2008).

Chapter 6

Results

6.1 Introduction

In this section, the results of the production study are presented. In the first subsection, between-group data at sentence level are presented. The mean values of duration, speaking rate and pitch span, averaged over speakers and sentences, will be used as indicators of the differences between the productions by NS and non-native speakers and of the acquisition patterns of the L2 speakers.

As for the word-level analysis, the results are presented by group of speakers. The purpose of this analysis will be to determine whether and how the three groups of speakers can mark narrow focus location by means of prosodic cues, namely duration and F_0. The Italian data set will be also analyzed in order to check for the effects of prosodic transfer from L1 to L2.

In each section, the results are presented first by showing tables and bar charts summarizing the descriptive statistics, followed by the results of the statistical tests. The results of the acoustic analyses will be discussed briefly at the end of each subsection.

6.2 Sentence-level analysis

As a first step, the data were analyzed at sentence level. The mean values and standard deviations of the suprasegmental aspects measured are summarized in Tab. 6.1 and presented one by one in the following subsections.

Table 6.1: Total number of sentences, with mean values and standard deviations for duration, speaking rate and pitch span, averaged over sentences and speakers, presented by group.

Group	**N**	**Duration (ms)**		**Speaking rate (syllables/s)**		**Pitch span (Hz)**	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
NS	40	1805.20	181.69	3.92	0.40	26.32	18.53
NNS1	40	2207.15	239.45	3.49	0.32	42.08	19.99
NNS2	40	2315.72	290.83	3.07	0.40	45.79	21.01

6.2.1 Duration

As shown in Fig. 6.1, the mean duration of the sentences produced by NS is shorter than those of both groups of non-native speakers. The sentences produced by NNS1 are longer than the ones produced by NS and shorter that the ones produced by NNS2.

A Kruskal-Wallis test was conducted to evaluate differences between the three groups of speakers (NS, NNS1 and NNS2), with duration as dependent variable and group as fixed factor. The test was significant (χ^2 (2, N=120) = 4.496, $p<0.01$). This non-parametric test was chosen instead of an Analysis of Variance (ANOVA) after a Levène's test of Equality of Error Variances had shown that the data distributions among groups were not homogeneous ($p<0.05$).

Follow-up Mann-Whitney U tests were conducted to obtain pairwise comparisons between the three groups, controlling for Type I error across tests

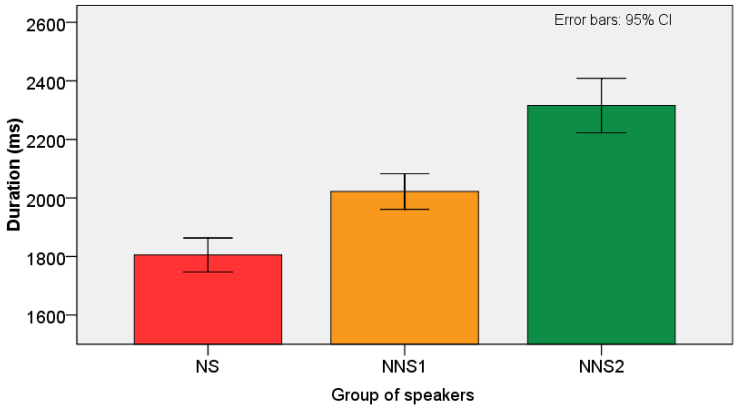

Figure 6.1: Bar chart showing the mean duration of sentences by group, averaged over speakers.

by using the Bonferroni correction (p = α/number of comparisons). Pairwise comparisons between the three groups showed significant differences in all cases, as summarized in Tab. 6.2.

Table 6.2: Results of Mann-Whitney U tests to determine pairwise differences in duration between groups of speakers.

Group	N	Z	p
NS vs. NNS1	80	-4.364	$<$**0.01**
NS vs. NNS2	80	-6.544	$<$**0.01**
NNS1 vs. NNS2	80	-4.446	$<$**0.01**

6.2.2 Speaking rate

The mean values of speaking rate, measured by dividing the number of syllables (7) by the total duration of each sentence, are summarized in the bar chart in Fig. 6.2.

The mean speaking rate in the sentences produced by NS is higher than that of both groups of non-native speakers. The speaking rate of NNS1

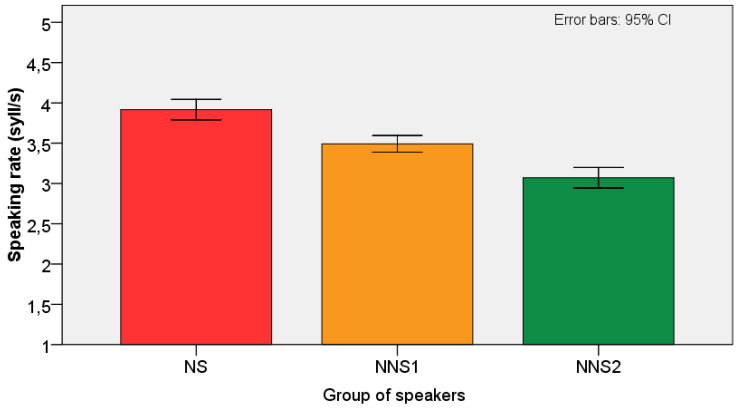

Figure 6.2: Bar chart showing the mean speaking rate of sentences by group, averaged over speakers.

sentences is higher than that of NNS2. Similarly to what was observed for duration, NNS1 present values that are between the ones measured for NS and NNS2. The statistical significance of the speaking rate values was tested with a one-way Analysis of Variance (ANOVA) with speaking rate as dependent variable and group as fixed factor. The ANOVA showed a significant effect of group on speaking rate ($F(2, 117)=50.707$, $p<0.01$). Pairwise comparisons between the three different groups showed significant differences in all cases ($p<0.01$, with Bonferroni correction).

6.2.3 Pitch Span

The mean values of pitch span, which was calculated as the difference between the local maximum and minimum F_0 values in each sentence, are summarized in the bar chart in Fig. 6.3.

The productions by NS present a narrower pitch span, as compared to both groups of non-native speakers. NNS1 present lower values in pitch span than NNS2, although the difference between NNS1 and NNS2 is less marked than the one between NNS1 and NS.

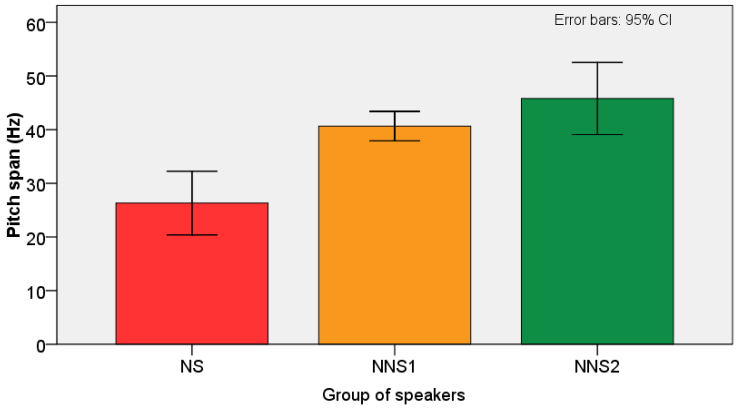

Figure 6.3: Bar chart showing the mean pitch span by group, averaged over speakers.

Since the Levène's test had shown that the data distribution among groups was not homogeneous ($p<0.05$), the pitch span values were analyzed by conducting a Kruskal-Wallis test with mean pitch span as dependent variable and group as fixed factor. The test was significant (χ^2 (2, N=120) =41.058, $p<0.01$).

Follow-up Mann-Whitney U tests were conducted to obtain pairwise comparisons between the three groups, controlling for Type I error across tests by using the Bonferroni correction. Pairwise comparisons between the three groups showed significant differences in two out of three cases, as summarized in Tab. 6.3.

Table 6.3: Results of Mann-Whitney U tests to determine pairwise differences in pitch span between groups of speakers.

Group	N	Z	p
NS vs. NNS1	80	-5.822	**<0.01**
NS vs. NNS2	80	-5.254	**<0.01**
NNS1 vs. NNS2	80	-0.25	0.802

The pitch span resulted significantly wider for both groups of non-native

speakers, when compared to NS, but the difference between NNS1 and NNS2 was not significant.

The results suggested that the difference between the Italian and the NS speakers might be the result of prosodic transfer from the L1. In order to verify this hypothesis, the mean pitch span values recorded in the Italian L1 data set were considered and compared to the NS ones. The mean pitch range in Italian was 88.66 Hz (SD=27.68), which is sizably higher that the one of NS, calculated in 26.32 Hz (SD=18.53). A series of Mann-Whitney U tests showed that the difference between the pitch span values found in the Italian L1 data set were significantly higher than the ones obtained not only for NS, but also for NNS1 and NNS2 ($p<0.01$, with Bonferroni correction).

6.2.4 Discussion

The results at sentence level show consistent differences between the L1 and L2 speakers.

The sentences produced by NS are significantly shorter than the ones produced by NNS1. In turn, the productions by NNS1 are significantly shorter than the ones by NNS2. The lack of vowel reduction and the addition of epenthetic vowels (see Section 6.4) have certainly contributed to the longer duration of the sentences produced by NNS2. This difference in duration between the productions of NNS1 and NNS2 can be seen as evidence for a progressive tuning towards the native model. NNS1 have indeed produced shorter sentences, which seem to imply that the acquisition of English rhythmic aspects is in progress.

NS showed a significantly higher speaking rate when compared to both groups of non-native speakers. NNS2 were the ones obtaining the lowest values, with NNS1 showing a significant higher speaking rate, again showing progress towards the target native model.

The analysis of pitch span showed that the NS have significantly lower values when compared to both groups of non-native speakers. Although

NNS1 speakers still showed a tendency towards the native values, the difference between the productions by NNS1 and NNS2 was not statistically significant.

The analysis of the Italian speakers' pitch span values in English and in Italian L1 showed that the mean pitch span values of the Italians are significantly higher than any English production. This means that, when speaking their L1, Italians use a wider pitch span in Italian L1 than in English L2. In both cases, the Italians' pitch span is higher than the English NS. This suggests that, in the first place, pitch span implementation seems to be structurally different in the two languages: it is wider in Italian and narrower in English; in the second place, this wider pitch span is transferred from the L1 to the L2, confirming H_4 (see Section 5.1).

6.3 Word-level analysis

In this section, the results will be presented by group, comparing the acoustic measurements for the keywords that are in focus to the ones that are not. As mentioned in section 5.2.1, the two keywords that will be analyzed in this study will be sentence subjects and verbs, which will be referred to as 'S' and 'V', respectively.

6.3.1 Native English speakers (NS)

The results of the acoustic analysis of the productions by NS are summarized in Tab. 6.4.

6.3.1.1 Duration

The results of the duration measurements are summarized in the two panels composing Fig. 6.4. Each panel corresponds to one focus condition ('S in focus' or 'V in focus').

Table 6.4: Mean values and standard deviations of duration and normalized F_0 for the NS group, averaged over sentences and speakers, presented by word in focus.

Native English speakers (NS)

Sentences with *subject* (S) in focus					
	N	**Duration (ms)**		**normalized F_0 (Hz)**	
		Mean	**SD**	**Mean**	**SD**
subject	20	402.88	91.28	32.15	14.02
verb	20	379.76	48.15	19.90	8.57

Sentences with *verb* (V) in focus					
	N	**Duration (ms)**		**normalized F_0 (Hz)**	
		Mean	**SD**	**Mean**	**SD**
subject	20	417.49	48.29	31.86	15.09
verb	20	403.08	29.50	42.06	16.30

When comparing the mean values of duration, S appears slightly longer than V, regardless of the focus condition. However, the differences between the duration of the two keywords are not statistically significant in either focus condition.

6.3.1.2 Fundamental frequently (F_0)

The results obtained for normalized F_0 are summarized in Fig. 6.5. Each panel corresponds to one focus condition ('S in focus' or 'V in focus').

When in focus, S is uttered with a significantly higher F_0 as compared to V. An independent-samples t-test was conducted to compare the duration of S and V with S in focus. The results of the test showed that there was a significant difference in normalized F_0 between S (M=32.15, SD=14.03) and V (M=19.91, SD=8.58) when S was in focus: t(31.46)=3.331, p=0.002.

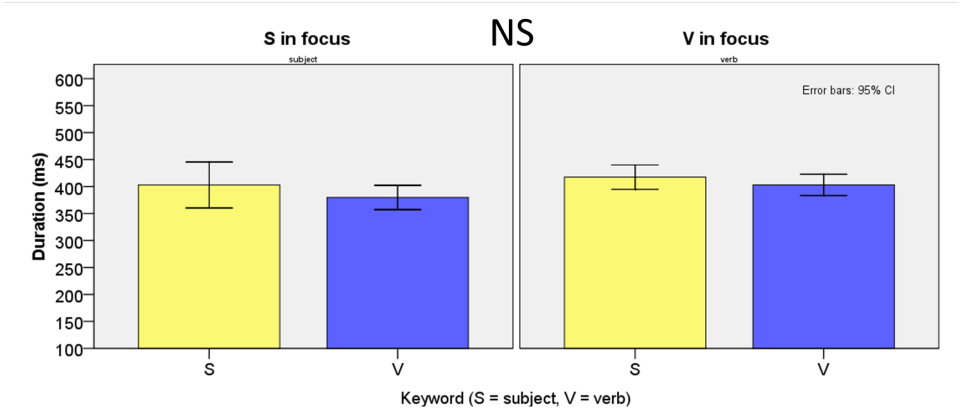

Figure 6.4: Mean duration of the keywords S and V for the NS group, averaged over speakers and sentences, with S (left panel) V (right panel) in focus.

When V is in focus the difference is smaller as compared to the case of S in focus. In addition the difference between the F_0 values of S and V is not statistically significant.

6.3.1.3 Discussion

Duration does not seem to play an active role in the phonetic realization of narrow focus by NS. There was no significant difference between keywords in either focus condition, and no definite patterns emerged from the data.

As for F_0, the results show that the marking of narrow focus location is indeed affected by modifications in pitch. When S is in focus, the difference in F_0 between S and V is significant, with S having a higher F_0 than V. In contrast, when V is in focus, the difference between S and V does not reach statistical significance. The latter is realized with sustained F_0 values that are very close to the ones that characterize the former.

This difference in F_0 between S and V seems to be the crucial factor in narrow focus marking from the point of view of production. Its perceptual relevance will be tested in the perception study (see Chapter 5).

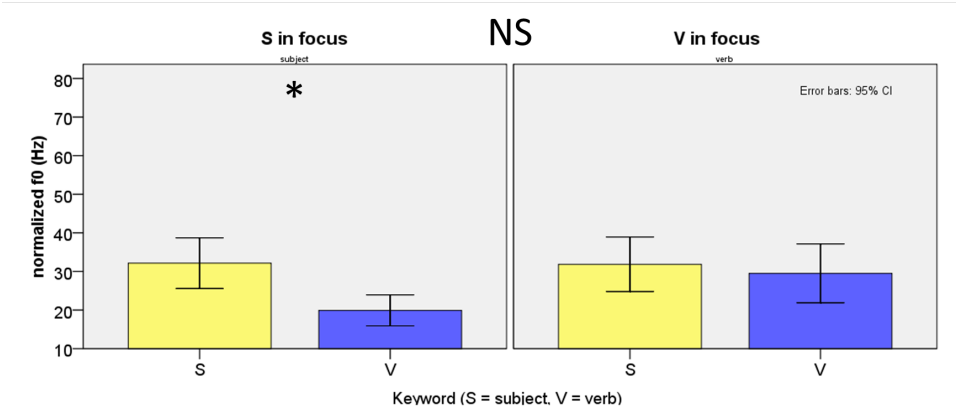

Figure 6.5: Mean normalized F_0 of the keywords S and V for the NS group, averaged over speakers and sentences, with S (left panel) V (right panel) in focus. The asterisk indicates a statistically significant difference ($p<0.05$).

To conclude, the results of the acoustic analysis confirmed that NS can mark narrow focus by prosodic means, in particular by modulating pitch, as shown in previous studies (see Chapter 2) and as predicted by H_1 (see Section 5.1).

6.3.2 Non-native speakers with higher competence in L2 (NNS1)

The results of the acoustic measurements of the NNS1 productions are summarized in Tab. 6.5.

6.3.2.1 Duration

The results of the duration measurements are summarized in Fig. 6.6. Each panel corresponds to one focus condition ('S in focus' or 'V in focus').

As observed for the NS group, S is produced with a somewhat longer duration when compared to V, regardless of its focus condition. However,

Table 6.5: Mean values and standard deviations of duration and normalized F_0 for the NNS1 group, averaged over sentences and speakers, presented by word in focus.

Non-native speakers with higher competence (NNS1)

Sentences with *subject* (S) in focus					
	N	**Duration (ms)**		**normalized F_0 (Hz)**	
		Mean	**SD**	**Mean**	**SD**
subject	20	477.65	94.54	61.98	15.48
verb	20	432.58	55.25	34.21	12.35

Sentences with *verb* (V) in focus					
	N	**Duration (ms)**		**normalized F_0 (Hz)**	
		Mean	**SD**	**Mean**	**SD**
subject	20	533,64	92.08	61.30	14.33
verb	20	476,50	70.66	31.27	9.57

the difference between the duration of S and V is not statistically significant in either focus condition.

6.3.2.2 Fundamental frequency (F_0)

The results concerning F_0 in the NNS1 productions are summarized in Fig. 6.7. Each panel corresponds to one focus condition ('S in focus' or 'V in focus').

S is uttered with a significantly higher F_0 than V in both focus conditions. An independent-sample t-test was conducted to compare F_0 in S and V when S is in focus. The results of the test showed that there was a significant difference in F_0 between S (M=61.98, SD=15.48) and V (M=34.21, SD=12.34) when S is in focus: $t(38, 14)=6.270$, $p<0.001$. A second t-test was conducted to compare F_0 between S and V with V in focus. The results

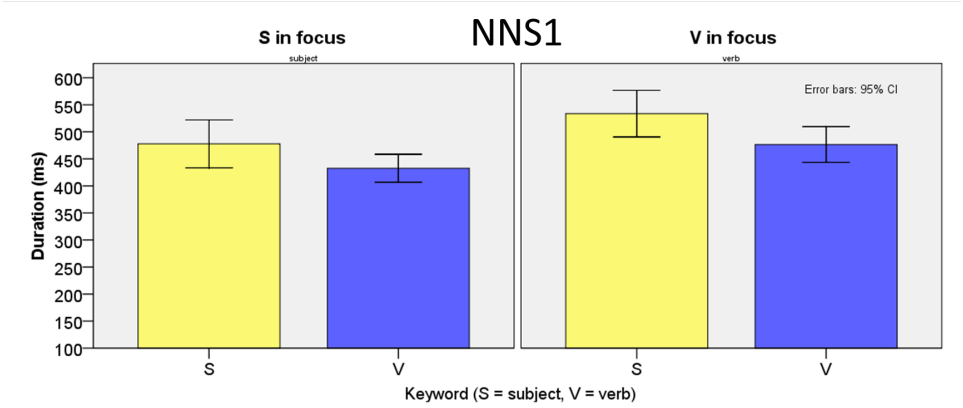

Figure 6.6: Mean duration of the keywords S and V for the NNS1 group, averaged over speakers and sentences, with S (left panel) V (right panel) in focus.

of the test showed that there was also a significant difference in F_0 between S (M=61.30, SD=14.32) and V (M=31.27, SD=9.57) when V is in focus: t(33, 14)=7.795, p<0.001.

6.3.2.3 Discussion

The significant differences in the F_0 values of S and V suggest that speakers in NNS1 have apparently learnt to differentiate words by modulating pitch, similarly to the NS productions. This confirms the hypothesis that NNS1 progressively tune towards the L2 model by learning to use pitch as a marker of prominent information (H_2, see Section 5.1). The hypothesis of a progressive tuning seems to be confirmed also by comparing the results obtained by NNS1 to the ones obtained by NNS2 (see Section 6.3.3).

However, it is important to point out that the differences in F_0 do not depend on the focus condition. These differences are rather determined by the position of the keyword in the sentence: S is systematically produced with a higher F_0 than V regardless of the focus condition. These results suggest that NNS1 have not completely acquired the prosodic strategies in

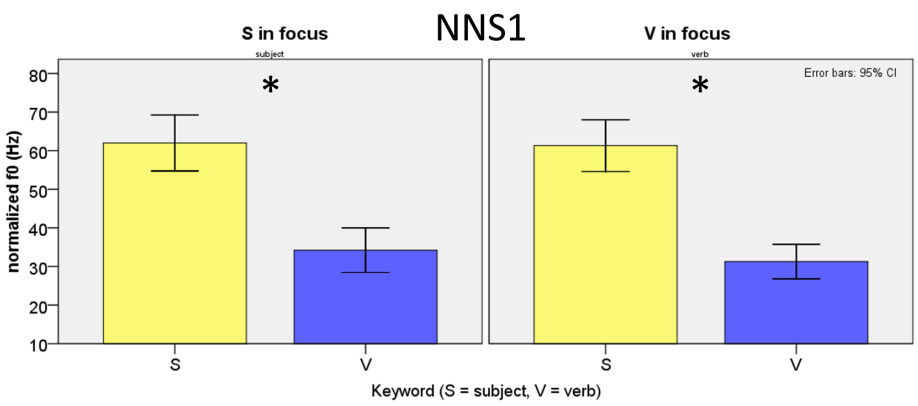

Figure 6.7: Mean normalized F_0 of the keywords S and V for the NNS1 group, averaged over speakers and sentences, with S (left panel) V (right panel) in focus. The asterisk indicates a statistically significant difference ($p<0.05$).

focus marking that characterize the productions by NS.

As for duration, no systematic patterns were found, suggesting that this acoustic cue was not actively used to mark narrow focus. This is in line with what was observed in the Italian L1 data set: even in their L1 the Italians did not actively use duration to mark narrow focus, as shown in Section 6.3.4.3.

To conclude, the NNS1 provide evidence of acquisition of native-like focus marking strategies, but have not achieved mastery of these strategies, as they lag behind the native speakers' models. This will be tested in perception study in Part IV.

6.3.3 Non-native speakers with lower competence in L2 (NNS2)

The results of the acoustic analysis of the productions by NS are summarized in Tab. 6.6.

Table 6.6: Mean values and standard deviations of duration and normalized F_0 for the NNS2 group, averaged over sentences and speakers, presented by word in focus.

Non-native speakers with lower competence (NNS2)

Sentences with *subject* (S) in focus					
	N	**Duration (ms)**		**normalized F_0 (Hz)**	
		Mean	**SD**	**Mean**	**SD**
subject	20	526.56	96.31	58.93	17.49
verb	20	564.47	98.82	54.34	15.30

Sentences with *verb* (V) in focus					
	N	**Duration (ms)**		**normalized F_0 (Hz)**	
		Mean	**SD**	**Mean**	**SD**
subject	20	572,09	114.35	54.51	25.59
verb	20	504,57	77.34	53.98	26.16

6.3.3.1 Duration

The results of the duration data are summarized in Fig. 6.8. Each panel corresponds to one focus condition ('S in focus' or 'V in focus').

The bar chart shows opposite tendencies for the two focus conditions: when S in in focus V is longer, when V in focus S is longer. The results of two independent-samples t-tests showed that the difference between the mean duration of S and V when S in focus is not significant, while the difference between S (M=572.10, SD=114.26) and V (M=504.45, SD=77.33) is significant when V is in focus: $t = 2.193$, $p<0.05$.

6.3.3.2 Fundamental frequency (F_0)

The results of normalized F_0 are summarized in Fig. 6.9. Each panel corresponds to one focus condition ('S in focus' or 'V in focus').

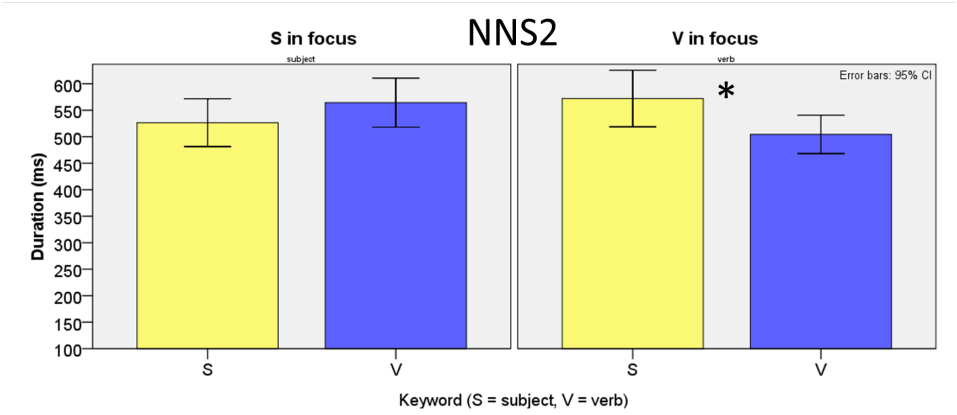

Figure 6.8: Mean duration of the keywords S and V for the NNS2 group, averaged over speakers and sentences, with S (left panel) V (right panel) in focus. The asterisk indicates a statistically significant difference.($p<0.05$)

When analyzing mean F_0 values, no significant difference and no systematic patterns were found in the productions of the NNS2 speakers. The keywords were uttered with small changes in F_0, with no sizable effects caused by focus condition.

6.3.3.3 Discussion

The results suggest that NNS2 do not mark focus with prosodic cues, at least not in a consistent way. The values of duration did change when S was in focus as compared to when V was in focus, but this change seems more likely to be due to chance rather than to a use of duration as a means to mark focus. Indeed, if duration were used to mark focus, one would expect the word in focus to be longer, while the NNS2 productions of V in focus show the opposite. A comparison with the results in Italian (see Section 6.3.4.3) excluded any systematic function of duration as a narrow focus marker for the NNS2.

As for F_0, the productions by the NNS2 appear undifferentiated. This

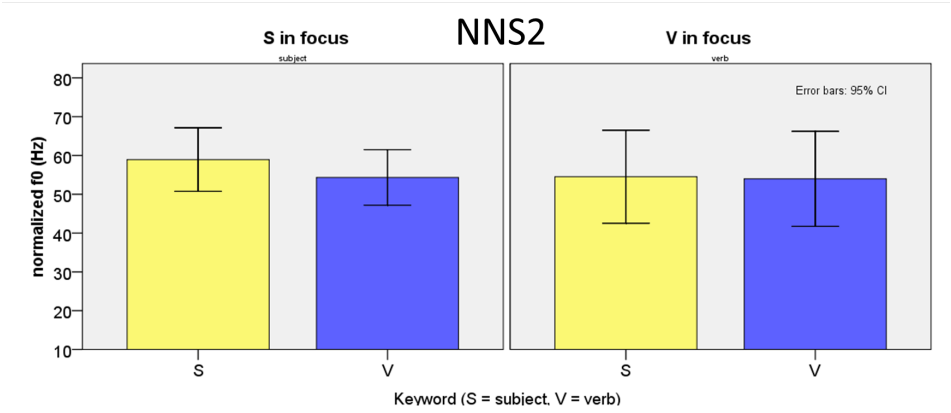

Figure 6.9: Mean normalized F_0 of the keywords S and V for the NNS2 group, averaged over speakers and sentences, with S (left panel) V (right panel) in focus.

suggests that F_0 is not used to mark focus by the NNS2.

To conclude, the results of the acoustic analysis confirmed the hypothesis that the NNS2 do not use pitch modulation as a focus marking strategy, in contrast with the results of the NNS1 (H_3, see Section 5.1).

6.3.4 Italian L1 speakers (IT)

The results of the acoustic measurements for data set in Italian L1 (IT) are summarized in Tab. 6.7. These data are base on the productions by all eight Italian speakers involved in the study.

6.3.4.1 Duration

The results for duration are summarized in Fig. 6.10. Each panel corresponds to one focus condition ('S in focus' or 'V in focus').

When comparing the mean values of duration in the Italian L1 data set, S is realized with longer durations than V, regardless of the focus condition. An independent-samples t-test was conducted to compare duration in S and

Table 6.7: Mean values and standard deviations of duration and normalized F_0 for the Italian L1 data set (IT), averaged over sentences and speakers, presented by word in focus.

Italian L1 speakers (IT)

Sentences with *subject* (S) in focus					
	N	**Duration (ms)**		**normalized F_0 (Hz)**	
		Mean	**SD**	**Mean**	**SD**
subject	20	424.57	76.71	78.03	43.54
verb	20	349.23	55.93	52.72	56.14

Sentences with *verb* (V) in focus					
	N	**Duration (ms)**		**normalized F_0 (Hz)**	
		Mean	**SD**	**Mean**	**SD**
subject	20	448.00	62.79	74.54	46.42
verb	20	415.75	72.36	67.86	39.77

V with S in focus: The test showed that there is a significant difference in duration between S (M=424.57, SD=76.70) and V (M=349.23, SD=55.93) when S is in focus: $t(78)=5.019$, $p<0.01$. A second independent-samples t-test was conducted to compare duration in S and V with V in focus. This test showed that there is a significant difference in duration also between S (M=448, SD=62.789) and V (M=415.75, SD=72.364) when V is in focus: $t(78)=2.129$, $p=0.036$.

6.3.4.2 Fundamental frequency (F_0)

The results of normalized F_0 are summarized in Fig. 6.11. Each panel corresponds to one focus condition ('S in focus' or 'V in focus').

Similarly to what was observed for duration, S is produced with a higher F_0 when compared to V, no matter if in focus or not. However, the differ-

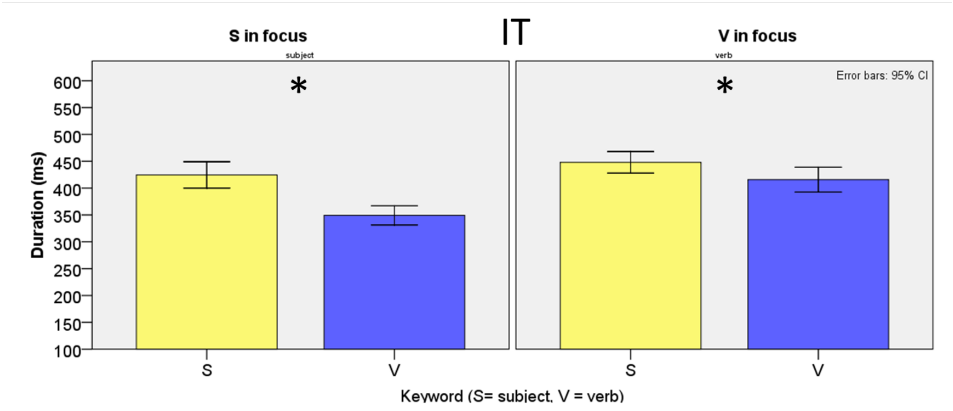

Figure 6.10: Mean duration of the keywords S and V for the IT group, averaged over speakers and sentences, with S (left panel) V (right panel) in focus. The asterisk indicates a statistically significant difference (p<0-05).

ences in F_0 between S and V are not statistically significant in either focus condition.

6.3.4.3 Discussion

When speaking their L1, the Italians produce S with a significantly longer duration than V, regardless of the focus condition of the word. It seems that this difference is related to the position of the keyword in the sentence, rather than to the focus condition. This is an interesting result, since it suggests that in Italian duration does not play a role in narrow focus marking. On the one hand, this was somewhat unexpected, as in Italian duration is the main acoustic correlate of prominence at word level, that is, in the realization of word stress (Magno Caldognetto et al., 1983; Bertinetto, 1981). On the other hand, other studies based on narrow contrastive focus have shown that F_0 can be a more reliable cue than duration for sentence level prominence in Italian (Kori & Farnetani, 1983; Magno Caldognetto & Fava, 1983). As for the Italian L1 data set presented here no pattern was found also in the results

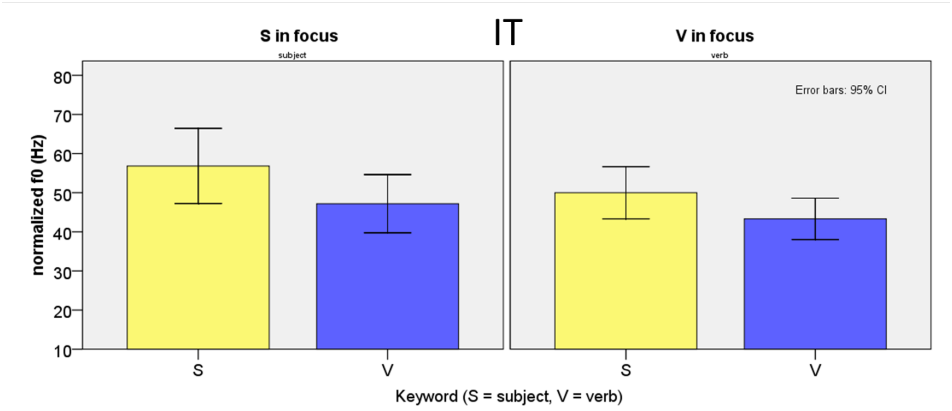

Figure 6.11: Mean normalized F_0 of the keywords S and V for the IT group, averaged over speakers and sentences, with S (left panel) V (right panel) in focus.

of the normalized F_0 measurement, suggesting that neither pitch nor duration play an active role in marking narrow focus in Italian. Further research is needed to determine the acoustic correlates of narrow non-contrastive focus in Italian. However, as suggested in Section 2.6, it is possible that the non-contrastive type of narrow focus is not at all prosodically characterized in Italian, and word order strategies would be used instead.

The results of the acoustic analysis therefore support the hypothesis that in Italian the marking of narrow focus location is not conveyed by prosodic means (H_4, see Section 5.1). As shown in Section 2.6, this lack of acoustic characterization of focus is compensated by the use of word order and syntax as preferential strategies for focus marking (cf. Ladd, 1996; Vallduvì, 1991).

6.4 Presence of epenthetic vowels

During the annotation process, it was found that the productions by NNS2 were characterized by a pervasive presence of epenthetic vowels in word-final position. An epenthetic vowel is a "vowel inserted into a phonological envi-

ronment to repair a marked or illegal structure" (Repetti, 2012); when this vowel is added in word-final position, it is also referred to as paragogic vowel. The addition of epenthetic vowels is frequently found in L2 speech, especially in early stages of second-language acquisition, where learners struggle to reproduce syllable structures and syllable clusters that are not present in their L1. Italian speakers of English L2 are particularly known to produce paragogic vowels (Duguid, 1997), and the addition of epenthetic vowels is often used in popular media as a stereotypical feature of Italian accent. The reason for this phenomenon is to be found in the syllable structure of Italian: since "[t]he native lexicon of Italian is characterized by the nearly total absence of consonant ending words" (Passino, 2005: 1), Italian speakers tend to accommodate the pronunciation of foreign words ending in consonants by adding a short vowel sound to adapt the unfamiliar sequence. These paragogical vowels are normally shorter than lexical vowels and produced as very short instances of [e] or [ə] (Repetti, 2012).

A full-scale acoustic analysis of the epenthetic vowels (e.g., plotting their formant structure) was beyond the scope of this study. In the production data presented in this study, every unexpected occurrence of a vowel sound longer than 30 ms was considered a paragogic vowel. In the data presented in this dissertation, epenthetic vowels appeared to be systematically added at the end of words with consonants or consonant clusters in final position in the productions by NNS2. In contrast, they were absent from the productions by NNS1. This result suggested that the production of epenthetic vowels decreases as the L2 competence increases.

The presence of epenthesis in the productions by NNS1 was quantified by using a measure called *epenthesis ratio*. The author devised this method to obtain a straightforward indication of the presence and impact of epenthetic vowels in the non-native productions. The epenthesis ratio was calculated by dividing the total number of actual occurrences of epenthetic vowels by the total number of potential occurrences of epenthetic vowels in the sentences,

as shown in (1).

(1) $$\text{Epenthesis ratio} = \frac{\text{number of actual occurrences}}{\text{number of potential occurrences}}$$

The potential occurrences were determined by counting all instances of words ending with: CVC (e.g. red), CVCC (e.g., runs), and CVCCC (e.g., walks) in the 20 sentences composing the NNS2 data set. The resulting total number of potential occurrences was 188 (88 for sentences with S in focus and 100 for sentence with V in focus).

The overall epenthesis ratio for NNS2 productions was 83/188=0,44, and the ratio is particularly high in the S + V sequences (46/68 = 0.67). However, it is after the sequences at the end of the main intonational phrase (i.e., after the verb and before the following prepositional phrase), that epenthesis is almost always present, reaching the following ratio: 36/40=0.9 (see Fig. 6.12 for an example).

These results suggest that the production of epenthesis might be triggered by the position of the word in the utterance: if the word is at an intonation boundary, there is a higher chance for the occurrence of an epenthetic vowel. In addition, impressionistic observations of the f0 contours and the corresponding transcriptions showed that epenthetic vowels at the end of an intonation boundary are often pronounced with a stray rising tone. Fig. 6.12 shows an example of this phenomenon, which was frequently found in the productions by NNS2.

As mentioned before, a more specific study of epenthetic vowels goes beyond the scope of this thesis. However, it was important to highlight the impact of this phonological phenomenon in the prosody of NNS2 productions, both in terms of duration and pitch. The addition of paragogical vowels surely played a role in determining the overall duration of the productions by NNS2. It is also possible that the peaks in f0 that were frequently found in connection with the epenthetic vowels contributed to the wide pitch span observed for NNS2.

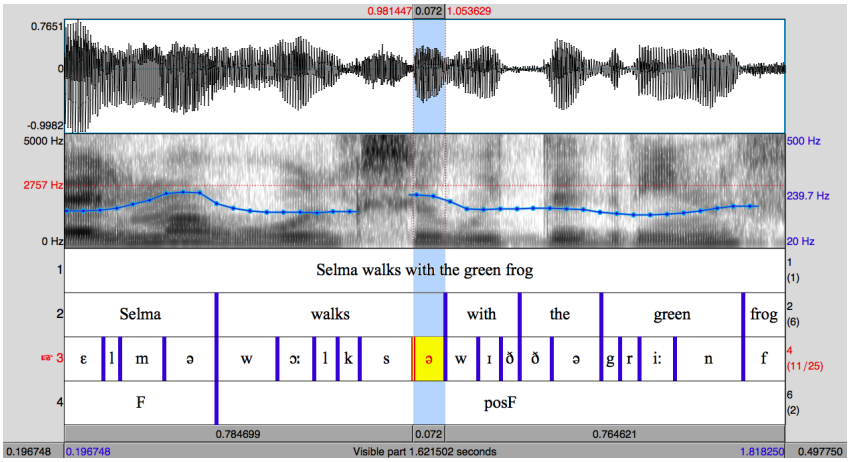

Figure 6.12: Detail of a sentence produced by a NNS2 speaker. The epenthetic vowel is highlighted.

Part III

Perception Study

Chapter 7

Experiment 1

7.1 Rationale and hypotheses

As presented in detail in Chapter 6, the results of the acoustic analysis showed the following major trends:

1. Native speakers (NS) systematically mark relevant information by modulating pitch;
2. Non-native speakers with a higher competence in English (NNS1) modulate pitch to mark prominence, but they implement it in a way that is not completely consistent with the native model;
3. Non-native speakers with a lower competence in English (NNS2) fail to mark focus prosodically;
4. When speaking their L1, Italian speakers do not to mark narrow focus prosodically;
5. Both non-native groups of speakers (NNS1 and NNS2) present a significantly wider pitch span when compared to NS.

Following the above findings, a perception experiment was conducted with the aim of answering the following question: can native listeners identify narrow focus when they listen to an utterance without any contextual information? To the author's knowledge, only a few studies have tackled similar questions from the perceptual perspective, and examined especially American English (e.g., Bishop, 2011). Moreover, none of these studies has investigated the perception of narrow non-contrastive focus in British English. The present perceptual experiment was also run on Italian listeners, in order to verify their capability of recognizing narrow focus when presented with sentences in English (uttered by native and non-native speakers) and in Italian (uttered by native speakers).

The experiment was set out to test the following hypotheses:

- H_1: When listening to productions by NS, native and non-native listeners can correctly recognize the location of narrow focus even without extra contextual information.

- H_2: When listening to productions by NNS1, native and non-native listeners can still correctly detect narrow focus, although with less success as compared to the productions by NS. Conversely, it is expected that none of the two groups of listeners can correctly identify focus in the NNS2 productions.

- H_3: When listening to productions in Italian L1, Italian listeners cannot correctly recognize the location of narrow focus in absence of any extra contextual information.

7.2 Methodology

7.2.1 Stimuli

The set of stimuli presented in this perception experiment consisted in the entire corpus of sentences that were analyzed in the production study. The sentences were produced by three groups of speakers, consisting of 4 speakers each: English native speakers (NS), non-native speakers with a higher competence in English (NNS1) and non-native speakers with a lower competence in English (NNS2). For each speaker, 5 sentences with narrow focus on the sentence subject (S in focus condition) and 5 sentences with narrow focus on the verb (V in focus condition) were used. As a result, the total number of stimuli was 120 (4 speakers x 5 sentences x 2 focus conditions x 3 groups = 120). Further information about the composition of the groups (e.g., gender, average age, level definition) and about the recording setup can be found in Section 5.2.1.

For the Italian listeners only, the experiment presented an extra block of sentences in Italian, extracted from the set recorded and analyzed in the production study. This set was composed like the other three blocks of stimuli (4 speakers x 5 sentences x 2 focus conditions x 1 group = 40). As a result, the Italian listeners were tested on 160 stimuli.

The sentences used in this experiment were natural, that is, no digital manipulation was applied. The stimuli corresponded to the original sentences that were recorded for the production study.

7.2.2 Subjects

The group of British English native listeners consisted of 22 individuals. Their average age was 24,5 years, and their professional background was varied. None of them reported any hearing problems. At the moment of taking the test no participants claimed to be able to speak Italian or that

they were living or had lived in Italy.

The group of Italian native listeners consisted of 22 individuals. Their average age was 30,6 years. Their professional background was also varied, and none of them reported any hearing problems. All listeners declared that they were able to speak and understand English, and self-reported levels of English L2 ranging from elementary to advanced.

7.2.3 Task and procedure

The experiment was presented using the *LimeSurvey* survey presentation software (Schmitz, 2012) on a laptop personal computer connected to a headset, in a silent environment. Before starting the experiment, the subjects were asked to fill in a consent form and complete a brief questionnaire to collect information about their geographical origin, age, profession and language background. The subjects were then presented with detailed on-screen instructions about the experimental procedure and the task they were asked to perform (see Appendix C).

The task was based on Büring's *Question-Answer Congruence* hypothesis (see Section 2.3): in a reply to a wh-question, narrow "foci correspond to the wh-expression in a preceding constituent question" (Büring, 2007: 447). Assuming the validity of this correspondence, the experimental task was built to ask the subjects the following question: 'when you listen to an answer out of its context, can you correctly guess the question that triggered that answer?'

The participants were asked to listen to the sentences one by one and to select which question was more likely to have triggered the sentence as an answer, choosing from two options. One option represented a question that prompted focus on the subject of the sentence (e.g., "Who runs with the green frog?" "Bobbie runs with the green frog."), the other one on the verb (e.g., "What does Bobbie do with the green frog?" "Bobbie jumps with the green frog."). The program automatically played each stimulus once,

but the subjects were allowed to listen to the sentences as many times as they wished, in order to make informed guesses and to reduce the risk of providing random responses. After expressing their choice by selecting one of the two options, the subjects had to press the "Next" button to prompt the presentation of the following stimulus. An example of the presentation of an item of Experiment 1 can be found in Fig. 7.1.

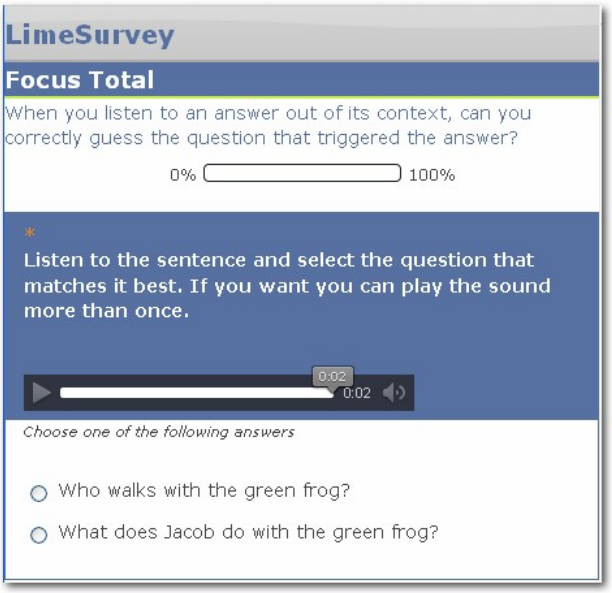

Figure 7.1: Screenshot of the presentation of a stimulus in Experiment 1 with the software *LimeSurvey*.

The actual experiment was preceded by a short training session, where the subjects could familiarize with the task and with the interface. The 24 sentences composing the training session were similar to the ones used in the actual experiment. The only difference was that the sentences of the training set were spoken by voices that were not included in the experimental set.

The 120 stimuli were pooled together in a single block of items, where the tokens were presented in a different randomized order for each participant to control for possible memory effects. At the end of the experiment, the

subjects received immediate feedback on their performance on a screenshot reporting the total number of correct responses. The Italian listeners were also presented a set of 40 extra stimuli in Italian. This set of stimuli was grouped in a second block presented after the one in English.

The average time to complete the whole experiment, including the training session, ranged from approximately 15 minutes (for the English listeners) to approximately 20 minutes (for the Italian listeners).

7.3 Results

The results of the experiment are summarized in Tab. 7.1 and 7.2.

Table 7.1: Total numbers of correct responses with mean and standard deviation, averaged by group of speakers over single speakers and sentences.

Speaker group	**English listeners**			**Italian listeners**		
	N	**Mean**	**SD**	**N**	**Mean**	**SD**
NS	40	31.73	1.78	40	28.64	3.37
NNS1	40	26.91	2.64	40	25	3.22
NNS2	40	22.73	3.56	40	20.05	3.5
IT	-	-	-	40	21.05	2.65

Tab. 7.1 shows the mean number of correct responses given by the two groups of native listeners, along with standard deviation, divided by the three (or four, in the case of Italian listeners) groups of speakers.

Tab. 7.2 shows the mean number of correct responses given by the two groups of native listeners along with standard deviation, divided by the three (or four, in the case of Italian listeners) groups of speakers and by focus condition (*S in focus* or *V in focus*).

The next two sections will discuss the results obtained by the two groups of native listeners, English and Italian, respectively.

Table 7.2: Total numbers of correct responses with mean and standard deviation, averaged by group of speakers over single speakers and sentences.

Speaker group	Focus	English listeners			Italian listeners		
		N	Mean	SD	N	Mean	SD
NS	S	20	14.09	2.11	20	12.23	2.20
	V	20	17.64	1.29	20	16.41	3.49
NNS1	S	20	8.18	3.62	20	6.00	4.36
	V	20	14.55	2.67	20	14.05	4.75
NNS2	S	20	11.27	2.78	20	10.23	3.05
	V	20	15.64	2.79	20	14.77	4.48
IT	S	-	-	-	20	6.45	3.60
	V	-	-	-	20	14.59	4.01

7.3.1 English listeners

Fig. 7.2 shows the mean number of correct responses given by English listeners. In this case the two focus conditions are pooled together to have a general vision of the results, presented by group of speakers.

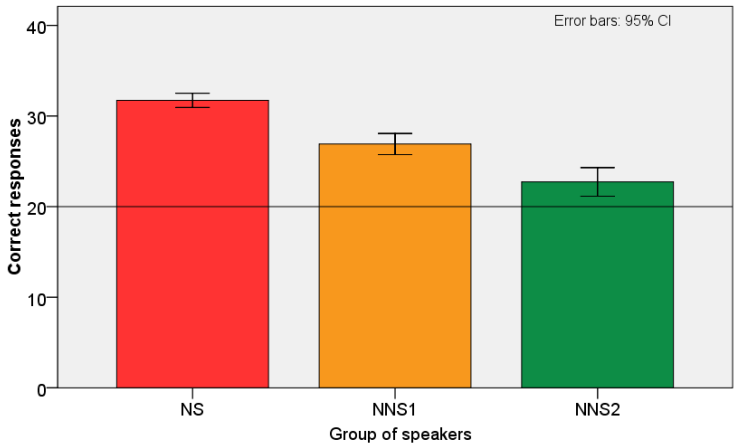

Figure 7.2: Mean number of correct responses (out of 40) given by English native listeners per group, averaged over sentences.

As the figure shows, the mean number of the English native listener's correct responses for the sentences produced by the NS is higher than the one observed for both groups of non-native speakers. As for the non-native productions, the results achieved for NNS1 appear higher than the ones achieved by English native listeners when judging NNS1 productions.

A series of one-sample t-tests was performed to test whether the number of correct responses of each group was significantly different from chance. Since the data sets consisted of 40 items each and the experiment was based on a two-alternative forced-choice paradigm, the chance level was 20 (50% of correct responses). The results of the one-sample t-tests are summarized in Table 7.3.

Table 7.3: Results of one-sample t-tests per group against chance level (=20).

Group	**N**	**mean**	**SD**	**t**	**p**
NS	40	31.73	1.78	30.94	$<$**0.01**
NNS1	40	26.91	2.64	12.91	$<$**0.01**
NNS2	40	22.73	3.56	3.59	$<$**0.01**

The results of the one-sample t-tests show that the number of correct responses obtained for all three groups was significantly above chance level ($p<0.01$).

The mean numbers of correct responses were analyzed by conducting a one-way Analysis of Variance (ANOVA) with mean number of correct responses as dependent variable and group as fixed factor. The ANOVA showed a significant effect for group on the mean number of correct responses ($F(2, 63) = 3.820$, $p<0.05$). Pairwise comparisons between the three different groups showed significant differences in all cases ($p<0.01$, with Bonferroni correction).

In order to have a deeper understanding of the results, the numbers of correct responses were also analyzed by keywords in focus, that is, sentence

subject (S) or verb (V), as summarized by the values reported in Tab. 7.2 and by the bar charts represented in Fig. 7.3.

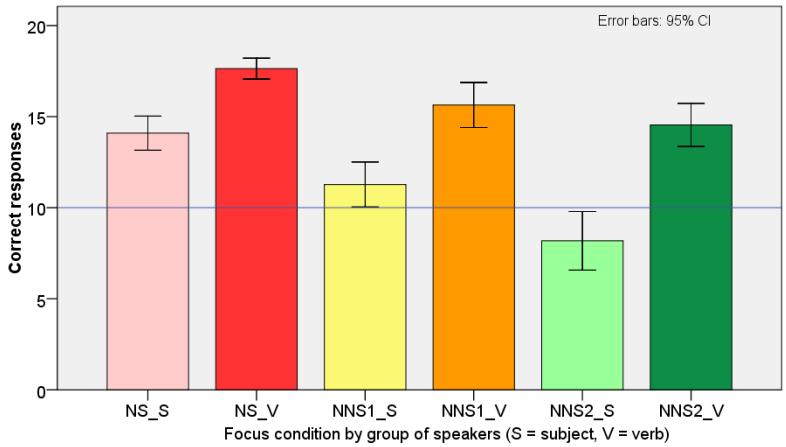

Figure 7.3: Number of correct responses (out of 20) given by English listeners and averaged by group and focus condition (S = subject in focus; V = verb in focus.

As Fig. 7.3 shows, the English listeners obtained a higher number of correct responses when responding to the productions where V was in focus as compared to the ones where S was in focus. This trend becomes more marked when the English listeners had responded to productions by NNS1 and even more when they had responded to productions by NNS2. The significance of these results was tested with a one-way Analysis of Variance (ANOVA) with mean number of correct responses as dependent variable and focus condition as fixed factor. The ANOVA showed a significant effect for focus condition on mean number of correct responses ($F(5, 126) = 35.529$, $p<0.01$). Pairwise comparisons within the three different groups showed significant differences in all oppositions between S vs. V focus conditions (NS_S vs. NS_V, NNS1_S vs. NNS1_V, NNS2_S vs. NNS2_V: for all pairs $p<0.01$, with Bonferroni correction).

A series of one-sample t-tests was performed to test whether the numbers of correct responses for all focus conditions were significantly above chance

level. The responses were given to sets of 10 stimuli for focus condition in a forced-choice paradigm, so the chance level was 5 (50% of correct responses). The results of the one-sample t-tests are summarized in Table 7.4.

Table 7.4: Results of one-sample t-tests by group of speaker and focus condition against chance level (=10)

Speaker group	**Focus**	**N**	**mean**	**SD**	**t**	**p**
NS	S	20	14.09	2.11	9.08	**<0.01**
	V	20	17.64	1.29	27.71	**<0.01**
NNS1	S	20	11.27	2.78	2.15	**0.044**
	V	20	15.64	2.79	9.49	**<0.01**
NNS2	S	20	8.18	3.62	-2.36	0.28
	V	20	14.55	2.67	7.99	**<0.01**

The results of the one-sample t-tests show that the numbers of correct responses were significantly above chance level for both focus conditions in NS and NNS1 productions, but not in the ones by NNS2.

7.3.2 Italian listeners

The mean number of correct responses given by the Italian native listeners are presented in Fig. 7.4.

As Fig. 7.4 shows, the Italian listeners gave a fairly high number of correct responses when judging NS and NNS1 productions, while the NNS2 and IT productions are close to chance level.

A series of one-sample t-tests was performed to test whether the responses of each group were significantly above chance level (20). The results of the one-sample t-tests are summarized in Table 7.5.

The results of the one-sample t-tests show that the responses obtained when judging productions by NS and NNS1 were significantly above chance level. In contrast, the number of correct responses provided for NNS2 and

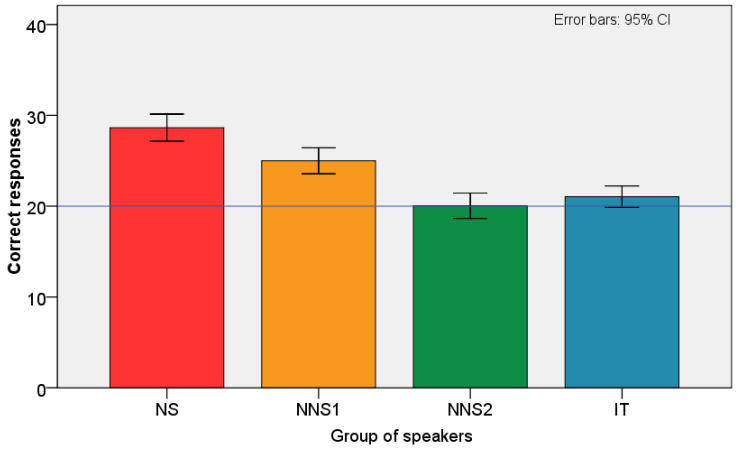

Figure 7.4: Mean number of corrected responses given by Italian listeners by group, averaged sentences.

Table 7.5: Results of one-sample t-tests per group against chance level (=20).

Group	N	mean	SD	t	p
NS	40	28.64	3.37	12.01	**<0.01**
NNS1	40	25	3.22	7.73	**<0.01**
NNS2	40	20.05	3.15	0.07	0.95
IT	40	21.05	2.65	1.85	0.78

IT were not significantly above chance level.

The significance of the results was tested with a one-way Analysis of Variance (ANOVA) with mean number of correct responses as dependent variable and group as fixed factor. The ANOVA showed a significant effect for group on mean number of correct responses ($F(3, 84) = 35.201$, $p<0.01$). Pairwise comparisons between the four different groups showed significant differences in all cases ($p<0.01$, with Bonferroni correction) except between the NNS2 and IT.

The responses by focus condition (S or V) were also analyzed were also analyzed by focus condition (S or V), as summarized by the values reported

in Tab. 7.2 and by data in Fig. 7.5.

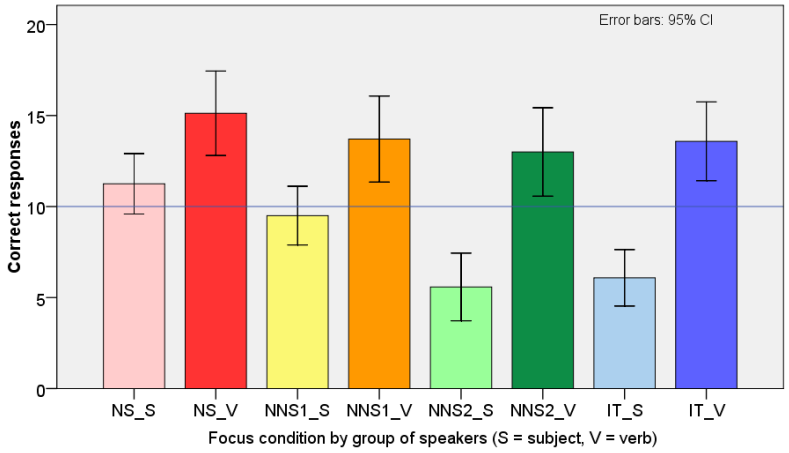

Figure 7.5: Number of correct responses (out of 20) given by the Italian listeners and averaged by group and by focus condition (S = subject in focus; V = verb in focus).

As Fig. 7.5 shows, the data of the Italian listeners replicate the tendency observed for the English listeners: the number of correct responses given for the productions with V in focus was higher than with the one with S in focus. This difference becomes more marked as the competence in L2 decreases. Finally, the data of the Italian listeners when responding to the Italian sentences are similar to the data of the NNS2 productions, showing sizably higher number of correct responses for the sentences with V in focus.

The significance of the results was tested with a one-way Analysis of Variance (ANOVA) with mean number of correct responses as a dependent variable and focus condition as a fixed factor. The ANOVA showed a significant effect for focus condition on mean number of correct responses ($F(7, 168) = 23.162$, $p<0.01$).

Pairwise comparisons within the three different groups showed significant differences in all pairs of sentences with S in focus vs. V in focus (NS_S vs. NS_V, NNS1_S vs. NNS1_V, NNS2_S vs. NNS2_V, IT_S vs. IT_V: for all pairs $p<0.05$, with Bonferroni correction).

The number of correct responses was significantly above chance level for both focus conditions only for the productions of NS. This was confirmed by the results of a one-Sample t-test comparing the results obtained for NNS1_S (M=12.23, SD =2.202) to chance level (=10); t(22, 21)=4.473. In all other cases, the productions with S in focus showed results that were not significantly above chance level.

7.4 Discussion

The results of Experiment 1 show that both English and Italian native listeners could guess well above chance level which were the questions that had originally prompted the sentences spoken by the NS. These results confirm the first hypothesis (H_1), which predicted that native and non-native listeners could correctly identify the information in focus when listening to NS productions, even in absence of the contextual information that is normally present in a conversation.

The second hypothesis (H_2) predicted that English and Italian native listeners would still be able to identify the information in focus in the productions by NNS1, although with worse precision. The results confirmed this hypothesis only for the English listeners. The English listeners were indeed able to recognize focus in the productions by NNS1 well above chance level. However, as predicted by H_2, the number of correct responses was significantly lower than the ones recorded for the productions by NS. Moreover, the number of correct responses obtained for NNS1 was significantly higher than the one obtained for NNS2, which were not significantly above chance level. As for the Italian listeners, they could only identify narrow focus in the productions by NS, while the productions by NNS1 and NNS2 resulted not significantly above chance level. In order to be fully understood, the responses given by both groups of listeners were broken down by focus condition. Both groups of listeners provided a higher number of correct responses

when judging sentences with V in focus as compared to the ones with S in focus. As will be explained in more detail in the General Discussion (Section 9.3.1), the data suggested that in absence of clear prosodic cues that mark word in focus (e.g., higher F_0 and/or longer duration), the listeners opted for the solution where the word in focus was closer to the end of the sentence. This can be caused by the fact that in both English and Italian the 'neutral' broad focus condition is marked with a pitch accent on the rightmost element of the sentence (Ladd, 1996, see Section 2.6). As a consequence, if narrow focus is not clearly marked by prosody or context, the listeners tend to consider the sentence as an instance of broad focus.

The third and last hypothesis (H_3) predicted that the Italian listeners would not be able to identify narrow focus in their L1 productions. The results confirm this hypothesis. The perceptual results are in accordance with the outcome of the acoustic analysis of the sentences in Italian (cf. Section 6.3.4), where no acoustic characterization of narrow focus was found. As observed for the productions in English, the results broken down by focus condition show a bias for V in focus. This shows that, also in Italian, the sentences that are poorly characterized in terms of prosodic focus marking could be interpreted as examples of broad focus.

To conclude, the results of Experiment 1 substantially confirm the results of the production study, by showing that a correct identification of focus is possible only for the productions where prominence is realized with sizable changes in the phonetic cues, especially F_0. While English listeners were also able to detect narrow focus in the productions by NNS1, the Italian listeners could successful detect narrow focus only in the productions by NS. As expected, none of the two groups of listeners could successfully detect focus in the productions by NNS2. Finally, the Italian listeners could not identify focus in the productions in their L1, confirming that the lack of prosodic characterization impedes the identification of narrow focus without extra contextual information.

The analysis of the results broken down by focus condition also shows that when narrow focus is not clearly marked with prosody, the listeners tend to interpret it as an instance of broad focus, both in English and in Italian. The results of the experiments will be discussed in further detail in the General Discussion (Section 9.3.1).

Chapter 8

Experiment 2

8.1 Rationale and hypotheses

The results of the production study showed that F_0 is the acoustic cue that is mainly responsible in the realization of informative narrow focus by NS. In contrast, the results from the two groups of non-native speakers show that the native focus marking strategies are difficult to acquire. The NNS1 show some awareness of the necessity of modulating pitch to signal narrow focus, resulting in an active use of pitch to mark focus location. However, they fail to consistently reproduce the native model, since they mark the first word with a significantly higher pitch, regardless if the word is in focus or not. As for NNS2, the results show no systematic use of pitch or duration as markers of narrow focus location, resulting in undifferentiated productions, heavily characterized by phenomena of transfer from L1 and by a high presence of epenthetic vowels (see Chapter 6). In addition, both NNS1 and NNS2 produce their sentences with a significantly wider pitch span as compared to NS over the whole length of the sentences.

The statistical analysis of the differences in pitch values for NS and NNS1 are discussed in detail in Sections 6.3.1.2 and 6.3.2.2 respectively and they are summarized here in Table 8.1. For the NS, when the S is in focus there

is a significant difference in pitch between the subject (S) and the verb (V). As for sentences with V in focus, the difference in pitch between S and V is smaller and not statistically significant. The NNS1 manage to produce differences in pitch between S and V, although a significant difference in pitch is observed regardless of the focus condition, while in the NS productions the high difference is only noticed when S is in focus condition.

Table 8.1: Mean values of normalized F_0 of the NS and NNS1 speaker groups, averaged by word in focus over sentences and speakers.

Sentences with *subject* (S) in focus

NS		**NNS1**	
	mean norm. F_0 (Hz)		mean norm. F_0 (Hz)
Subject	32.15	Subject	61.98
Verb	19.90	Verb	34.21
F_0 difference	**12.25**	F_0 difference	**27.77**

Sentences with *verb* (V) in focus

NS		**NNS1**	
	mean norm. F_0 (Hz)		mean norm. F_0 (Hz)
Subject	31.86	Subject	61.30
Verb	29.50	Verb	31.27
F_0 difference	**2.36**	F_0 difference	**30.03**

As for the perceptual dimension, the results of Experiment 1 show that native listeners can indeed recognize narrow focus location by prosody alone, both in native and non-native productions, although the numbers of correct responses is significantly higher when judging native productions. The results from the production study and from Experiment 1 were the basis for the design of Experiment 2. The experiment was set up to test the three following hypotheses:

- H_1: English listeners will be able to better identify narrow focus lo-

cation when judging productions by NS as compared to non-native productions, in accordance with the results of Experiment 1;

- H_2: Listeners' ability to detect focus location will be boosted when judging sentences produced by NS presenting the differences in pitch found in the productions by NS; conversely, their ability will be hindered if the sentences uttered by NS present the pitch difference realized by NNS;

- H_3: Listeners' ability to recognize focus location will be hindered when judging sentences produced by NNS presenting the difference in pitch found in the productions by NNS; it is expected that this ability will improve when judging productions by NNS presenting the pitch differences realized by NS.

8.2 Methodology

8.2.1 Stimuli

The stimuli created for this experiment were based on a subset of the sentences analyzed in the production study. The productions of two speakers were considered: one male native speaker and one female non-native speaker. The non-native speaker was chosen from the NNS1 group. Speakers from NNS2 were excluded, based on the results of the production study and Experiment 1, which had both shown that NNS2 were not able to successfully differentiate the location of narrow focus by using prosodic cues (pitch or duration). The selected productions consisted in 10 sentences per speaker, equally distributed in two focus conditions: 5 with S in focus and 5 with V in focus. The resulting number of sentences was therefore 20 (5 sentences x 2 focus conditions x 2 speakers). The selected set of sentences was digitally manipulated using Praat. The normalized F_0 values corresponding to

the pitch peak on the words in focus were manipulated locally for each sentence in order to obtain two opposite experimental conditions, together with a third intermediate condition. The resulting set of stimuli included:

1. Productions where the difference in pitch between S and V was set to the average difference in F_0 calculated for the group which they belonged to. In other words, this manipulation resulted in a match between sentences and group: sentences produced by NS were matched with the NS average F_0 differences and sentences produced by NNS were matched with the NNS average F_0 difference;

2. Stimuli where the difference in F_0 between S and V was set to the average difference calculated for the group which they did not belong to. In other words, this manipulation resulted in a mismatch between sentences and group: sentences produced by NS were modified with the NNS pitch differences and sentences produced by NNS were modified with the NS pitch difference);

3. Stimuli where the difference in pitch span between S and V was set to the values of the F_0 difference standing between NS and NNS. This intermediate step was determined by locating a value that was at midpoint in the difference between the F_0 values of NNS and NS for the two focus conditions.

The six experimental conditions are described in Tab. 8.2, together with the corresponding number of stimuli.

The visual Manipulation Editor of *Praat* was used to modify pitch by manually raising or lowering the F_0 values in accordance with the calculations summarized in Tab. 8.3.

Table 8.2: Summary of the six experimental conditions of Experiment 2, with description and number of stimuli.

Condition	Description	Number of stimuli
NS_ a	NS sentences with NS F_0 difference	10
NS_ b	NS sentences with the intermediate value between NNS and NS F_0 differences	10
NS_ c	NS sentences with NNS F_0 difference	10
NNS_ a	NNS sentences with NNS F_0 difference	10
NNS_ b	NS sentences with the intermediate value between NNS and NS F_0 differences	10
NNS_ c	NNS sentences with NS F_0 difference	10

8.2.2 Subjects

The participants were 20 British English speakers. Their average age was 23,5 years, and they had varied professional backgrounds. None of the listeners had reported any hearing impairments or familiarity with Italian.

8.2.3 Task and procedure

The experiment was presented using *LimeSurvey* (Schmitz, 2012) on a laptop personal computer connected to a headset. The experiment was performed in a silent environment at the University of York Library (UK).

The task was the same used in Experiment 1. The recognition of focus location was prompted by asking the participants the question: 'when you listen to an answer out of its context, can you correctly guess the question that triggered that answer?' As in Experiment 1, the subjects' task was to listen to the responses presented individually and to select the question that was more likely to have triggered the answer. The listeners expressed their choice by choosing the most appropriate response out of two options, each corresponding to one focus condition (S or V in focus). Each stimulus was

Table 8.3: Determination of intermediate steps in the differences in F_0 between NNS and NS. Values approximated to the closest integers.

***Subject* (S) in focus**	
	mean norm. F_0 (Hz)
NNS F_0 difference	28
NS F_0 difference	12
NNS F_0 - NS F_0	16
Step = (NNS F_0 - NS F_0) / 2	8
(NNS F_0 - NS F_0) + step	20

***Verb* (V) in focus**	
	mean norm. F_0 (Hz)
NNS F_0 difference	30
NS F_0 difference	2
NNS F_0 - NS F_0	28
Step = (NNS F_0 - NS F_0) / 2	14
(NNS F_0 - NS F_0) + step	16

played automatically once, although the subjects were given the possibility to listen to the sentences again by using a button in the graphic user interface to replay the audio files. The instructions that were provided to the listeners are reported in Appendix C.

In order to reduce the risk of introducing the possible bias caused by memory effects, it was decided to precede every item with a short beeping sound (100 ms) followed by 500 ms of silence. The beeping sound was generated as a pure tone and attached to the files by running a *Praat* script written by the author.

At the end of the experiment, the subjects could see their results in the

form of a feedback message reporting the number of correct responses.

8.3 Results

The results of Experiment 2 are summarized in Tab. 8.4 and Tab. 8.5.

Table 8.4: Total number, mean and standard deviations of correct responses, averaged by experimental condition over speakers and sentences.

Number of correct responses			
Condition	**N**	**mean**	**SD**
NS_ a	10	8.80	1.24
NS_ b	10	8.15	1.35
NS_ c	10	7.25	1.55
NNS_ a	10	5.90	1.21
NNS_ b	10	5.75	1.45
NNS_ c	10	5.85	1.14

Tab. 8.4 shows the mean number of correct responses given by the English native listeners, along with standard deviation, divided by the six experimental conditions.

Tab. 8.5 shows the mean number of correct responses given by the listeners along with standard deviation, divided by experimental condition and by focus (S in focus or V in focus).

The bar chart in Fig. 8.1 shows that the listeners can correctly identify narrow focus location in all conditions, while the responses given for all productions by NNS are close to chance level.

As for the differences between the six experimental conditions, the responses given to NS show a clear ranking between conditions, with the highest number of correct responses for condition NS_a, a slightly lower number for condition NS_b and the lowest number for condition NS_c. In contrast,

Table 8.5: Total number, mean and standard deviations of correct responses, averaged by experimental condition and by focus over speakers and sentences.

Number of correct responses				
Condition	**Focus**	**N**	**mean**	**SD**
NS_ a	S	5	4.20	1.15
	V	5	4.60	0.68
NS_ b	S	5	4.10	1.02
	V	5	4.05	0.89
NS_ c	S	5	4.10	1.21
	V	5	3.15	1.27
NNS_ a	S	5	2.20	1.28
	V	5	3.70	1.17
NNS_ b	S	5	2.25	1.21
	V	5	3.50	1.15
NNS_ c	S	5	1.85	1.27
	V	5	4.00	0.80

the responses given to NNS do not present sizable trends differentiating the three experimental conditions.

The mean numbers of correct responses were analyzed by conducting a one-way Analysis of Variance (ANOVA) with mean number of correct responses as dependent variable and group as fixed factor. The ANOVA showed a significant effect for condition on mean number of correct responses ($F(5, 114) = 19.690$, $p<0.01$). Pairwise comparisons between the six different groups showed that there is a significant difference between NS_a and NS_c ($p<0.01$, with Bonferroni correction); in contrast, the results achieved in the intermediate condition NS_b do not differ significantly from conditions NS_a and NS_c. As for the non-native productions, the pairwise comparisons between the results obtained in the three different conditions showed no significant differences between NNS_a, NNS_b and NNS_c.

The results were also broken down by focus condition (S or V in focus).

The values reported in Tab. 8.4 and plotted in Fig. 8.2 show that the numbers of correct responses for NS native productions with S in focus are almost constant, while the ones with V in focus show a downward trend from condition NS_a to condition NS_c. As for the productions by the NNS, sentences with V in focus show a systematically higher number of correct responses as compared to sentences with S in focus in all conditions.

A series of one-sample t-tests was performed to test whether the numbers of correct responses for all focus conditions were significantly above chance level. The responses were given to sets of 5 stimuli for focus condition in a forced-choice paradigm, so the chance level was 2.5 (50% of correct responses). The results of the one-sample t-tests are summarized in Table 8.6.

Table 8.6: Results of one-sample t-tests for each focus condition against chance level (=2.5).

Condition	**Focus**	**N**	**mean**	**SD**	**t**	**p**
NS_a	S	5	4.20	1.15	6.60	**<0.01**
	V	5	4.60	0.68	13.80	**<0.01**
NS_b	S	5	4.10	1.02	7.01	**<0.01**
	V	5	4.05	0.89	7.82	**<0.01**
NS_c	S	5	4.10	1.21	5.92	**<0.01**
	V	5	3.15	1.27	2.29	**<0.01**
NNS_a	S	5	2.20	1.28	-1.05	**0.033**
	V	5	3.70	1.17	4.57	0.308
NNS_b	S	5	2.25	1.21	-0.93	**<0.01**
	V	5	3.50	1.15	3.90	0.367
NNS_c	S	5	1.85	1.27	-2.29	**0.03**
	V	5	4.00	0.80	8.44	**<0.01**

The results of the one-sample t-tests show that the numbers of correct responses were significantly above chance level for all NS conditions. However, in the case of NNS, in none of the experimental conditions the numbers of correct responses were above chance level for both focus conditions. The

p value for NNS_c in Tab. 8.6 must not be considered as a proof of significance: the mean number of correct responses is significantly below chance level, as shown by the negative value of t and by Fig. 8.2.

The mean numbers of correct responses were analyzed by conducting a one-way Analysis of Variance (ANOVA) with mean number of correct responses as dependent variable and focus condition as fixed factor. The ANOVA showed a significant effect for condition on number of correct responses ($F(11, 228) = 13.486$, $p<0.01$). Pairwise comparisons within the NS productions showed that there is a significant difference between NS_a_V and NS_c_V ($p=0.03$, with Bonferroni correction); in contrast, the results achieved in the intermediate condition NS_b_V do not differ significantly from conditions NS_a_V and NS_c_V.

8.4 Discussion

The first hypothesis tested in this experiment predicted that the native listeners could detect focus more efficiently in native productions than in non-native productions, regardless of the way the stimuli had been manipulated (H_1). The results confirm this hypothesis, as the number of correct responses given by the listeners when judging NS productions were significantly higher than the ones given when listening to NNS productions.

The second hypothesis predicted that the native listeners' ability to detect narrow focus would be enhanced when judging NS sentences realized with NS F_0 difference between S and V. In contrast, narrow focus would be more difficult to identify in NS sentences presenting NNS F_0 difference between S and V (H_2). The results confirm also this hypothesis, showing that a match between the native status of the sentences and the native differences in F_0 indeed facilitated the listeners. The listeners achieved significantly higher numbers of correct responses when judging all native stimuli as compared to native stimuli with non-native F_0 values.

The third hypothesis predicted that native listeners' ability to identify narrow focus would be enhanced when judging NNS sentences realized with NS F_0 difference between S and V. In contrast, narrow focus would be more difficult to identify in NNS sentences presenting a matching NNS F_0 difference between S and V (H_3). The results did not confirm this hypothesis: in this case the differences between the numbers of correct responses given under the different conditions were not significant. Moreover, the analysis of the results, broken down by focus condition, showed that the listeners could not identify narrow focus above chance level in any of the three non-native conditions. It could therefore be concluded that the detection of narrow focus was not successful for non-native productions.

The lack of significant results in the NNS productions can be explained by considering the sentences at a global level: the significantly wider pitch span observed for the NNS productions could have masked the small differences in F_0 that were introduced with the signal manipulation, thus reducing their perceptual impact. The differences were still easy to perceive in NS productions, which were characterized by a narrow pitch span. However, the identification was difficult when dealing with NNS productions, where the fine-grained differences in F_0 could have been lost in the wider pitch span of their utterances.

As in Experiment 1, the analysis of the results broken down by focus condition provided interesting findings. In the productions by NS, where the native sentences were matched with the NS differences in F_0, the number of correct responses for V in focus was higher. Conversely, when the sentences were modified with the NNS F_0 difference, the number of correct responses for V in focus resulted significantly lower as compared to S in focus. This outcome can be explained by observing the difference in F_0 realized by the NNS. The productions by the NNS presented the same difference in F_0 between S and V in both focus conditions (see Tab. 8.2), while NS realized this difference only when S was in focus. For this reason, NS tended to identify

the differences in pitch between S and V in the NNS productions as cues for focus on the sentence subject. As a consequence, the tendency to assign focus on the rightmost constituent in the sentence was neutralized.

As observed in the results of Experiment 1 (see Section 7.4), the preference for V in focus in the NNS productions was probably caused by the lack of a proper prosodic characterization of narrow focus. As a consequence, the intended realizations of narrow foci seemed to be mistaken for examples of broad focus. This interesting possibility will be discussed further in the General Discussion (see Section 9.3.2).

To conclude, the results of Experiment 2 suggest that pitch differences play an important role in the detection of narrow focus, especially in native productions. As for non-native productions, the manipulation of the signal and the global differences in pitch span seem to have neutralized any sizable impact of pitch differences on focus detection. A more detailed discussion of the results of the experiment will follow in Section 9.3.2.

Part IV

Interpreting the results

Chapter 9

General Discussion

9.1 Introduction

This chapter will discuss the results of the production and perception studies, outlined in the relevant sections of Chapters 6, 7 and 8. After a brief summary of the methodology used in the production study (Section 9.2), the discussion will tackle the data from production, starting from the results of the acoustic and statistical analyses at sentence level (Section 9.2.1). The discussion will then deal with the results of the word-level analysis (Section 9.2.2).

The results of the perception study will be discussed in Section 9.3, which will be divided into two subsections that will discuss the results of Experiment 1 (9.3.1) and Experiment 2 (9.3.2). Each section will be preceded by a short summary of the methodology used in the respective perception experiment. Finally, Section 9.4 will discuss the relation between the results of the production and the perception study.

9.2 Production study

This section and the respective subsections will be dedicated to the full-scale discussion of the results of the production study (see Chapters 6 and 7). The

study was aimed to analyze a set of short sentences spoken by a group of four native British English speakers (NS) and two groups of Italian speakers of English L2, composed by four speakers each: one group of Italian native speakers with a higher competence in English L2 (NNS1) and one group of Italian native speakers with a lower competence in English L2 (NNS2). A total of 120 sentences in English (40 sentences x 3 groups) were recorded for this study using an elicitation protocol that was aimed to prompt the prosodic marking of narrow focus on sentence subjects (S) or on the verb (V). An extra set of similar sentences in Italian was also elicited from the Italian speakers. All sentences were segmented and annotated using *Praat* (Boersma & Weenink, 2013). The program was also used to acoustically analyze the productions at sentence and word level. The acoustic analysis was based on the measurement of duration, speaking rate and pitch range for the sentence-level analysis; and of duration and normalized F_0 for the word-analysis. The following sections will discuss the results of the two levels of analysis.

9.2.1 Sentence-level analysis

The results of the acoustic analysis at sentence level successfully confirmed the hypothesis that NNS1 would tune their productions towards the native model as a function of their higher proficiency in L2. This process of progressive tuning to the prosodic system of English was clearly visible by observing all three acoustical measurements that were considered at sentence level, namely duration, speaking rate and pitch span.

The sentences produced by NS resulted shorter than the ones produced by both groups of Italians. The fact that the mean duration values by NNS1 speakers were significantly lower than the ones by NNS2 can be seen as a reliable indicator of a progression towards a more native-like prosody. The longer duration measured in the productions by the Italian speakers possibly reflect the structural differences between the rhythmic structures of the

two languages involved. As shown in Section 2.6, the English and Italian respectively occupy places near the two extremes in the continuum between syllable-timed and stress-timed languages (Dauer, 1983). Moreover, as already shown in the literature (see Busà, 1995; Flege et al., 1999), the English spoken by Italians is often characterized by the lack of vowel reduction and by the addition of epenthetic vowels. In the data presented in this study, these two phenomena certainly contributed to the longer duration observed in the productions by NNS2.

The results obtained for duration was mirrored by the speaking rate values, with the difference that the relation between the three groups was symmetrically reversed. As expected, NS have the highest speaking rate, followed by NNS1 and NNS2. Again, the statistically significant differences between NNS1 and NNS2 show that a convergence towards the native model is in progress. As expected, the productions by NNS2 of English L2 are characterized by the lowest speaking rate. This is in line with the findings reported in the literature on the perception of foreign accent, where speaking rate has been considered a reliable indicator of limited L2 proficiency. For example, it has been suggested that "L2 speech is typically delivered more slowly" (Munro et al., 2010: 627) as compared to L1 speech. Moreover, lower speaking rate values have been related to a high degree of perceived foreign accent (cf. Trofimovich & Baker, 2006) and it has been shown that a slower speaking rate also results in a smaller amount of information conveyed (Hincks, 2010).

The results for pitch span present an interesting difference between native and non-native speakers. The productions by NS are characterized by a significantly narrower pitch span when compared to the productions of both non-native groups. As for the two groups of non-native speakers, the differences between NNS1 and NNS2 are not significant, although NNS1 still show a tendency towards the native values. These results are in contrast with findings reported in the literature regarding the comparison of pitch

span between native and non-native speakers of English. In this regard, Hincks (2004), Ramírez Verdugo (2006) and Mennen (2007, Mennen et al., 2012) have claimed that non-native productions of English are characterized by a narrower pitch span when compared to the values expected for the target language. In contrast, the data collected in this study show an opposite trend: the non-native productions are characterized by a significantly wider pitch span as compared to the native ones. The results are also in contrast with the empirical data collected in recent studies comparing the productions of Italian speakers of English L2 and the productions by American English NS (Busà & Urbani, 2011; Urbani, 2013). In these studies, the productions by non-native speakers have a narrower pitch span when compared to the native productions. However, preliminary results presented in Stella & Busà (in press) suggest that speakers of British English do have a narrower pitch span when compared to Italian speakers of English L2. This difference might be the result of a sloppy control over pitch span by non-native speakers, as compared to the tight control over pitch span characterizing the native productions. In the case of NNS2, by inspecting spectrograms and F_0 contours, it was found that the presence of epenthetic vowels also affects the overall pitch span. As shown in Section 6.4, epenthetic vowels are often pronounced with an erratic rise in F_0 which makes them stand up as compared to the rest of the utterances.

However, wider pitch span is not an exclusive prerogative of NNS2, but it characterizes the productions of both levels, which present similarly high values of pitch span when compared to the native productions. The analysis of the Italian L1 data set showed that the Italians' pitch span is significantly wider also in their L1, as compared to the one observed in the productions of the English NS. These relatively high values of pitch range could also be originated from the characteristic of the regional variety of Italian considered in this study, which is the same that was analyzed in Stella & Busà (in press). In this regard, it would be interesting to investigate in more detail

the differences in production and perception of narrow focus location by speakers coming from different regional areas of Italy (see Chapter 10).

9.2.2 Word-level analysis

The results of the acoustic measurements performed at word level were used to verify if NS could mark narrow focus with the use of prosodic cues. The results show that duration does not seem to play an active role. In contrast, words in focus are indeed affected by modifications in pitch. When in focus, S are produced with a significantly wider F_0 when compared to V. In contrast, when V is in focus, the difference between S and V becomes smaller and not statistically significant. These results are in line with what found in the previous literature, where it was shown that pitch is the most reliable phonetic cue in focus marking in English (cf. Büring, 2005, see Section 2.5), both in terms of the presence of pitch peak on the focused constituents and in terms of *pitch obtrusion* (Cruttenden, 1997). This latter concept has been defined as "the step up or down in pitch immediately following the focused constituent" (Ramírez Verdugo, 2006:11) and such a "step down" after the word in focus is exactly what could be observed in the production by NS to mark S in focus. This drop in F_0 following focus material was also reported in Xu & Xu (2005, see Section 2.5.2).

The NNS1 data suggests that a process of progressive tuning to the native model is in action. NNS1 present systematic differences in F_0 between S and V, suggesting that the speakers have apparently learnt to activate pitch differences to mark narrow focus. However, the results show that NNS1 have not yet achieved mastery in focus marking. Indeed, the differences do not reflect the focus condition of the words, but are determined by the position of the words in the sentence: S is always produced with a higher F_0 as compared to V, regardless of the focus condition (S in focus or V in focus). This could be seen as empirical evidence for the difficulty of acquiring such a fine-grained phonetic implementation even for experienced speakers of English L2. NNS1

might have been aware of the need for marking focus with pitch modulation, but they could not correctly use because of the influence of their L1, where narrow focus is more often marked with syntax and word order than by means of prosody (see Ladd, 1996 and Face & D'Imperio, 2005, discussed in Section 2.6).

NNS2 experienced serious problems in differentiating focus by prosodic means. In particular, the results of the acoustic analysis did not show any emerging systematic pattern, rather suggesting an erratic, or random, prosodic behavior. This inconsistency in focus marking confirms the expectation that NNS2 would not be able to signal prominence by prosodic means. The results observed for NNS2 reflect the findings reported in Busà (1995) for the acquisition of English vowels by Italian speakers with a lower competence in L2. This analogy suggests that the difficulties in the acquisition of L2 prosody go in parallel with the ones in L2 segments acquisition.

In general, both NS show very fine-grained differences in F_0 to mark narrow focus location. The small range of these differences is probably due to the nature of the phenomenon studied. Ladd (1996) reported that narrow contrastive focus is by definition produced with more emphasis. As a consequence, the phonetic characterization of its informative, non-contrastive, counterpart is expected to be more elusive, resulting in smaller changes in the phonetic cues as compared to contrastive narrow focus. Interestingly, the majority of empirical studies dealing with the phonetic realization of narrow focus are based on the differences between narrow and broad focus, but not on the different realization of the two types of narrow focus, namely contrastive vs. non-contrastive. Therefore, the results presented in this study seem to provide empirical evidence that could justify the theoretical distinction between the two types of narrow focus.

In order to have a complete vision of the phenomenon of the prosodic marking of narrow focus and to verify the existence of effects of prosodic transfer from L1 to L2, the Italian L1 data set was also analyzed. As for

duration, the results showed that in their L1 the Italian speakers produce significant differences in duration between S and V, but these differences depend on the position in the sentence, regardless of the narrow focus location. This suggests that in Italian duration does not play a role in narrow focus marking. This result ss particularly interesting, as in Italian duration is the main prosodic cue involved in the realization of prominence at word level, that is, in the realization of word stress (Bertinetto, 1981; Magno Caldognetto et al., 1983). It seems therefore that in Italian duration is not involved in the marking of narrow non-contrastive focus. As for F_0, the results do not show any sizable trend, excluding an active role of fundamental frequency in the phonetic realization of narrow focus in Italian. This result is also in contrast with previous literature on the realization of narrow contrastive focus, where F_0 was identified as the main acoustic correlate for narrow contrastive focus in Italian (Magno Caldognetto & Fava, 1974; Kori & Farnetani, 1983). To conclude, the data presented in this study suggest that in Italian narrow non-contrastive focus is not prosodically marked. This outcome is in line with the definition of Italian as a non-plastic language (Vallduví, 1991), that is, a language that relies more on syntax and word order strategies rather than on prosody in marking prominence at sentence level.

9.2.3 Epenthetic vowels

Although an extensive analysis of epenthesis is beyond the scope of this thesis, it is important to note its impact on the productions by NNS2. For its nature, vowel epenthesis has been traditionally treated as a segmental phenomenon (Repetti, 2012), although it certainly affects the prosodic domain too. The impact of epenthesis on the temporal organization of the productions by NNS2 is evident: adding a vowel results in the creation of new syllables, consequently prolonging duration and changing the overall rhythm of sentences (cf. Section 9.2.1). In addition, the data analyzed in this study show that the impact of epenthesis on prosody is not limited to the temporal

aspects, but that it also influences the overall pitch of the productions. It was already mentioned that epenthetic vowels were often pronounced with a stray rising tone (cf. Sections 6.4 and 9.2.1).

In the production data analyzed in this study it was found that F_0 peaks were particularly evident when the epenthetic vowel was at the boundary of an intonational unit. These rises seem to correspond to the suspended tones that are normally used for lists or to signal continuation in a speech turn in English (Wells, 2006). This suggests that epenthetic vowels can be considered at the borderline between actual vowels and filled pauses (such as *hum* or *err*). Besides, this combined used of epenthesis and rises in pitch also suggests that NNS2 fail to produce the sentence as a single intonation unit and that they have to break the single intonation phrase composing the sentences into smaller, more manageable, intermediate phrases. The limited ability to correctly parse information in a single intonation phrase and the consequent tendency to divide the intonational structure into smaller units has been documented for Japanese and Korean speakers of English L2 by Ueyama & Jun (1998). The productions by Italian speakers of English L2 could also be characterized by this behavior. Further research based on empirical data is needed to shed more light on this possibility.

To conclude, the data presented in this study suggest that epenthesis should not be treated as an only segmental phenomenon, but that it should instead be considered as a two-fold interface phenomenon, between the segmental and suprasegmental levels, and between the two fluency-based (speech rate, duration of pauses) and melody-based (stress timing, pitch) dimensions of L2 prosody (Trofimovich & Baker, 2006).

9.3 Perception study

The perception study was composed by two experiments. The methodology used in experiments 1 and 2 was described in detail in Chapters 7 and 8,

respectively. This section will present general comments on the common features of the two experiments. The results of the single experiments, along with relevant comments, will be discussed in more detail in Section 9.3.1 (Experiment 1) and Section 9.3.2 (Experiment 2).

In both experiments, the task of identifying narrow focus consisted of a task where the listeners were asked to guess the question that had originated the sentence as an answer in a two-alternative forced choice. This procedure was devised in order to present the listeners with a straightforward task that could elicit their "metalinguistic judgments" (Gili Fivela, 2012: 20, see Section 3.4.3) without the need for too technical instructions and training. The robustness of the results seems to confirm the efficiency of this experimental paradigm. The informal feedback received from the participants after the experiment also hinted at its success in catering the subjects with a stress-free and at the same time thought-provoking experience.

Another common feature of the experiments was the choice not to use heavily manipulated or substantially resynthesized stimuli for the study of focus marking. Considering also the inconclusive results found in the pilot studies documented in Chapter 4 (see Sections 4.3.3 and 4.5.3), it was decided to use original speech (Experiment 1) or speech where only a part of a manipulated F_0 contour (Experiment 2). This decision was also based on the indication that "using synthetic speech stimuli may be inappropriate for studying the perception of focus in everyday speech" (Vaissière, 2005: 242). This choice had the twofold purpose of reducing frustration and to present the listeners with more natural (and, therefore, realistic) stimuli.

This section has presented general comments regarding both perception experiments and the experimental procedures that were used. The following sections will discuss in detail the results of the two individual experiments.

9.3.1 Experiment 1

The purpose of the first experiment was to test the perception of narrow focus on the basis of the prosodic cues used in prominence marking by native and non-native speakers of English. Based on the results of the production study, it was expected that the listeners could successfully identify narrow focus in the productions by NS and NNS1, since these were the two groups of speakers that were capable of marking focus with prosodic cues (in particular, with pitch). On the other hand, it was expected that the listeners could not identify narrow focus in the productions by NNS2, as this group of speakers did not show any active use of prosodic cues in focus marking. The experiment was presented to two groups of listeners: a group of English native listeners and a group of Italian native listeners. It was expected that the sensitivity to the prosodic marking of narrow focus would be higher for English native speakers than for Italian ones.

As for the experimental procedure, the experiment presented the participants with the complete set of the 120 original, non-manipulated sentences produced by the three groups of speakers considered in the production study (NS, NNS1 and NNS2). The participants were asked to listen to a sentence and to guess the question that had prompted the sentence as an answer, choosing one of the two options presented in a two-alternative forced choice. The Italian listeners were also asked to respond to an extra set of 40 sentences in Italian by performing the same experimental task.

The results of Experiment 1 show that English native listeners can successfully identify the questions that originally prompted the sentences for the productions by NS and by NNS1. This outcome confirms the hypothesis that, when listening to NS productions, English listeners can correctly identify the information in focus only by attending to prosodic cues. This means that the acoustic cues in the productions by the two groups are enough to recognize narrow focus location even in absence of the contextual information that is normally present in a conversation. As for NNS2, the listeners

could not successfully identify narrow focus. The analysis of the results by focus condition showed that the poorly characterized realizations of narrow focus by NNS2 were often mistaken for instances of broad focus. This will be explained in detail in the next paragraphs of this section.

As expected, the comparison between the results obtained for each group show that English listeners can identify focus in the productions by NS with a significantly higher accuracy than when responding to the productions by NNS1. This shows that the productions by the non-native speakers could still be understood by English native listeners, but with more difficulty as compared to those by the NS. Moreover, the fact that the productions by NNS1 could still be understood reflects the trends found in the production study, where it was shown that NNS1 are able to activate pitch differences in the direction of the native model (cf. Section 6.3.2). Conversely, the productions by NNS2 failed to be understood, confirming the results of the production study, which show that NNS2 are not able to differentiate narrow focus information by using prosody (cf. Section 6.3.2). Beside the lack of prosodic characterization, other factors that might have hindered the identification of narrow focus in the NNS2 include the frequent occurrence of epenthetic vowels, the significantly wider pitch span and the slower speaking rate.

As expected, the Italian listeners' ability to identify narrow focus is not as good as the English listeners': the analysis of the results by focus condition showed that the Italian listeners were only able to successfully recognize narrow focus in the productions by NS. The results of the perception experiment therefore suggest that the sensitivity to narrow focus is lower for non-native speakers (see Section 9.4).

The Italian listeners were also asked to identify narrow focus in the Italian L1 data set. As for the stimuli in Italian, the Italian listeners also failed to recognize focus location. This is in line with the results of the production study, where duration and F_0 did not seem to play an active role in focus marking in the productions by Italian L1 speakers.

The analysis of the results of the experiment broken down by focus condition shows that both groups of listeners give a significantly higher number of correct responses when judging sentences with V in focus as compared to the ones with S in focus. This outcome might be explained by considering that for both English and Italian the broad focus condition is characterized by the location of focus on the rightmost element of the sentence (Ladd, 1996, Wells, 2006; Gagliardi et al., 2012, see Sections 2.3 and 2.6). In Experiment 1, the forced choice was between S in focus and V in focus. If one considers that the subject is invariably located at the beginning of the sentences, therefore in the leftmost position, it seems that, in absence of evident changes in prosody, the listeners preferred to choose the option where focus was marked on the rightmost constituent of the sentence (in the case of the options available in Experiment 1, the verb).

To conclude, the results of Experiment 1 confirm the hypotheses that were based on the results of the production data. First, it shows that both English and Italian listeners could successfully identify narrow focus in the productions by NS. Second, English listeners were still able to recognize focus in the productions by NNS1, but they could not detect focus in the productions by NNS2. Italian listeners, instead, could successfully identify focus only in the productions by NS. Thus, the analysis by focus condition provides evidence for a deeper understanding of the dynamics involved in the perception of narrow and broad focus in both English and Italian.

9.3.2 Experiment 2

The purpose of the second perception experiment was to determine the impact of the correct pitch modulation in the detection of narrow focus by English native listeners. Based on the results of the production study, it was expected that use of pitch (the perceptual correlate of F_0) would be crucial in the detection of narrow focus in absence of any extra contextual information. Moreover, the results of Experiment 1 had shown that English native listen-

ers could successfully recognize narrow focus in the productions by NS and NNS1, who were the two groups of speakers that were capable of marking focus with the modulation of F_0 differences between S and V. The experiment was therefore aimed at determining if the correct implementation of these differences in F_0 would be enough to successfully perceive narrow focus. In this experiment the productions by NNS2 were not considered, so the native and non-native status of the speakers used in the stimuli was referred to as NS and NNS, respectively. A subset of the sentences collected in the production study was acoustically modified with *Praat*. The differences in F_0 between S and V were manipulated so that in each sentence the F_0 difference would correspond to the average values found in the production study for native or non-native speakers. The six experimental conditions obtained with the acoustic manipulation, together with the calculations and the methodology used to generate the corresponding stimuli are presented in detail in Section 8.2.1. It was expected that the listeners could identify focus with higher accuracy when dealing with sentences where the native status was matched with native F_0 values than when judging sentences with a mismatch between native and non-native F_0 values. On the other hand, non-native sentences presenting native F_0 differences between S and V should be understood with more success than the ones where the non-native status was matched with non-native F_0 differences.

The experiment was based on the same paradigm used in Experiment 1, that is, a two-alternative forced choice between two questions that could have triggered the sentence as an answer: one with S in focus and the other with V in focus. The results of Experiment 2 confirm that the native listeners are more successful in identifying narrow focus in the native productions than in the non-native ones. The participants gave a significantly higher number of correct responses when listening to NS productions as compared to NNS productions. As for the latter, the analysis of the results by focus condition confirms that the listeners are not able to recognize focus above chance level.

It was expected that the native listeners' ability in recognizing narrow focus would be facilitated when judging NS sentences realized with the native F_0 difference between S and V as compared to NS sentences realized with non-native F_0 difference between S and V. The results of the experiment confirm this expectation, showing fewer correct responses in the condition where native status and F_0 differences between S and V were matched than in the condition where the native status was modified with non-native F_0 differences.

However, the results of the NNS productions did not show any significant difference between the single experimental conditions. Moreover, as mentioned above, none of the NNS conditions reached significance above chance level, showing that the listeners could not successfully identify narrow focus in neither of the NNS conditions regardless of a match or mismatch between the non-native status and the differences in F_0. The lack of significant results in the NNS productions can be explained by considering the sentences at a global level: the significantly wider pitch span observed for NNS productions could have masked the small differences in F_0 inserted with the signal manipulation, thus reducing the perceptual impact of these differences. While the differences were still easy to perceive in NS productions, which were characterized by a narrow pitch span, the identification was difficult when dealing with NNS productions, where the fine-grained differences in F_0 could have been lost in the wider pitch range.

The literature on the so-called *just noticeable differences* (JND), or the "differential threshold of pitch change" (t'Hart & Collier, 1990: 33), has attempted to define the smallest changes in F_0 that can be perceived by a listener with conflicting results. It has been suggested that differences as small as 2 Hz are enough to perceive a categorical change in the perception of speech (Klatt, 1973), although most of the data come from experiments done with synthetic speech. As for natural speech, the literature has provided a variety of possible values, which seem to be influenced by the interaction

of a number of parameters (such as speaking rate or musical training, cf. Quené, 2007 and Marotta et al., 2012). When observing the values used in Experiment 2 (see Tab. 8.3), it is reasonable to think that such small differences could have been lost when implemented in the productions by NNS, characterized by sizably higher pitch span values. By contrast, the same fine-grained differences could have been easier to detect in the NS productions, characterized by a very narrow pitch span. Auditory impressions seem to confirm this idea: by listening to the NS productions, differences between the single experimental conditions can be clearly heard. In contrast, by listening to NNS productions differences between conditions are difficult to perceive.

More evidence of the effect of the lack of proper prosodic characterization of narrow focus in the listeners' perception can be found in the analysis of the results broken down by focus condition. As for NNS, the listeners replicated the results observed in Experiment 1: the number of correct answers was significantly higher for the sentences with V in focus. This suggests again that, when in absence of a clear prosodic characterization of narrow focus, the listeners tend to select the constituent that is closer to the right periphery of the sentence (in the case of the experiments, the verb). These results suggested that the productions by NNS, as modified in Experiment 2, were not enough prosodically characterized to allow narrow focus identification.

In contrast, the analysis of the NS productions by focus condition shows significant differences in the results in the different conditions. The sentences with S in focus received an about equal number of correct resposes across all conditions, while the number of correct answers for sentences with V in focus changed significantly depending on the experimental condition. When the NS sentences were matched with the NS differences in F_0, the number of correct responses for V in focus was higher, while when the sentences were modified with the NNS F_0 difference, the correct responses for V in focus were significantly lower than the ones for S in focus. Therefore this

higher number of correct responses for S in focus for the sentences with NNS F_0 values seemed to override the tendency to prefer V in focus that was found in the results of both experiments. This outcome can be explained by observing the difference in F_0 realized by NNS.

As observed in Section 6.3.2, NNS1 (the group of speakers considered in Experiment 2 as NNS) manages to produce differences in pitch that are not present in Italian, showing that a partial attunement to the native model is in progress. However, this attunement is not achieved completely; the productions by NNS1 present the same difference in pitch from S to V in both focus conditions (see Section 6.3.2.2), whereas NS realize this difference only when S is in focus (see Section 6.3.1.2). Therefore, it is not surprising to see that the default preference for focus location on the verb is neutralized by the presence of differences in pitch that are identified by NS as characteristic cues for focus on the sentence subject (i.e., a sizable F_0 difference between S and V).

In sum, the results of Experiment 2 suggest that pitch differences have an important role in detecting narrow focus location. This was shown by the results obtained for the productions by NS, where an incorrect implementation of F_0 changes the perception of narrow focus location. As for the productions by NNS, the global characteristics of pitch span, which is significantly wider as compared to the productions by NNS, seem to have neutralized any sizable impact of the fine-grained F_0 differences on focus detection.

To conclude, the productions by NS with NNS differences in F_0 show that an incorrect implementation of F_0 might result in the misunderstanding of the intended focus. Future research should be carried out with the aim of studying the effects of this kind of misunderstanding in the communication between native and non-native speakers.

9.4 Relation between production and perception

The relation between speech production and perception is not fully understood and it has been argued that "the closeness of the fit between the activities of speaking and perceiving speech has not been frequently addressed" (Fowler & Galantucci, 2005: 633). However, the study of both dimensions of speech is necessary to have a better understanding of any phonetic phenomena. The question of the relationship between production and perception has been frequently discussed in studies on L2 speech acquisition, especially in the study of the acquisition of L2 phonemes (see Llisterri, 1995 for a review). However, "the relationship between the perception of L2 speech sounds and their production by non-native speakers is still far from being understood" (Rochet, 1995: 406). This is particularly true for the acquisition of L2 prosody, which has only recently started to be studied from both the production and the perception perspectives (cf. Chun, 2002).

As for this dissertation, the decision to collect and analyze empirical data from both production and perception was aimed to have a deeper understanding of the realization of narrow focus by native and non-native speakers of English. In particular, it was expected that the results from production and perception would converge, resulting in a mutual validation of the respective findings.

The results of the production and perception study presented here indeed do show a certain convergence. This can be observed in the fact that the English native speakers were able to successfully realize and perceive narrow focus. As for non-native speakers, the production data of NNS1 show that the speakers were able to tune their productions to the native model, although not completely. This progress was confirmed perceptually by the results of Experiment 1, where English native listeners were still able to successfully identify narrow focus in the productions by NNS1. In contrast, the acoustic

analysis shows that NNS2 cannot not clearly mark focus by the sole use of prosodic cues. As expected, the lack of distinctive prosodic cues in the productions by NNS2 results in a difficult identification of focus from the perceptual point of view. The acoustic analysis of the sentences in Italian L1 also shows that neither duration nor F_0 were used to mark narrow focus. As for perception, the data from the Italian L1 listeners confirm the expectation that they are not able to identify narrow focus in absence of clear prosodic cues marking focus.

Furthermore, Experiment 1 also gave some perceptual evidence of the differences between perception in L1 and L2: the English native listeners were more successful at identifying narrow focus than the Italian listeners in English productions. In other words, the English native listeners were able to successfully identify focus in the productions by NS and by NNS1, while the Italians could recognize focus only in the productions by NS. The Italians' lower sensitivity seems to reflect the lower ability in the prosodic marking of focus that was generally observed in the production study, suggesting a link between production and perception.

To conclude, the results of the acoustic analysis and of the perception study are highly compatible and they confirm the expectation that the instances of narrow focus that are clearly marked prosodically are also the ones that are easier to be identified by the listeners. On the other hand, narrow focus result more difficult to be recognized when its realization is not properly marked by prosodic means, as in the cases of NNS2 and for the speech material in Italian.

Chapter 10

Conclusions

The research presented in this dissertation has implications both for theories of L2 speech acquisition and for L2 language instruction.

All the L2 speech acquisition models currently in use are based on a comparison between the phonological systems of L1 and L2. In particular, the models are principally focused on the acquisition of L2 phonemes. As mentioned in Chapter 3, the testing of L2 phoneme acquisition is based on experimental paradigms that cannot be readily adapted to the study of L2 prosody (Vaissière, 2005). For example, the perception tests on L2 phoneme acquisition can be performed without providing any contextual information to the subjects (Strange, 1995). This is not the case of the acquisition of L2 prosody, since the perception of prosody is context-dependent (see Section 3.3). Moreover, through prosody information is conveyed on a variety of different levels (Chun, 2002), where individual variation often hinders systematic generalizations (Grabe 2004).

Further research is needed to adapt the existing models or to create new ones to account for the acquisition of L2 prosody. This dissertation has hopefully provided empirical evidence that can contribute to the elaboration of models that can account for the acquisition of prosodic features of L2.

The results of this study show that the acquisition of English prosodic

focus marking is difficult for Italian speakers of English L2, suggesting that it should be specifically highlighted in language instruction so as to enhance its acquisition.

It is likely that the difficulties experienced by Italian learners are mainly generated by the structural differences in the prosodic systems of English and Italian. As for the the results presented in this study, in English narrow focus is marked with differences in f0, while in Italian the production data suggest that narrow non-contrastive focus is not prosodically marked.

The importance to learn correct prominence marking strategies has been acknowledged by Jenkins (2000), who listed correct prominence marking as one of the *core* aspects of pronunciation to acquire in order to avoid miscommunication in English. Jenkins included "nuclear stress production and placement" (Jenkins, 2000: 159), where 'nuclear stress' is used as a synonym for prominence (Celce-Murcia et al., 2010). In a recent study on the intonation of urban varieties of British English, including the SSBE variety used in this study, Grabe et al. (2008) found empirical evidence to support Jenkins's approach, concluding that "it is worth learning where native speakers place nuclear accents and why native listeners are used to consistency in nuclear accent placement" (Grabe et al., 2008: 22). It is also interesting to note that in Jenkins (2000) prominence marking is considered more important than the acquisition of pitch movements, which, in contrast, are considered non-core features.

From the pedagogical perspective, language instructors should insist on the correct acquisition of all levels of focus marking (information structure, prominence and acoustics, cf. Baker (2010) discussed in Section 3.3) with extensive explanations and practice activities, possibly based on the perception and production of the different types of focus. In particular, it has been suggested that since "[p]rominence is very sensitive to meaning, discourse, lexical stress, and syntactic boundaries", it "must be taught in rich contexts that permit learners to see what is new and what is important or contrastive

information" (Celce-Murcia et al., 2010: 226).

The first step in teaching how to mark prominence in English is to build conscious awareness on the mechanism of focus marking (Gilbert, 2008). The author of the present study speculates that the task proposed in the perception experiments presented in Chapters 8 and 9 could be adapted for a pedagogical context. Accompanied with proper instructions, a classroom activity could be based on listening to a sentence and then attempting to guess the question that could have prompted it as its answer. This could be a possible way to build a global awareness of how focus marking works in English. The robust results of the perception experiments and the positive feedback received from the participants represent encouraging starting points for carrying out further research to test such an activity in the classroom.

A significant finding based on the data presented this study is that English and Italian present significant differences in the implementation of pitch span. British English speakers present a significantly narrower pitch span as compared to what characterizes the productions by the non-native speakers, who, in turn, produce sentences with a significantly wider pitch span than the native speakers. This difference can also have consequences in communication, as pitch span is connected to the attitudinal level of meaning of intonation (see Mennen, 2007; Busà & Urbani, 2011; Urbani, 2013).

However, it is very difficult to imagine a way to teach the right implementation of pitch span. One way to deal with this problem, which could also be useful for learning prosodic focus marking strategies, is the use of the visual display of pitch contours with pitch tracking software, such as Praat or similar programs (e.g., Anderson-Hsieh, 1994; Chun, 1998; Levis & Pickering, 2004; Busà, 2007; Rocca, 2007; Hincks & Edlund, 2009). However, the initial enthusiasm that welcomed the use of visual aids for teaching intonation has been curbed by the difficulty to establish standardized methods and by the lack of studies showing results on long-term learning (Chun, 1998; Busà, 2008). In sum, more empirical research is required to prove the success

of these methods in the teaching/learning process.

In conclusion, the question on how to successfully teach the prosodic marking of focus in English remains unanswered. The main problem of teaching prominence marking, like other aspects connected with intonation and prosody, is that methods based on empirical data have not been sufficiently developed yet.

This dissertation has provided new data on both the production and the perception of Italian-accented English. However, the author is aware that this research could be enhanced and improved in several directions.

In the production study, only a small range of differences in the acoustic cues were measured. Such small differences can be attributed to the nature of narrow non-contrastive focus, which is less emphatic than its contrastive counterpart. However, this could also have been a byproduct of the elicitation protocol, and it might have been caused by the nature of the speech material that was collected, which was highly controlled. In this regard, Bishop (2011) observed that in the study on the perception of focus there might be a tradeoff in recurring to highly controlled speech material, which is possibly not optimal for eliciting fine-graded phonetic differences. In this regard, Bishop argues, "it may be that speakers [...] do not encode robust phonetic cues to the contrast when the context is highly salient, especially when reading printed materials" (Bishop, 2011: 313). The highly redundant context provided by the written and visual prompts used in this study could have limited the need for a clear characterization of focus. This could have been a cause for the small differences in production, regardless of the focus condition.

As for the speech material that was elicited from native and non-native speakers, the initial plan was to test the phonetic realization of narrow focus on four keywords per sentence, not only on subjects and verbs. However, as explained in Section 5.2.1.1, the last two keywords of each sentence (i.e., attribute and complement) were discarded from the analysis because they

presented longer values of duration and lower f0. These values were caused by the combined action of final lengthening and declination. The impossibility to use these keywords in a fair comparison was the reason why the analysis was limited to the first two keywords in the sentences, namely S and V.

As for the perception study, the main limitation of the two experiments resides in the use of a two-alternative forced-choice paradigm. This experimental paradigm has the intrinsic characteristic of limiting the participants' freedom of choice, so that their judgments are always to a certain extent guided to pre-decided options. However, the robust results obtained in the two experiments shows that the forced-choice paradigm was a viable heuristic for the tasks presented in the tests.

Further investigations should be based on the elicitation of sentences with more than two keywords, as was the original plan for the present data set. In a future study, a new data set should be designed by controlling for the presence of final lengthening and declination. The data set could also be made more homogeneous by using only monosyllabic words as keywords (cf. Xu & Xu, 2005; Breen et al., 2010).

From the point of view of the main research topic, this dissertation was aimed to study narrow non-contrastive focus. More dimensions of focus (e.g., contrastive vs. non-contrastive focus, narrow vs. broad focus...) could be studied in the future by adopting a methodological approach similar to the one followed in this study, collecting data from both production and perception.

The finding that Italian speakers have a significantly wider pitch span as compared to British English native speakers triggers a question from the perceptual point of view: what is the impact of such wide pitch span not only in the detection of focus, but also in the perception of Italian accent in English? The perception test presented in Pilot Study 4 (see Section 4.5) was an attempt to answer this question, but the heavy manipulation of the stimuli used in the experiment prevented from obtaining enlightening results

(cf. 4.5.3).

In order to study the perceptual impact of pitch span, it would also be interesting to collect speech material from speakers coming from different regional areas of Italy, to get a deeper understanding of the structural differences in pitch range found in the data presented in this dissertation. In particular, it would interesting to see if this prosodic behavior is a prerogative of the variety analyzed in this dissertation (North-East Italian) or if it can be considered as characteristic of Italian in general. It is clear that further research is required in order to define the role of the Italians' wider pitch span implementation in the perception of focus marking.

This thesis was aimed to investigate the phonetic realization of English narrow focus marking by Italian speakers at two different stages of their L2 acquisition. The production and perception data presented in this study converged in showing that the structural differences between the prosodic systems of the two languages result in difficulties for learners of English L2 in acquiring the focus marking strategies that characterize the target language. In particular, for the learners it is difficult to successfully adopt the plastic use of f0 to mark focus found in English productions, as in Italian word order strategies are normally preferred to mark prominence.

The findings reported here are particularly interesting not only for research in L2 speech acquisition, but also for their implications for language instruction, where prosodic aspects have recently started to be studied and taught with renewed interest (Busà, 2012).

Appendix A

DIPARTIMENTO DI STUDI LINGUISTICI E LETTERARI (DiSLL)

UNIVERSITÀ DEGLI STUDI DI PADOVA

Sede di via Beato Pellegrino, 26
35137 Padova
tel +39 049 8274951
fax +39 049 8274955

MODULO DI CONSENSO ALLA PARTECIPAZIONE A STUDIO LINGUISTICO E AL TRATTAMENTO DEI DATI PERSONALI

Con la presente io sottoscritto/a _____

Acconsento che la mia voce sia audioregistrata nell'ambito dello studio linguistico intrapreso dal ricercatore dottorando Rognoni Luca.

Acconsento inoltre al trattamento dei miei dati personali ai sensi della Legge 196/03, nella consapevolezza che i risultati del test verranno pubblicati anonimamente e che i dati non verranno in nessun caso divulgati per scopi diversi da quelli della ricerca scientifica.

In fede,

_____ (firma del partecipante) Padova, _____

Età	
Luogo di nascita	
Dove vivi?	
Professione	
Livello di studio	
e-mail	

Quali lingue straniere parli? A che livello (indicativamente)?
A che età hai iniziato a studiare inglese?
Hai mai vissuto per più in un paese anglofono? Se sì, dove e per quanto?

SPAZIO A CURA DEL RICERCATORE	
Dialang score	

DIPARTIMENTO DI STUDI LINGUISTICI E LETTERARI (DiSLL)

UNIVERSITÀ DEGLI STUDI DI PADOVA

Sede di via Beato Pellegrino, 26
35137 Padova
tel +39 049 8274951
fax +39 049 8274955

CONSENT FORM

I _____ (name and surname)

understand tha my voice will be recorded by the researcher Luca Rognoni, PhD student at the University of Padova, Italy as part of a control group for a study in the phonetics of foreign-accented English.

I also understand that my personal data will be treated anonymously and for the sole purpose of scientific research.

_____ (signature) London, _____ (date)

Date of Birth	
Place of Birth	
Where do you live?	
Profession	
Level of education	
e-mail	

How many languages do you speak?

Appendix B

English L1 and L2 sentences

Subject in focus

Who walks with the green frog?

Carlos walks with the green frog.
Jacob walks with the green frog.
Bobbie walks with the green frog.
Ginny walks with the green frog.
Selma walks with the green frog.

Verb in focus

What does Carlos do with the red fox?

Carlos walks with the red fox.
Carlos runs with the red fox.
Carlos eats with the red fox.
Carlos jumps with the red fox.
Carlos drinks with the red fox.

Attribute in focus

What cat does Bobbie run with?

Bobbie runs with the green cat.
Bobbie runs with the black cat.
Bobbie runs with the red cat.
Bobbie runs with the blue cat.
Bobbie runs with the pink cat.

Object in focus

What animal does Martha speak to?

Martha speaks to the black frog.
Martha speaks to the black hen.
Martha speaks to the black cat.
Martha speaks to the black fox.
Martha speaks to the black dog.

Italian L1 sentences

Subject in focus

Chi gioca con la rana verde?

Luca gioca con la rana verde.
Salvo gioca con la rana verde.
Giorgio gioca con la rana verde.
Marta gioca con la rana verde.
Carla gioca con la rana verde.

Verb in focus

Che cosa fa Salvo con la volpe rossa?

Salvo gioca con la volpe rossa.

Salvo corre con la volpe rossa.
Salvo mangia con la volpe rossa.
Salvo salta con la volpe rossa.
Salvo beve con la volpe rossa.

Attribute in focus

Con quale gatto corre Carla?

Carla corre con il gatto verde.
Carla corre con il gatto nero.
Carla corre con il gatto rosso.
Carla corre con il gatto giallo.
Carla corre con il gatto rosa.

Object in focus

Con che animale parla Emma?

Emma parla con la rana nera.
Emma parla con il pollo nero.
Emma parla con il gatto nero.
Emma parla con la volpe nera.
Emma parla con il cane nero.

Appendix C

Instructions for Experiment 1

Instructions for English native listeners

When you listen to an answer out of its context, can you correctly guess the question that triggered that answer?

When speaking English, we concentrate attention on particular parts of the message according to the communication needs of our conversation by using our intonation (that is, the "melody" and "tempo" in our speech). In particular, when we are asked a question, in our answer we normally emphasize, or highlight, the most relevant piece of information using intonation. As a result, the same sentence can be uttered in slightly different ways depending on the context.

Typically, the most relevant piece of information is the element of the sentence corresponding to the wh-element in the question. For example, if a sentence is an answer to a question like: "Who's eating a pear?", the answer would be: "Bobbie's eating a pear." Similarly, when replying to a question like: "What's Bobbie eating?", the answer would sound like: "Bobbie is eating a pear".

The task

In this experiment you will be presented with a series of short sentences produced by native and non-native English speakers as answers to

wh-questions.

You will be asked to select which question is more likely to have triggered the answer. Your choice will be limited to two options. The system will play each sentence automatically, but you are allowed to listen to the sentences as many times as you wish; you are invited to make an informed guess even when the correspondence is not straightforward. The task normally takes around 15 minutes to be completed and it is preceded by a short training phase, where you can familiarize with your task and with the interface.

Click Next when you are ready to begin.

Instructions for Italian native listeners

Quando ascolti una risposta fuori dal suo contesto, sei in grado di individuare la domanda che ha provocato la risposta?

I parlanti nativi di inglese, quando parlano la loro lingua, concentrano la loro attenzione su particolari parti del messaggio, in base alle necessità comunicative della conversazione in atto, facendo uso dell'intonazione (la "melodia" e il "tempo" del discorso parlato). In particolare, quando si risponde a una domanda, in inglese si enfatizza, cioè si rende più evidente, l'informazione più rilevante utilizzando l'intonazione. Di conseguenza, una frase può essere pronunciata in modi leggermente diversi a seconda del contesto.

Generalmente, l'informazione più rilevante si identifica con l'elemento della frase che corrisponde all'elemento *wh-* in una domanda (per esempio: "what", "who", "where"...).

Ad esempio, se una frase è la risposta alla domanda: "Who's eating a pear?", la risposta sarebbe: "Bobbie's eating a pear." Così, quando si risponde a una domanda come: "What's Bobbie eating?", la risposta dovrebbe essere: "Bobbie is eating a pear".

Il compito

In questo esperimento vi sarà presentata una serie di brevi risposte

realizzate da parlanti nativi e non di inglese come risposte a domande *wh-*.

Vi sarà richiesto di selezionare la domanda che più probabilmente ha provocato la risposta. La vostra scelta sarà ristretta a due opzioni. Il sistema riprodurrà automaticamente ogni frase una volta, ma avrete la possibilità di riprodurre le frasi manualmente, se lo ritenete necessario.

Le indicazioni che accompagneranno ogni frase saranno in inglese: "Listen to the sentence and select the question that matches it best. If you want you can play the sound more than once." Questa è la traduzione: "Ascolta la frase e seleziona la domanda che meglio corrisponde. Se lo desideri, puoi riprodurre il suono più di una volta".

Questo esperimento durerà circa 15 minuti e sarà preceduto da una breve fase di *training* nella quale potrete familiarizzare con il compito e con l'interfaccia del programma.

Cliccate su Next quando siete pronti.

Instructions for the Italian L1 block of stimuli

In questa fase dell'esperimento vi sarà presentata una serie di brevi risposte realizzate da parlanti italiani come risposte a domande parziali, cioè del tipo "chi?" o "che cosa?". Vi sarà richiesto di selezionare la domanda che più probabilmente ha provocato la risposta solo sulla base dell'ascolto della frase, senza ulteriore contesto. La vostra scelta sarà ristretta a due opzioni. Il sistema riprodurrà automaticamente ogni frase una sola volta, ma avrete la possibilità di riprodurre le frasi manualmente, se lo ritenete necessario.

Questo esperimento durerà circa 10 minuti.

Cliccate su Next quando siete pronti.

Instructions for Experiment 2

When you listen to an answer out of its context, can you correctly guess the question that triggered that answer?

When speaking English, we concentrate our attention on particular parts of the message according to the communication needs of our conversation by using our intonation (that is, the "melody" and "tempo" in our speech). In particular, when we are asked a question, in our answer we normally emphasize, or highlight, the most relevant piece of information using intonation. As a result, the same sentence can be uttered in slightly different ways depending on the context.

Typically, the most relevant piece of information is the element of the sentence corresponding to the wh-element in the question. For example, if a sentence is an answer to a question like: "Who's eating a pear?", the answer would be: "Bobbie's eating a pear." Similarly, when replying to a question like: "What's Bobbie eating?", the answer would sound like: "Bobbie is eating a pear".

The task

In this experiment you will be presented with a series of short sentences produced as answers to *wh-questions* by two voices: one native and one non-native speaker of English.

Some characteristics of the two voices have been digitally modified, so you are asked to pay particular attention: the sentences might sound the same, but they are all slightly different one from the other.

You will be asked to select which question is more likely to have triggered the answer. Your choice will be limited to two options. The system will play each sentence automatically, but you are allowed to listen to the sentences as many times as you wish; you are invited to make an informed guess even when the correspondence is not straightforward. The task normally takes around 10 minutes to be completed.

Click Next when you are ready to begin.

References

Adams, C., & Munro, M. J. (1978). In search of the acoustic correlates of stress: Fundamental frequency, amplitude, and duration in the connected utterances of some native and nonnative speakers of English. *Phonetica*, *35*, 125-156.

Albano Leoni, F. (2009). *Dei suoni e dei sensi*. Bologna: Il Mulino.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., ... Weinert, R. . (1991). The HCRC Map Task Corpus. *Language and Speech*, *34*, 351-366.

Anderson-Hsieh, J. (1994). Interpreting visual feedback on suprasegmentals in computer assisted pronunciation instruction. *CALICO Journal*, *11*(4), 5-22.

Anderson-Hsieh, J., Johnson, R., & Kohler, K. J. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, *42*(4), 529-555.

Avesani, C., & Vayra, M. (2003). Broad, narrow and contrastive focus in Florentine Italian. *Proceedings of 15th ICPhS, Barcelona, Spain*, 1803-1806.

Avesani, C., & Vayra, M. (2005). Accenting, deaccenting and information structure in Italian dialogue. *Proc. 6th DIGdial Workshop on Discourse and Dialogue, Lisbon, Portugal*, 19-24.

Azzaro, G. (2006). *Sounds right. comprensione, pronuncia, apprendimento dell'inglese L2*. Rome: Aracne.

Baker, R. E. (2010). *The acquisition of English focus marking by non-native speakers*. Unpublished doctoral dissertation.

Bartels, C., & Kingston, J. (1994). Salient pitch cues in the perception of contrastive focus. *Proc. of J. Sem. conference on Focus. IBM Working Papers*, 94-106.

Bent, T., & Bradlow, A. (2003). The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America*, *114*(3), 1600-1610.

Bertinetto, G. M. (1981). *Strutture prosodiche dell'Italiano*. Firenze: Accademia della Crusca.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research* (p. 13-45). Timonium, MD: York Press.

Best, C. T., & Tyler, M. (2007). Nonnative and second-language speech perception. commonalities and complementarities. In O.-S. Bohn (Ed.), *Language experience in second language speech learning. in honor of James Emil Flege* (p. 13-34). Amsterdam: John Benjamins.

Bigi, B., & Hirst, D. (2012). SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody. *Proc. Speech Prosody, Shanghai, China*.

Bocci, G., & Avesani, C. (2008). Deaccent given or define focus? where Italian doesn't sound like English. *Paper presented at 6th Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Naples, 3-5 February 2010*.

Bocci, G., & Avesani, C. (2010). Givenness, deaccentazione e il ruolo di l* nell'Italiano di toscana. *Paper presented at 34th Incontro di Grammatica Generativa, Padua, 21-23 February 2008*.

Boersma, P., & Weenink, D. (2014). *Praat: doing phonetics by computer* [Computer Program]. Retrieved from `http://www.praat.org/`

Bohn, O.-S. (1995). Cross-language speech perception. first language transfer doesn't tell it all. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research* (p. 275-300). Timonium, MD: York Press.

Boula de Mareüil, P., Brahimi, B., & Gendrot, C. (2004). Role of segmental and suprasegmental cues in the perception of Maghrebian-accented French. *Proceedings of Interspeech, Jeju Island, Korea*, 341-344.

Boula de Mareüil, P., & Vieru-Dimulescu, B. (2006). The contribution of prosody to the perception of foreign accent. *Phonetica*, *63*(4), 247-267.

Breen, M., Dilley, L. C., Kraemer, J., & Gibson, E. (2012). Inter-transcriber reliability for two systems of prosodic annotation: Tobi (tones and break indices) and rap (rhythm and pitch). *Corpus Linguistics and Linguistic Theory*, *8*(2), 277-312.

Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes*, *25*(7), 1044-1098.

Büring, D. (2007). Semantics, intonation, and information structure. In G. Ramchand & C. Reiss (Eds.), *The oxford handbook of linguistic interfaces* (p. 445-474). Oxford: Oxford University Press.

Büring, D. (2009). Towards a typology of focus realization. In M. Zimmermann & C. Fary (Eds.), *Information structure* (p. 177-205). Oxford: Oxford University Press.

Busà, M. G. (1995). *L'inglese degli Italiani. l'acquisizione delle vocali.* Padua: Unipress.

Busà, M. G. (2007). New perspectives in teaching pronunciation. In A. Baldry, M. Pavesi, & C. Taylor Torsello (Eds.), *From didactas to ecolingua: an ongoing research project on translation and corpus linguistics* (p. 171-188). Trieste: Edizioni UniversitàĂ di Trieste.

Busà, M. G. (2008). Teaching prosody to Italian learners of English: working towards a new approach. In C. Taylor (Ed.), *Ecolingua: The role of e-corpora in translation, language learning and testing* (p. 113-126). Trieste: Edizioni Università di Trieste.

Busà, M. G. (2010). Effects of L1 on L2 pronunciation: Italian prosody in English. In A. Gagliardi & A. Maley (Eds.), *ILS, ELF, Global English: Teaching and learning processes, linguistic insights: Studies in language and communication* (p. 200-228). Bern: Peter Lang.

Busà, M. G. (2012). The role of prosody in pronunciation teaching: A growing appreciation. In M. G. Busà & A. Stella (Eds.), *Methodological perspectives on second language prosody. papers from ML2P 2012* (p. 101-106). Padua: Cleup.

Busà, M. G., & Rognoni, L. (2012). Italians speaking English: The contribution of verbal and non-verbal behavior. In H. Mello, M. Pettorino, & T. Raso (Eds.), *Proceedings of the 7th GSCP international conference: Speech and corpora* (p. 313-317). Florence: Firenze University Press.

Busà, M. G., & Stella, A. (2012a). Intonational variations in focus marking in the English spoken by north-east Italian speakers. In M. G. Busà & A. Stella (Eds.), *Methodological perspectives on second language prosody. papers from ML2P 2012* (p. 31-35). Padua: Cleup.

Busà, M. G., & Stella, A. (2012b). *Methodological perspectives on second language prosody. papers from ML2P 2012* [Edited Book]. Padua: Cleup.

Busà, M. G., & Urbani, M. (2011). A cross linguistic analysis of pitch range in English L1 and l2. *Proc. 17th International Conference of Phonetic Sciences (ICPhS), Hong Kong, China*, 380-383.

Celce-Murcia, M., Brinton, D. M., Goodwin, J. M., & Griner, B. (2010). *Teaching pronunciation. A course book and reference guide.* Cambridge: Cambridge University Press.

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics and points of view. In C. N. Li (Ed.), *Subject and topic* (p. 27-55). New York: Academic Press.

Chun, D. M. (1998). Signal analysis software for teaching discourse intonation. *Language Learning & Technology*, *2*(1), 61-77.

Chun, D. M. (2002). *Discourse intonation in l2. from theory and research to practice.* Amsterdam: John Benjamins.

Cooper, W. E., Eady, S. J., & Mueller, P. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America*, *77*.

Council of Europe. (2001). *Common european framework of reference for languages: Learning, teaching, assessment.* Strasbourg: Council of Europe.

Couper-Kuhlen, E. (1984). A new look at contrastive intonation. In R. Watts & U. Weidman (Eds.), *Modes of interpretation: Essays presented to Ernst Leisi* (p. 137-158). Tübingen: Gunter Narr Verlag.

Cruttenden, A. (1997). *Intonation.* Cambridge: Cambridge University Press.

Darcy, I. (in press). Phonological attention control, inhibition, and second language speech learning. *Proc. New Sounds 2013, Concordia University, Montreal, Canada*.

Darcy, I., Dekydtspotter, L., Sprouse, R. A., Glover, J., Kaden, C., McGuire, M., & Scott, J. H. G. (2012). Direct mapping of acoustics to phonology: On the lexical encoding of front rounded vowels in L1 English-l2 French acquisition. *Second Language Research*, *28*, 1-36.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalysed. *Journal of Phonetics*, *11*, 51-62.

De Meo, A. (2012). How credible is a non-native speaker? prosody and surroundings. In M. G. Busà & A. Stella (Eds.), *Methodological perspectives on second language prosody. papers from ML2P 2012* (p. 3-9). Padua: Cleup.

De Meo, A., Pettorino, M., & Vitale, M. (2012). Transplanting credibility into a foreign voice. an experiment on synthesized l2 Italian. In H. Mello, M. Pettorino, & T. Raso (Eds.), *Proceedings of the 7th GSCP international conference: Speech and corpora* (p. 281-284). Florence: Firenze University Press.

De Meo, A., Vitale, M., Pettorino, M., Cutugno, F., & Origlia, A. (2013). Imitation/self-imitation in computer- assisted prosody training for Chinese learners of L2 Italian. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 4th pronunciation in second language learning and teaching conference* (p. 90-100). Ames, IA: Iowa State University.

De Meo, A., Vitale, M., Pettorino, M., & Martin, P. (2012). Acoustic-perceptual credibility correlates of news reading by native and Chinese speakers of Italian. *Proc. 17th International Congress of Phonetic Sciences (ICPhS), Hong Kong, China*, 1366-1369.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility and comprehensibility. evidence from four l1s. *Studies in Second Language Acquisition*, *19*(1), 1-16.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*(4), 476-490.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, *63*(2), 163-185.

D'Imperio, M. (2002). Italian intonation: an overview and some questions. *Probus*, *14*(1), 37-69.

Drullman, R., & Collier, R. (1991). On the combined use of accented and unaccented diphones in speech synthesis. *Journal of the Acoustical Society of America*, *90*, 17-66-1775.

Duguid, E. (2001). Italian speakers. In M. Swan (Ed.), *Learner english: A teacher's guide to interference and other problems* (2nd ed., p. 73-89). Cambridge: Cambridge University Press.

Eady, S., Cooper, W. E., Klouda, G., MÃijller, P., & Lotts, D. (1986). Acoustical characteristics of sentential focus: Narrow vs. broad and single vs. dual focus environments. *Language and Speech*, *29*(3), 233-251.

Elliott, A. R. (1995). Field independence/dependence, hemispheric specialization, and attitude in relation to pronunciation accuracy in Spanish as a foreign language. *The Modern Language Journal*, *79*(3), 356-371.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Escudero, P. (2005). *Linguistic perception and second language acquisition. explaining the attainment of optimal phonological categorization.* Doctoral dissertation, University of Utrecht.

Face, T. L. (2003). Intonation in Spanish declaratives: differences between lab speech and spontaneous speech. *Catalan Journal of Linguistics*, *2*, 115-131.

Face, T. L. (2007). The role of intonational cues in the perception of declaratives and absolute interrogatives in castilian Spanish. *Estudios de fonètica experimental*, *16*, 185-225.

Face, T. L., & D'Imperio, M. (2005). Reconsidering a focal typology: Evidence from Spanish and Italian. *Italian Journal of Linguistics*, *17*, 271-289.

Flege, J. E. (1984). The detection of French accent by American listeners. *Journal of the Acoustical Society of America*, *76*(3), 692-707.

Flege, J. E. (1987). The production of "new" and 'similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics*, *15*, 47-65.

Flege, J. E. (1995). Second language speech learning. theory, findings and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research* (p. 229-273). Timonium, MD: York Press.

Flege, J. E. (1999). Age of learning and second-language speech. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (p. 101-132). Hillsdale, NJ: Lawrence Erlbaum.

Flege, J. E. (2002). Interactions between the native and second-language phonetic systems. In P. Burmeister, T. Piske, & A. Rohde (Eds.), *An*

integrated view of language development: Papers in honor of Henning Wode (p. 217-244). Trier: Wissenschaftlicher Verlag.

Flege, J. E., & Fletcher, K. L. (1992). Talker and listener effects on degree of perceived foreign accent. *Journal of the Acoustical Society of America*, *9*(1), 370-389.

Flege, J. E., MacKay, I. R. A., & Meador, D. (1999). Native Italian speakers' production and perception of English vowels. *Journal of the Acoustical Society of America*, *106*, 2973-2987.

Fowler, C. A., & Galantucci, B. (2005). The relation of speech perception and speech production. In D. Pisoni & R. Remez (Eds.), *The handbook of speech perception* (p. 633-652). London: Blackwell.

Frascarelli, M. (2004). L'interpretazione del focus e la portata degli operatori sintattici. In F. Albano Leoni, F. Cutugno, M. Pettorino, & R. Savy (Eds.), *Il parlato Italiano. atti del convegno nazionale (napoli 13-15 febbraio 2003)*. Napoli: D'Auria Editore.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, *27*, 765-768.

Gagliardi, G., Lombardi Vallauri, E., & Tamburini, F. (2004). La prominenza in Italiano: demarcazione più che culminazione. *Atti del VIII Convegno dell'Associazione Italiana Scienze della Voce*, 255-270.

Gass, S., & Varonis, E. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, *34*, 65-89.

Gilbert, J. B. (2008). *Teaching pronunciation using the prosody pyramid.* Cambridge: Cambridge University Press.

Gili Fivela, B. (2002). Tonal alignment in two Pisa Italian peak accents. *Proc. Speech Prosody, Aix-en-Provence, France*, 339-342.

Gili Fivela, B. (2012). Testing the perception of L2 intonation. In M. G. Busà & A. Stella (Eds.), *Methodological perspectives on second language prosody. papers from ML2P 2012* (p. 17-30). Padua: Cleup.

Gili Fivela, B., Avesani, C., Barone, M., Bocci, G., Crocco, C., D'Imperio, M., ... Sorianello, P. (to appear). Varieties of Italian and their intonational phonology. In S. Frota & P. Prieto (Eds.), *Intonation in romance.* Oxford: Oxford University Press.

Giordano, R. (2006). Note sulla fonetica del ritmo dell'Italiano. *Atti del II Convegno Nazionale dell'Associazione Italiana di Scienze della Voce (AISV), Salerno*, 233-244.

Grabe, E. (2004). Intonational variation in urban dialects of English spoken in the British Isles. In P. Gilles & J. Peters (Eds.), *Regional variation in intonation* (p. 9-31). Tübingen: Niemeyer.

Grabe, E., Kochanski, G., & Coleman, J. (2008). The intonation of native accent varieties in the British Isles. potential for miscommunication? In K. Dziubalska-Kolaczyk & J. Przedlacka (Eds.), *English pronunciation models: a changing scene* (p. 311-337). Bern: Peter Lang.

Grice, M., & Baumann, S. (2007). An introduction to intonation. functions and models. In J. Trouvain & U. Gut (Eds.), *Non-native prosody. phonetic description and teaching practice* (p. 25-51). Berlin: Mouton De Gruyter.

Grice, M., D'Imperio, M., Savino, M., & Avesani, C. (2005). Strategies for intonation labelling across varieties of Italian. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (p. 362- 389). Oxford: Oxford University Press.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, *28*, 267-283.

Gussenhoven, C. (1983). Testing the reality of focus domains. *Language and Speech*, *26*, 61-80.

Gut, U. (2012). Rhythm in L2 speech. In D. Gibbon (Ed.), *Speech and language technology* (p. 83-94). Poznan.

He, X., Hanssen, J., Van Heuven, V. J., & Gussenhoven, C. (2011). Phonetic implementation must be learnt: native versus Chinese realization of focus accent in Dutch. *Proc. 17th International Conference of Phonetic Sciences (ICPhS), Hong Kong, China*, 843-846.

Heldner, M. (2003). On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in swedish. *Journal of Phonetics*, *31*, 39-62.

Hincks, R. (2004). Processing the prosody of oral presentations. *Proceedings of InSTIL/ICALL Symposium on Computer Assisted Language Learning, Venice*, 63-69.

Hincks, R. (2010). Speaking rate and information content in English lingua franca oral presentations. *English for Specific Purposes*, *29*(1), 4-18.

Hincks, R., & Edlund, J. (2009). Promoting increased pitch variation in oral presentations with transient visual feedback. *Language Learning and Technology*, *13*(3), 32-50.

Holm, S. (2007). The relative contributions of intonation and duration to intelligibility in norwegian as a second language. *Proc. 16th International Conference of Phonetic Sciences (ICPhS), Saarbrücken, Germany*, 1653-1656.

Hongyan, W., & van Heuven, V. J. (2007). Quantifying the interlanguage speech intelligibility benefit. *Proc. 16th International Conference of Phonetic Sciences (ICPhS), Saarbrücken, Germany*, 1729.1732.

Ito, K., Speer, S., & Beckman, M. (2004). Informational status and pitch accent distribution in spontaneous dialogues in English. *Proc. Spoken Language Processing, Nara, Japan*, 279-282.

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford: Oxford University Press.

Jesney, K. (2004). The use of global foreign accent rating in studies of L2 acquisition. *Language Research Centre University of Calgary Working Papers*.

Jilka, M. (2000). *The contribution of prosody to the perception of foreign accent*. Doctoral dissertation, University of Stuttgart.

Jilka, M. (2007). Different manifestations and perceptions of foreign accent in intonation. In J. Trouvain & U. Gut (Eds.), *Non-native prosody. phonetic description and teaching practice* (p. 77-96). Berlin: Mouton de Gruyter.

Jun, S.-A. (2005). *Prosodic typology: The phonology of intonation and phrasing* [Edited Book]. Oxford: Oxford University Press.

Kawahara, H. (2008). Tandem-straight, a research tool for L2 study enabling flexible manipulations of prosodic information. *Proc. Speech Prosody, Campinas, Brazil*, 619-628.

Klatt, D. H. (1973). Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. *Journal of the Acoustical Society of America*, *53*, 8-16.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, *67*(3), 971-995.

Kohler, K. J. (2003). Neglected categories in the modelling of prosody. pitch timing and non-pitch accents. *Proc. 15th International Conference of Phonetic Sciences (ICPhS), Barcelona, Spain*, 2925-2928.

Kohler, K. J. (2006). What is emphasis and how is it coded? *Proc. Speech Prosody, Dresden, Germany*, 748-751.

Kori, S., & Farnetani, E. (1983). Acoustic manifestation of focus in Italian. *Quaderni del Centro di Sudio per le Ricerche di Fonetica*, *2*, 323-338.

Krahmer, E., & Swerts, M. (2001). On the alleged existence of contrastive accents. *Speech Communication*, *34*, 391-405.

Krahmer, E., & Swerts, M. (2004). More about brows. In Z. Ruttkay & C. Pelachaud (Eds.), *From brows to trust: Evaluating embodied conversational agents* (p. 191-216). Dordrecht: Kluwer Academic Press.

Kuhl, P. K. (1991). Human adults and human infants show a perceptual magnet effect for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics*, *50*(93-107).

Ladd, D. R. (1980). *The structure of intonational meaning: evidence from English.* Bloomington: Indiana University Press.

Ladd, D. R. (1996). *Intonational phonology* (1st ed.). Cambridge: Cambridge University Press.

Ladd, D. R. (2008). *Intonational phonology* (2nd ed.). Cambridge: Cambridge University Press.

LePage, A., & Busà, M. G. (in press). Intelligibility of English L2: The effects of lack of vowel reduction and incorrect word stress placement in the speech of French and Italian learner. *Proc. New Sounds 2013, Concordia University, Montreal, Canada*.

Lepschy, A. L., & Lepschy, G. (1977). *The Italian language today.* London: Hutchinson.

Levis, J., & Pickering, L. (2004). Teaching intonation in discourse using speech visualisation technology. *System*, *34*, 505-524.

Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America*, *32*(4), 451-454.

Llisterri, J. (1995). Relationships between speech production and speech perception in a second language. *Proc. 13th International Conference of Phonetic Sciences (ICPhS), Stockholm, Sweden*, 92-99.

Magen, H. S. (1998). The perception of foreign-accented speech. *Journal of Phonetics*, *26*(4), 381-400.

Magno Caldognetto, E., & Fava, E. (1974). Studio sperimentale delle caratteristiche elettroacustiche dell'enfasi su sintagmi in Italiano. *Atti del VI Congresso Internazionale di Studi. Fenomeni morfologici e sintattici nell'italiano contemporaneo*, 441-156.

Magno Caldognetto, E., Ferrero, F., Vagges, K., & K., C. (1983). Indici acustici della struttura sintattica: un contributo sperimentale. In *Scritti linguistici in onore di g.b. pellegrini* (p. 1127-1156). Pisa: Pacini.

Mairano, P. (2011). *Rhythm typology: acoustic and perceptive studies.* Doctoral dissertation, University of Turin.

Major, R. (1987). Phonological similarity, markedness, and rate of L2 acquisition. *Studies in Second Language Acquisition*, *9*(1), 63-82.

Major, R. (2001). *Foreign accent: The ontogeny and phylogeny of second language homology.* Mahwah, NJ: Lauwrence Erlbaum Associates.

Marotta, G. (1985). *Modelli e misure ritmiche.* Bologna: Zanichelli.

Marotta, G. (2008). Sulla percezione dell'accento straniero. In U. Lazzeroni (Ed.), *Diachronica et synchronica. studi in onore di anna giacalone ramat* (p. 327-345). Pisa: ETS.

Marotta, G., Calamai, S., & Sardelli, E. (2004). Non di sola lunghezza. la modulazione di f0 come indice socio-fonetico. In A. De Dominicis, L. Mori, & M. Stefani (Eds.), *Costituzione, gestione e restauro di corpora vocali. atti delle xiv giornate del fgs* (p. 210-215). Rome: Esagrafica.

Marotta, G., Molino, A., & Bertini, C. (2012). Lunghezza e frequenza nell'espressione e nella percezione della prominenza. un'analisi empirica. *L'Italia Dialettale*, *73*(67-99).

Marotta, G., & Sardelli, E. (2007). Prosodic parameters for the detection of regional varieties in Italian. *Proc. 16th International Conference of Phonetic Sciences (ICPhS), Saarbrücken, Germany*, 682-704.

Martin, P. (2004). Winpitchpro. a tool for text to speech alignment and prosodic analysis. *Proc. Speech Prosody, Nara, Japan*.

Mathot, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314-324.

McCullogh, E. A. (2013). *Acoustic correlates of perceived foreign accent in non-native English*. Doctoral dissertation, Ohio State University.

Medina, E., & Solorio, T. (2006). Wavesurfer: a tool for sound analysis. *Departmental Technical Reports (CS). University of Texas at El Paso*.

Mennen, I. (1999). The realisation of nucleus placement in second language intonation. *Proc. 14th International Conference of Phonetic Sciences (ICPhS), San Francisco, CA*, 555-558.

Mennen, I. (2007). Phonological and phonetic influences in non-native intonation. In J. Trouvain & U. Gut (Eds.), *Non-native prosody. phonetic description and teaching practice* (p. 53-76). Berlin: Mouton de Gruyter.

Mertens, P. (1991). Local prominence of acoustic and psychoacoustic functions and perceived stress in French. *Proc. 12th International Conference of Phonetic Sciences (ICPhS), Aix-en-Provence, France*, 218-221.

Mertens, P. (2013). Automatic labelling of pitch levels and pitch movements in speech corpora. *Proc. TRASP 2013, Aix-en-Provence, France*, 42-46.

Molnar, V. (2002). Contrast - from a contrastive perspective. In H. Hallelgard, S. Johansson, B. Behrens, & C. Fabricius-Hansen (Eds.), *Proc. symposium on information structure in a cross-linguistic perspective* (p. 147-161).

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

Munro, M. J. (1995). Nonsegmental factors in foreign accent: Ratings of filtered speech. *Studies in Second Language Acquisition*, *17*(1), 17-34.

Munro, M. J. (2008). Foreign accent and speech intelligibility. In E. Hansen & M. L. Zampini (Eds.), *Phonology and second language acquisition* (p. 199-218). Amsterdam: John Benjamins.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73-97.

Munro, M. J., & Derwing, T. M. (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication*, *52*, 626-637.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, *28*(1), 111-131.

Ohala, J., & Gilbert, J. B. (1981). Listeners' ability to identify languages by their prosody. *Studia Phonetica*, *19*, 123-131.

Origlia, A., & Alfano, I. (2012). Prosomarker: a prosodic analysis tool based on optimal pitch stylization and automatic syllabification. *Proc. 8th LREC, Istanbul, Turkey*.

Passino, D. (2005). *Aspects of consonantal lengthening in Italian*. Doctoral dissertation, University of Padua.

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Perrier, P., Vick, J., ... Zandipour, M. (2000). A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *Journal of Phonetics*, *28*, 233-272.

Petrone, C. (2008). *From targets to tunes: Nuclear and prenuclear contribution in the identification of intonation contours in Italian*. Doctoral dissertation, Universitè de Provence.

Pettorino, M., & Vitale, M. (2012). Transplanting prosody into non-native speech. In M. G. Busà & A. Stella (Eds.), *Methodological perspectives on second language prosody. papers from ML2P 2012* (p. 11-16). Padua: Cleup.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. Doctoral dissertation, M.I.T.

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, & M. Pollack (Eds.), *Intentions in communication* (p. 273-311). Cambridge, MA: M.I.T. Press.

Piske, T., MacKay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: a review. *Journal of Phonetics*, *29*(2), 191-215.

Quenè, H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics*, *35*, 353-362.

Ramìrez Verdugo, M. D. (2006). Prosodic realization of focus in the discourse of Spanish learners and English native speakers. *Estudios ingleses de la Universidad Complutense*, *14*, 9-32.

Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America*, *105*(1), 512-521.

Rasier, L., & Hiligsmann, P. (2007). Prosodic transfer from L1 to L2. theorical and methodogical issues. *Nouveaux cahiers de linguistique française*, *28*, 41-66.

Repetti, L. (2012). Consonant-final loanwords and epenthetic vowels in Italian. *Catalan Journal of Linguistics*, *11*(167-188).

Rocca, P. D. A. (2007). New trends on the teaching of intonation of foreign languages. *Proceedings of New Sounds, Florianopolis, Brazil*, 420-428.

Rochet, B. L. (1995). Perception and production of L2 speech sounds by adults. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research* (p. 379-410). Timonium, MD: York Press.

Rognoni, L. (2012). The impact of prosody in foreign accent detection. a perception study of Italian accent in English. In M. G. Busà & A. Stella (Eds.), *Methodological perspectives on second language prosody. papers from ML2P 2012* (p. 89-93). Padua: Cleup.

Rognoni, L., & Busà, M. G. (in press). Testing the effects of segmental and suprasegmental phonetic cues in foreign accent rating. an experiment using prosody transplantation. *Proc. New Sounds 2013, Concordia University, Montreal, Canada*.

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, *1*, 75-116.

Rump, H. H. (1996). *Prominence of pitch-accented syllables.* Doctoral dissertation, Technische Universiteit Eindhoven.

Rump, H. H., & Collier, R. (1996). Focus conditions and the prominence of pitch-accented syllables. *Language and Speech*, *39*, 1-17.

Schmitz, C. (2012). *LimeSurvey: An open source survey tool* [Computer Program]. LimeSurvey Project. Retrieved from `http://www.limesurvey.org/`

Schröder, M., & Trouvain, J. (2003). The german text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology*, *6*, 395-377.

Schwarzschild, R. (1999). Givenness, Avoid F, and other constraints on the placement of accent. *Natural Language Semantics*, *7*, 141-77.

Selkirk, L. (1972). *Phonology and syntax: the relation between sound and structure.* Cambridge, MA: M.I.T. Press.

Signorello, R., Poggi, I., & Demolin, D. (2012). Charisma perception in political speech: a case study. In H. Mello, M. Pettorino, & T. Raso (Eds.), *Proceedings of the 7th GSCP international conference: Speech and corpora* (p. 281-284). Florence: Firenze University Press.

Silverman, K., Beckman, M., Pierrehumbert, J., Ostendorf, M., Wightman, C., Price, P., & Hirschberg, J. (1992). Tobi: A standard scheme for labeling prosody. *Proc. International Conference of Spoken Language Processing, Banff, Canada*, 867-869.

Slowiaczek, M. L. (1994). Semantic priming in a single-word shadowing task. *American Journal of Psychology*, *107*, 245-260.

Sluijter, A., & van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, *100*, 2471-2485.

Sonntag, G. P., & Portele, T. (1998). Comparative evaluation of synthetic prosody with the purr method. *Proc. ICSLP, Sydney, Australia*, 3-6.

Sorianello, P. (2006). *Prosodia. Modelli e ricerca empirica.* Rome: Carocci.

Stella, A., & Busà, M. G. (in press). Transfer intonativo nell'inglese L2 prodotto da parlanti padovani: il caso delle domande polari. *Atti del IX Convegno Nazionale AISV (Associazione Italiana di Scienze della Voce), Venice, Italy*.

Stella, A., & Gili Fivela, B. (2009). L'intonazione nell'Italiano dell'area leccese: prime osservazioni dal punto di vista autosegmentale-metrico. In L. Romito, V. Galatà, & R. Lio (Eds.), *La fonetica sperimentale: metodo e applicazioni. atti del iv convegno nazionale AISV (Associazione Italiana di Scienze della Voce)* (p. 259-292). Torriana (Rimini): EDK Editore.

Strange, W. (1995). Cross-language studies of speech perception. a historical review. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research* (p. 3-45). Timonium, MD: York Press.

Swerts, M., Krahmer, E., & Avesani, C. (2002). Prosodic marking of information status in Dutch and Italian: a comparative analysis. *Journal of Phonetics*, *30*, 629-654.

Tajima, K., Port, R., & Dalby, J. (1996). Foreign-accented rhythm and prosody in reiterant speech. *Journal of the Acoustical Society of America*, *99*(4), 2493-2500.

Tajima, K., Port, R., & Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign-accented English. *Journal of Phonetics*, *25*, 1-24.

Tamburini, F. (2009). Prominenza frasale e tipologia prosodica: un approccio acustico. In G. Ferrari, R. Benatti, & M. Mosca (Eds.), *Linguistica e modelli tecnologici di ricerca. atti del xl congresso internazionale di studi della SocietàĂ di Linguistica Italiana* (p. 437-455). Rome: Bulzoni.

Tancredi, C. (1992). *Deletion, deaccenting and presupposition.* Doctoral dissertation, M.I.T.

Terken, J. (1991). Fundamental frequency and perceived prominence. *Journal of the Acoustical Society of America*, *89*, 1768-1776.

t'Hart, J., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation.* Cambridge: Cambridge University Press.

Thompson, I. (1991). Foreign accents revisited: The English pronunciation of russian immigrants. *Language Learning*, *41*(2), 177-204.

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, *28*(1), 1-30.

Trouvain, J., & Gut, U. (2007). *Non-native prosody: phonetic description and teaching practice.* Berlin: Mouton de Gruyter.

Ueyama, M. (2012). *Prosodic transfer: An acoustic study of L2 English and L2 Japanese.* Bologna: Bononia University Press.

Ueyama, M., & Jun, S.-A. (1998). Focus realization in Japanese English and Korean English intonation. In H. Hajime (Ed.), *Japanese and Korean linguistics* (p. 629-645). CSLI: Stanford University Press.

Urbani, M. (2013). *The pitch range of Italians and Americans. a comparative study.* Doctoral dissertation, University of Padua.

Vaissière, J. (2005). Perception of intonation. In D. Pisoni & R. Remez (Eds.), *The handbook of speech perception* (p. 236-263). Oxford: Blackwell.

Vallduvi, E. (1991). The role of plasticity in the association of focus and prominence,. *Proc. Eastern States Conference on Linguistics (ESCOL)*, *7*, 295-306.

Van Els, T., & de Bot, K. (1987). The role of intonation in foreign accent. *The Modern Language Journal*, *71*(2), 147-155.

Van Heuven, V. J. (1994). Introducing prosodic phonetics. In C. Odè & V. J. Van Heuven (Eds.), *Phonetic studies of indonesian prosody* (p. 1-26). Leiden: LOT Publications.

Volìn, J., & Skarnitzl, R. (2010). The strength of foreign accent in czech English under adverse listening conditions. *Speech Communication*, 1010-1021.

Wagner, P. (2005). Great expectations. introspective vs. perceptual prominence ratings and their acoustic correlates. *Proc. Interspeech 2005, Lisbon, Portugal*, 2381-2384.

Wang, H., Zhu, L., Li, X., & Van Heuven, V. J. (2011). Relative importance of tone and segments for the intelligibility of Mandarin and cantonese. *Proc. 17th International Conference of Phonetic Sciences (ICPhS), Hong Kong, China*, 2090-2093.

Wayland, R. (1997). Non-native production of thai: Acoustic measurements and accentedness ratings. *Applied Linguistics*, *18*(3), 345-373.

Wells, J. C. (1962). *A study of the formants of the pure vowels of British English.* Master's dissertation.

Wells, J. C. (2006). *English intonation. an introduction.* Cambridge: Cambridge University Press.

Wightman, C. W. (2002). Tobi or not tobi? *Proc. Speech Prosody, Aix-en-Provence, France*, 25-29.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics*, *27*, 55-105.

Xu, Y. (2011a). Speech prosody: a methodological review. *Journal of Speech Sciences*, *1*, 85-115.

Xu, Y. (2011b). Post-focus compression: Cross-linguistic distribution and historical origin. *Proc. 17th International Conference of Phonetic Sciences (ICPhS), Hong Kong, China*, 152-155.

Xu, Y., & Xu, C. M. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*, *33*, 159-197.

Yoon, K. (2007). Imposing native speakers' prosody on non-native speakers' utterances: The technique of cloning prosody. *Journal of the Modern British & American Language & Literature*, *25*(4), 197-215.

Zipp, L., & Dellwo, V. (2011). Reading-speech-normalization: A method to study prosodic variability in spontaneous speech. *Proc. 17th International Conference of Phonetic Sciences (ICPhS), Hong Kong, China*, 2328-2331.