Localising foreign accents in speech perception, storage and production

Yuki Asano
Department of Linguistics, Faculty of Humanities
University of Konstanz, Germany

This dissertation is submitted for the degree of
Doctor of Philosophy

2016

ACKNOWLEDGEMENTS

For their insightful comments, suggestions, questions, ideas, for this thesis

Bettina Braun
Ryoko Hayashi
Aditi Lahiri
René Kager

For making my research project financially possible

Bettina Braun
The German Academic Exchange Service (DAAD)
The German National Academic Foundation
University of Konstanz

For their general interest, comments, and inspiring ideas

Bettina Braun
Miriam Butt
Nicole Dehé
Bijarke Frellesvig
Rolf Grawert
Janet Grijzenhout
Michele Gubian
Carlos Guessenhoven
Ryoko Hayashi
Bariş Kabak
René Kager

Aditi Lahiri
Frans Plank
Dominik Sasha
Giuseppina Turco

For correcting my English

Filippo Cervelli

For their technical support

Michele Gubian
Joachim Kleinmann
Dominik Sasha
Research assistants

For their attendance of my experiments

Participants

For trusting in my research potential

Bettina Braun
Masahiko & Kyoko Asano

For their mental support

Masahiko & Kyoko Asano
Bettina Braun
Emilia & Filippo Cervelli
Hannelore & Rolf Grawert

Thank you very much!

ZUSAMMENFASSUNG

Die Dissertationsschrift beschäftigt sich mit der Suche nach möglichen Quellen eines fremdsprachlichen Akzents in der Prosodie, insbesondere in F_0 und segmentaler Länge. Anhand der in den Kapiteln 2 bis 4 dargestellten Experimente wurde überprüft, ob ein fremdsprachlicher Akzent seine Ursache in der Perzeption, in der Speicherung oder in der Produktion der fremdsprachlichen Prosodie findet. Die Untersuchung konzentriert sich auf die Sprachen Japanisch (als Zielsprache, L2) und Deutsch (als Ausgangssprache, L1), da diese ein kontrastives prosodisches System aufweisen: In Bezug auf F_0, is im Japanischen ein Tonhöhenakzent lexikalisch, während es im Deutschen mit einer post-lexikalischen oder paralinguistischen Bedeutung verwendet wird. In Bezug auf die segmentale Länge, zeichnet sich Japanisch sowohl durch seinen vokalischen als auch durch seinen konsonantischen lexikalischen Kontrast aus, während Deutsch nur einen begrenzten vokalischen Kontrast aufweist. Deshalb interessiert in dieser Studie das Erlernen von lexikalischem Tonhöhenakzenten und von lexikalischen konsonantischen Längekontrasten bei deutschen Japanischlernenden.

Ausgangspunkt der Untersuchung ist die Feststellung der nicht-normentsprechenden Produktionen sehr häufig verwendeter japanischer Wörter im Experiment 1 (im Kapitel 2), die bei deutschen Lernenden auftraten. Im Experiment wurde ein halb-spontanes Produktionsexperiment bei deutschen Japanischlernenden und japanischen Muttersprachlern durchgeführt. Dabei produzierten die Teilnehemenden japanische und deutsche Wörter (*Sumimasen, Konnichiwa* und *Entschuldigung*, alle bedeutend „Entschuldigung", um jemanden zu rufen) in gegebenen fiktiven Situationen und wiederholten sie dieselben Wörter dreimal. Die Analyse der Realisierung des japanischen lexikalischen Tonhöhenakzents und der segmentalen Länge zeigte, dass die deutschen Japanischlernenden 1) den japanischen lexikalischen fest definierten Tonhöhenakzent phonologisch stets variierten und 2) die japanische segmentale Längestruktur nicht normentsprechend produzieren konnten. Diese Abweichungen der L2-Produktionen von den L1-Produktionen führten zu der Annahme, dass die L2-

Lernenden entweder L2-prosodische Information anders hörten (= Schwierigkeiten in der Anfangsphase der Sprachperzeption) oder diese in ihrem mentalen Lexikon anderes speicherten als L1-Sprecher (= Schwierigkeiten, die mit mentalen Repräsentationen verbunden sind). Oder möglicherweise hatten sie Schwierigkeiten bei der Artikulation. Diese drei Etappen der Sprachverarbeitung wurden Schritt für Schritt in den Experimenten 2, 3 und 4 genauer untersucht. Die Sprachperzeption sowie die Sprachproduktion, die entweder den Zugriff auf die mentalen Repräsentationen erforderten, oder die, das nicht notwendigerweise erforderten, wurden durch die Manipulation der kognitiven Belastung des Arbeitsgedächtnisses im Hinblick auf die Speicherauslastung getestet. Darüber hinaus wurden aufgaben-irrelevante prosodische Dimensionen den Stimuli hinzugefügt, um zu testen, ob die erhöhte Steuerung der Aufmerksamkeit zur Instabilität der L2-Sprachverarbeitung im Vergleich zu der L1 Sprachverarbeitung führen würde. Gleiche Stimuli wurden in den folgenden Experimenten verwendet. Dabei wurden dieselben Teilnehmenden getestet.

In den Experimenten 2 und 3 (im Kapitel 3 und 4) wurden zwei der drei oben genannten Etappen, nämlich Sprachperzeption und mentale Repräsentationen untersucht. Genauer gesagt, ging es um die Frage, ob L2-Lernende Schwierigkeiten haben, akustische Korrelate eines prosodischen Kontrastes wahrzunehmen oder diese phonologisch zu speichern und abzurufen. Im Experiment 2 wurden segmentale Längekontraste (vokalische und konsonantische Längekontraste) und im Experiment 3 Kontraste eines Tonhöhenakzentes (flacher vs. fallender F_0) untersucht. In beiden Experimenten wurden jeweils AX (same-different)-Diskriminationsaufgaben durchgeführt, bei denen die Teilnehmenden entweder eine Antwort „gleich" oder „unterschiedlich" zu jedem Stimuluspaar gegeben hatten. Eine Besonderheit der beschriebenen Experimenten bestand darin, dass der Zeitabstand zwischen den zwei Stimuli (A und X) (= Interstimulus-Intervall, ISI) variiert wurde (300 ms vs. 2500 ms). Dabei wurde angenommen, dass die Bedingung mit einem kürzeren ISI die Sprachperzeption des akustischen Korrelates des Kontrastes testete, während die Bedingung mit einem längeren ISI die Sprachperzeption testete, die mentalen Repräsentationen in einem größeren Ausmaß mit aktiviert hat. Unter der Bedingung mit einem längeren ISI hätten die phonetischen Informationen des ersten Stimulus gespeichert werden müssen, um diese mit denen des zweiten Stimulus vergleichen zu können. Diese Annahme basiert auf der Theorie des Arbeitsgedächtnises, die besagt, dass eine phonetische Information nach etwa 2 Sekunden verloren geht. Zusätzlich wurde die Komplexität der Stimuli dadurch erhöht, dass aufgaben-irrelevante flache und fallende Tonhöhebewegung für die Diskriminierung der segmentalen Längekontras-

te (im Experiment 2) und aufgaben-irrelevante segmentale Länge für die Diskriminierung der Tonöhenkontraste (im Experiment 3) hinzugefügt wurden. Getestet wurden 24 japanische Muttersprachler, 48 deutsche L2-Lernende (des Japanischen) und 24 deutsche Nicht-Lernende (=naïve Hörer). Analysiert wurden die *d*'-Werte (das Maß für die Sensitivität zu Kontrasten) und Reaktionszeiten.

Die Ergebnisse des Experimentes 2 zeigten, dass die *d*'-Werte und Reaktionszeiten der japansichen Muttersprachlern in allen experimentellen Bedingungen konstant gleich hoch waren. Was die L2-Lernenden und die Nicht-Lernenden anbetrifft zeigten sie unter den Bedingungen mit niedrigsten Aufgabenanforderungen (mit kürzerem ISI, ohne aufgaben-irrelevante fallende Tonhöhebewegung) genauso hohe Sensitivität für den nicht-muttersprachlichen konsonantischen Längekontrast wie japanische Muttersprachler. Sogar die Nicht-Lernenden konnten dies unterscheiden, weil sich der Kontrast durch den Vergleich auf der phonetischen Ebene erkennen ließ. Allerdings konnte eine solche Abhängigkeit vom phonetischen Vergleich nicht lange dauern. Sobald das ISI länger wurde und die Speicherauslastung höher wurde, so dass die phonetische Information mehr phonologisch verarbeitet werden musste, verringerte sich die Diskriminationsfähigkeit der L2-Lernenden und der Nicht-Lernenden. Die *d*'-Werte der L2-Lernenden verringerten und unterschieden sich von denen der Muttersprachler, und die der Nicht-Lernenden sanken deutlich, so dass sich die Werte der L2-Lernenden und Nicht-Lernenden voneinander unterschieden. Wenn die aufgaben-irrelevante Tonhöhebewegung ins Spiel kam unterschieden sich die *d*'-Werte der L2-Lernenden und Nicht-Lernenden bereits in der kürzeren ISI-Bedingung von denen der japanischen Muttersprachlern, und dies galt auch unter der längeren ISI-Bedingung. Das gleiche Ergebnis wurde in der Analyse der Reaktionszeiten gefunden. Die Reaktionszeiten der Nicht-Lernenden unterschieden sich nicht von denen der japanischen Muttersprachler in der flachen Tonhöhe- und kürzeren ISI-Bedingung. In der längeren ISI-Bedingung wurden nur die Reaktionszeiten der Nicht-Lernenden länger. In der fallenden Tonhöhebedingung unterschieden sich die Reaktionszeiten der drei Gruppen bereits in der kürzeren ISI-Bedingung voneinander. Zusätzlich hat die Analyse der Reaktionszeiten gezeigt, dass die L2-Lernenden für die Entscheidung generell länger brauchten als die Nicht-Lernenden. Die Reaktionszeiten der Nicht-Lernenden waren genauso kurz wie die der japanischen Muttersprachler wenn die aufgaben-irrelevante Tonhöhebewegung nicht vorhanden war. Mit der aufgaben-irrelevanten fallenden Tonhöhebedingung verlängerten sich die Reaktionszeiten von den Nicht-Lernenden, so dass sie sich von denen der japanischen Muttersprachler unterschieden. Der Vergleich zwischen den Er-

gebnissen der flachen und fallenden Tonhöhebedingungen zeigte eine konsistente Wirkung der aufgaben-irrelevanten Tonhöhebewegung auf die Diskriminierung des nichtmuttersprachlichen segmentalen Längekontrastes: Der konsonantische Längekontrast dargestellt mit der fallenden Tonhöhenbewegung bereitete den Lernenden und den Nicht-Lernenden größere Schwierigkeiten als der konsonantische Kontrast mit der flachen Tonhöhenbewegung. Die japanischen Muttersprachler wurden von der aufgabenirrelevanten prosodischen Dimension nicht betroffen.

Zusammengefasst legen diese Ergebnisse nahe, dass die Exposition gegenüber der L2 den L2-Lernenden half, die phonologische Repräsentationen der nichtmuttersprachlichen konsonantischen Längekontraste herzustellen (da die Lernenden durch die erhöhte Speicherauslastung in geringerem Maße betroffen wurden). Jedoch wurden sowohl die L2-Lernende als auch die Nicht-Lernende stark durch die aufgabenirrelevante prosodische Dimension beeinträchtigt. Ihre Diskriminationsfähigkeit wurde durch die höheren Anforderungen an die Steuerung der Aufmerksamkeit gestört. Für die Lernenden und Nicht-Lernenden war es schwierig, die aufgaben-irrelevante Tonhöhebewegung zu ignorieren und ihre Aufmerksamkeit nur auf die aufgabe-relevanten Informationen zu konzentrieren. Das Ergebnis zeigt die Schwierigkeit, die L2-Sprachperzeption zu stabilisieren, auch wenn L2 phonologische Repräsentationen aufgebaut wurden.

Das Experiment 3 testete die Diskriminationsfähigkeit des Tonhöhenkontrastes. Die Ergebnisse des Experimentes 3 zeigten, dass die *d'*-Werte der japanischen Muttersprachler höher waren als die von den L2-Lernenden, gefolgt von denen der Nicht-Lernenden. Die Reaktionszeitanalyse zeigte kürzere Reaktionszeiten für die japanischen Muttersprachler als für die Nicht-Lernenden gefolgt von den L2-Lernenden. Die Reaktionszeiten der japanischen Muttersprachler wichen unter beiden ISI-Bedingungen nicht ab. Ein Effekt der aufgaben-irrelevanten segmentalen Längestrukturen auf die Diskrimination des Tonhöhekontrastes wurde in den Reaktionszeiten der Lernenden und Nicht-Lernenden gefunden. Ihre Reaktionszeiten wurden unter der längeren ISI-Bedingung länger, mit Ausnahme der Langvokal-Paare, die phonologisch dem Deutschen ähnlich waren. Die Ergebnisse deuten an, dass Muttersprachler mit lexikalischen Tonhöhekontrasten eine höhere Sensitivität für akustische Korrelate der Tonhöhekontraste bilden. Die Unterschiede zwischen den japanischen Muttersprachlern und den beiden Nicht-Muttersprachlern wurden bereits unter der kürzen ISI-Bedingung gefunden. Unter der längeren ISI-Bedingung, unter der phonologische Repräsentationen der Kontraste in einem größeren Ausmaß aktiviert wurden, nahm die Leistung der beiden Gruppen der

Nicht-Muttersprachler nur dann ab, wenn die Paare mit fremden segmentalen Längestrukturen dargestellt wurden, die jedoch aufgaben-irrelevant waren.

Im Experiment 4 (im Kapitel 5) handelte es sich um eine unmittelbare und eine verzögerte Imitationsaufgabe. Dabei wurde untersucht, ob die L2-Lernenden Schwierigkeiten bei der Produktion nicht-muttersprachlicher segmentaler Länge und Tonhöhekontraste haben. Die unmittelbare Imitationsaufgabe wurde eingerichtet, um die Fähigkeit zu testen, die Stimuli ohne Vermittlung der phonologischen Repräsentationen zu imitieren. Die verzögerte Imitationsaufgabe (die Verzögerungszeit war 2500 ms) erforderte hingegen den Zugriff auf die phonologischen Repräsentationen. Die Korrektheit der Imitation wurde durch die Messung der Dauerverhältnisse von kurzen und langen Konsonanten und Vokalen sowie der Steilheit des Tonfalls analysiert. Die Analyse der konsonantischen Dauerverhältnisse zeigte, dass sich die Dauerverhältnisse der japanischen Muttersprachler nicht von Dauerverhältnissen der Stimuli unterschieden, während die Dauerverhältnisse der L2-Lernenden und der Nicht-Lernenden bereits in der unmittelbaren Imitationsaufgabe kleiner waren als die Dauerverhältnisse der Muttersprachler. Die konsonantischen Dauerverhältnisse der Lernenden waren größer als die Dauerverhältnisse der Nicht-Lernenden; das kann man als einen positiven L2-Lerneffekt ansehen. Im Gegenteil, die vokalischen Dauerverhältnisse zeigten, dass die Lernenden und die Nicht-Lernenden größere Dauerverhältnisse als die japanischen Muttersprachler produzierten. Die nicht-muttersprachlichen Sprecher übertrieben den Kontrast, wie dieser ihnen aus ihrer Muttersprache vertraut war.

Die Analyse der Tonsteilheit zeigte eine generell größere Steilheit bei den japanischen Muttersprachlern als bei den beiden deutschen Gruppen. Die artikulatorische Schwierigkeit wurde bereits in der unmittelbaren Imitation gefunden. Das Ausmaß, inwieweit die nicht-muttersprachlichen Sprecher die Stimuli korrekt imitieren konnten, war von den segmentalen Längestrukturen abhängig, da dieser Faktor Einfluss darauf hatte, inwieweit die Nicht-Muttersprachler die phonetischen Details der Stimuli beibehalten konnten, wenn auditorische Informationen nach einer Weile verloren gingen.

Das Ziel dieser Arbeit war es, die Quellen der Schwierigkeiten festzustellen, die bei den Produktionen der L2-Prosodie gefunden wurden. Die Sprachverarbeitung von nicht-muttersprachlichen prosodischen Kontrasten in der Anfangsphase der Sprachperzeption, in der Verarbeitung, die mit mentalen Repräsentationen verbunden ist, und in der Artikulation wurde durch die Manipulation der kognitiven Belastung des Arbeitsgedächtnisses im Hinblick auf die Speicherauslastung und Aufmerksamkeitskontrolle getestet. Die Erkenntnisse aus den Experimenten führte mich zu folgenden Schlussfolgerungen.

Erstens waren die Nicht-Muttersprachler in der Anfangsphase der Sprachperzeption erfolgreich, wenn die Aufgabenanforderungen die niedrigste waren. Sobald die kognitive Belastung erhöht wurde, nahm ihre Leistung ab. In der Artikulationsphase unterschieden sich die Nicht-Muttersprachler bereits von den Muttersprachlern (und von den Stimuli). Deshalb legen diese Ergebnisse nahe, dass die Quellen für einen fremdsprachlichen Akzent in der Verarbeitung, die mit mentalen Repräsentationen zu tun hat, und in der Artikulation zu finden sind. Zweitens wurde durch die Manipulation der kognitiven Belastung gezeigt, dass die Verarbeitung der L2-Prosodie unter den erhöhten kognitiven Belastung instabil wurde, während die der L1-Prosodie stabil blieb. Aufgrund des Mangels an phonologischen Repräsentationen in der L2 war das Ausmaß der Gedächtniskapazität der Nicht-Muttersprachler bei der L2-Verarbeitung kleiner. Außerdem wurde ihre Sprachverarbeitung durch ein erhöhtes Ausmaß der Aufmerksamkeitskontrolle leicht beeinträchtigt. Dies lässt darauf schließen, dass es schwierig war, die L2-Verarbeitung zu automatisieren. Drittens schnitten die Leistung der L2-Lernenden in der Regel besser ab als die der Nicht-Lernenden. Darin kann man einen positiven Lerneffekt erkennen. Jedoch zeigten die Reaktionszeitanalysen, dass die Nicht-Lernenden schneller waren als die Lernenden, was nahelegt, dass sie verschiedene Strategien für die Aufgabe verwendeten. In der Sprachverarbeitung der L2-Lernenden konkurrierten die L1- und L2-Repräsentationen miteinander, so dass sie vermutlich längere Zeit für eine Antwort brauchten. Viertens hat die Studie gezeigt, dass es keine allgemeine „prosodische Verarbeitung" gibt, die für alle Arten von prosodischen Eigenschaften maßgebend ist. Die aufgaben-irrelevante segmentale Länge und die aufgaben-irrelevante Tonhöhebewegung zeigten unterschiedliche Effekte. Das deutet darauf hin, dass die zugrundeliegenden Verarbeitungsmechanismen der segmentalen Länge und der Tonhöhebewegung anders waren. Prosodische Eigenschaften weisen in verschiedenen Sprachen auf verschiedenen linguistischen Ebenen unterschiedliche Kombination auf. In dieser Studie wurde gezeigt, dass eine prosodische Eigenschaft nicht erfolgreich von einer L1 zu einer L2 übertragen werden kann, wenn die Eigenschaft in der L1 und der L2 auf unterschiedlichen linguistischen Ebenen Kontraste bilden.

Über die theoretischen Thesen dieser Arbeit hinaus tragen deren Ergebnisse praktisch anwendbare Erkenntnisse für einen Fremdsprachenunterricht. Ein Beispiel ist, dass die gefundene instabile L2-Verarbeitung darauf hinweist, dass es wichtig ist, eine klassische Diskriminationsaufgabe mit Minimalpaaren, die oft in einem Unterricht des Japanischen als L2 durchgeführt wird, unter verschiedenen störenden Faktoren, die die kognitive Belastung erhöhen, zu präsentieren. So könnte man Minimalpaare in einen Satz

einfügen oder mit den störenden Hintergrundgeräuschen oder mit variierten Sprechgeschwindigkeiten verbinden.

Im Bereich der psycholinguistischen Untersuchung der L2-Prosodie ist Folgendes festzuhalten: Erstens hat die Querschnittsuntersuchung durch die Verwendung der gleichen Stimuli die Beziehungen zwischen Sprachperzeption, die mentalen Repräsentationen und die Sprachproduktion in einer L1 und L2 erfasst. Zweitens betraf die Analyse mehrere prosodische Eigenschaften (F_0 und segmentaler Länge), während die meisten Studien auf diesem Gebiet nur eine Eigenschaft innerhalb einer Studie thematisierten. Drittens zeigte die Studie die Instabilität der L2-Verarbeitung durch die Manipulation der experimentellen Variablen. Sie unterstrich die Bedeutung der Rolle der kognitiven Belastung, die beim Erlernen einer L2 zu berücksichtigen ist. Die Untersuchung der Folgen einer zunehmenden kognitiven Belastung durch mehrere experimentelle Variablen wurde hier erstmals unternommen. Viertens hat es die Verwendung von (kognitiv gesehen) einfachen experimentellen Aufgaben (Diskriminationsaufgaben und Imitationsaufgaben) ermöglicht, die Aufgabenanforderungen und infolgedessen das Ausmaß der kognitiven Belastung zu variieren. Diese experimentelle Methode konnte fein abgestufte Unterschiede bei der Verarbeitung von verschiedenen prosodischen Eigenschaften in der L1- und L2-Verarbeitung aufweisen. Fünftens wurden im Rahmen ein und derselben Studie zwei prosodische Eigenschaften untersucht. Das erste Experiment testete die Integration der beiden Eigenschaften in der Sprachproduktion. Die anderen Experimente testeten, wie sich diese Eigenschaften voneinander trennen lassen und wie sie einander beeinflussen. Da jede Sprache eine fast einzigartige obengenannte Kombination der prosodischen Eigenschaften („welche Eigenschaft auf welcher linguistischen Ebene") bildet, ist nicht nur die Untersuchung über die Unterschiede der gleichen prosodischen Eigenschaften zwischen einer L1 und L2, sondern auch die Unterschiede zwischen mehreren prosodischen Eigenschaften innerhalb einer Sprache für weitere theoretische Beiträge von Bedeutung.

CONTENTS

Contents **xiii**

Nomenclature **xv**

1 General introduction **1**
1.1 Rationale of the thesis 1
1.2 Outline of the dissertation 4
1.3 Languages in focus: Japanese and German 6
1.3.1 Pitch 6
1.3.2 Segmental length and rhythm 9
1.4 Prosodic transfer in the models and theories of L2 acquisition 10
1.4.1 Influential theories and models up to date 10
1.4.2 Possible extensions and limitations to L2 prosodic research 16
1.5 Stages of speech processing under investigation 21
1.6 Cognitive load to understand L2 processing 30
1.6.1 Working memory 30
1.6.2 Factors affecting cognitive load in L2 processing 34
1.7 Summary 35

2 Coordinating lexical and paralinguistic use of F_0 in L2 production **37**
2.1 Introduction 37
2.2 Experiment 41
2.2.1 Methods 41
2.2.2 Results 45
2.3 Discussion 55

3 Discrimination of nonnative segmental length contrasts **59**
3.1 Introduction 59

3.2 Experiment . . . 67
3.2.1 Methods . . . 67
3.2.2 Results . . . 73
3.3 Discussion . . . 80

4 Discrimination of pitch contrasts 87
4.1 Introduction . . . 87
4.2 Experiment . . . 95
4.2.1 Methods . . . 95
4.2.2 Results . . . 97
4.3 Discussion . . . 102

5 Immediate and delayed imitation of segmental length contrasts and pitch falls 109
5.1 Introduction . . . 109
5.2 Experiment . . . 117
5.2.1 Methods . . . 117
5.2.2 Results . . . 119
5.3 Discussion . . . 123

6 General discussion and outlook 127
6.1 General discussions . . . 128
6.1.1 Summary of the results . . . 128
6.1.2 The relationships between the stages of speech processing . . . 132
6.1.3 F_0 and segmental length contrasts . . . 134
6.1.4 Lexical, post-lexical and paralinguistic prosody (F_0) . . . 138
6.1.5 Cognitive load . . . 139
6.1.6 Definition of (di)simmilarities of a cross-linguistic prosodic contrast 140
6.2 Outlook . . . 142
6.3 An exploratory example: data-driven analyses of F_0 . . . 143
6.3.1 FPCA . . . 144
6.3.2 SOM . . . 148
6.3.3 Evaluation of the methods . . . 151
6.4 Conclusions . . . 154

Bibliography 157

Appendix A Rapid and Smooth Pitch Contour Manipulation 183

Appendix B Participants' demographics (Experiment 2–4) **191**

Appendix C Results (Experiment 2–4) **193**

Appendix D *M*odel specifications in the statistical analyses **199**

CHAPTER 1

GENERAL INTRODUCTION

1.1 Rationale of the thesis

The difficulties to acquire second language (henceforth L2) sounds and sound organisation are manifested in an immediately identifiable foreign accent retained by almost every adult L2 learner[1] Such a foreign accent is also observed in otherwise highly proficient L2 speakers who mastered the grammatical system very well. This is known as "Joseph Conrad Phenomenon" (Bongaerts, 1999; Bongaerts et al., 1995; Guiora, 1990) referring to the Polish-born novelist's native-like abilities in English grammar, vocabulary and writing style being opposed to his strong foreign accent. One of the most extensively examined factors causing a foreign accent is negative language transfer from one's L1 (Lado, 1957). Negative transfer in foreign accent is not limited to the acquisition of novel sounds (for instance the distinction between /r/ and /l/ by Japanese learners of German, e.g. Takagi, 2002), but also extends to the prosodic domain.

In the following, I use the term "prosody" referring to the set of features consisting of pitch, stress and quantity and to the phonological organisation of individual sounds (i.e., segments) into higher-level constituents, which is manifested by variation of F_0, duration, amplitude and segment quality (Shattuck-Hufnagel and Turk, 1996; Ueyama, 2000).

[1] In my thesis, the term "L2" is used as an umbrella term for any language learned in addition to one's first language (henceforth L1) including also nonnative language, foreign language, secondary language and weaker language, while "L1" includes native language, mother tongue, primary language and stronger language, following the distinctions made by Stern (1983). These two sets of terms indicate a subjective relationship between a language and an individual or a group (Stern, 1983, 9) and implies that an L1 is available prior to an L2. It also refers to the lower level of language proficiency and command in comparison with an L1. In the same way, an L2 learner refers to someone who has learned/ is learning an L2 after having acquired the L1 and who passed a "critical period", the time window in which a language develops readily and after which its acquisition is much more difficult and ultimately less successful (e.g. Lenneberg, 1967; Scovel, 1988).

I am aware of the term "suprasegmentals" that can be used as a synonym according to this definition of prosody. In this thesis, I will use the term "prosody" and not "suprasegmentals" for the following reason: The term "suprasegmentals" is exclusively used to refer to the features whose domain extends over more than one segment (Lehiste, 1969) in contrast to "segmentals" taking a segmental phonetic idealisation as a starting point (Trager and Bloch, 1941). However, the distinction between suprasegmentals and segmentals poses problems to explain some phenomena such as a lexical tone or features of place, voicing or nasality. A lexical tone is categorised as "suprasegmentals", but appears in a syllable consisting of one vowel, which is exactly one segment. Features of place, voicing or nasality are treated as segmentals, but can apply to two or three successive segments, namely at the "suprasegmental" level (Ladd, 2015, 70).

Previous studies demonstrate that foreign accent with deviant L2 prosody affects our communication and thus appropriate prosody is important for successful communication. For example, L1 speakers experience difficulties in comprehending L2 speakers with deviant L2 prosody (Braun et al., 2011; Bürki-Cohen et al., 2011; Gutknecht, 1979; Holm, 2007; Munro and Derwing, 1995a,b). Braun et al. (2011) conducted word-monitoring, lexical decision and semantic categorisation tasks by Dutch participants presenting Dutch sentences with normal intonation contours and with unfamiliar ones. In all tasks, it was found that the unfamiliar intonation contour slowed the participants' response down. Their findings demonstrate that an unfamiliar intonation contour hinders lexical access and negatively affects speech comprehension. Bürki-Cohen et al. (2011) also conducted a series of monitoring experiments. Their major results demonstrate that the processing of L2 speech under adverse listening conditions is aggravated than that of L1 speech. A sentence verification task conducted by Munro and Derwing (1995b) also showed that L1 listeners generally took longer to verify the truth value of sentences spoken by L2 speakers than the same sentences spoken by L1 speakers, even though both types of speech were highly intelligible. All these empirical findings consistently support the claim that L2 accented speech is more difficult to process for L1 listeners. Moreover, the importance of the accurate L2 prosodic forms in speech comprehension is not limited to the intelligibility of L2 speech, but it is even claimed that deviant prosodic forms contribute relatively more to the impression of foreign accent than segmental accuracy (e.g., Anderson-Hsieh and Koehler, 1988; Johansson, 1978; Pennington and Richards, 1986).

As for speech production, the presence of a foreign accent distinguishes the L2 learner from the L1 speaker of a language regardless of one's desires and it may lead to

negative attitudes or even social exclusion (Gluszek and Dovidio, 2010; Hirschfeld and Trouvain, 2007). This is because the individual way of speaking conveys a speaker's personality, from which L1 listeners deduce the educational status, the social affiliation, the degree of intelligence and even certain traits of the individual character (Hirschfeld, 1994; Hirschfeld and Trouvain, 2007).

This wide range of communicative and attitudinal impediments is caused by the fact that prosodic properties deliver us not only lexical or post-lexical but also paralinguistic information, and even extralinguistic information. Post-lexical information refers to information in a domain of phonology that may not interact with morphological rules and is ordered after the rules of syntax (Kaisse, 1984; Kaisse and Shaw, 1985), while paralinguistic information refers to a speaker's emotion and intention excluding non-linguistic features as those that cannot be used intentionally (Liscombe, 2007) and extralinguistic information refers to demographic and cultural information about a speaker (Chun, 2002; Couper-Kuhlen, 1986; Liscombe, 2007).

Recently, minimising prosodic interference has begun to be an important issue in L2 teaching (Mennen, 2007). More and more teachers and teaching materials emphasise the importance of acquiring prosody (Hirschfeld and Trouvain, 2007). However, only a limited number of studies have so far investigated foreign accent manifested in deviant L2 prosodic forms (e.g., Chen and Mennen, 2008 and Flege et al., 1995b for L2 English - L1 Italian; Gårding, 1981 for L2 French - L1 Swedish and Greek; Jilka et al., 2007 for L2 English - L1 German; Jun and Oh, 2000 for L2 Korean - L1 English; Mennen, 1998, 2004 for L2 Greek - L1 Dutch; Mennen et al., 2010a for L2 English - L1 Punjabi or Italian; Ueyama and Jun, 1998 for L2 English - L1 Korean or Japanese). It is notable that most of them investigated English as L2 and analysed learners' foreign accent in their L2 productions.

Despite the documentation of learners' deviant production in L2 prosody in previous studies, little is known yet about the question where foreign accent comes from - about the stages in L2 processing that contribute to the deviant forms. The learners' difficulties might relate to the lack of ability to perceive acoustic correlates of nonnative prosodic information in sensory memory (e.g. Atkinson and Shiffrin, 1968). Difficulties might otherwise relate to the failure in maintaining the prosodic information in short-term memory storage by communicating with their long-term mental representations of (lacking) L2 categories, i.e., in working memory[2] (e.g. Atkinson and Shiffrin, 1968; Bad-

[2] In this thesis, short-term memory merely refers to the short-term storage of information in a theory-neutral manner without entailing the manipulation or organisation of information in memory (Cowan, 2008). On the contrary, the term working memory implies complex cognitive activities such as the moment-to-moment monitoring processing and its rehearsal system by communicating

deley, 2003). Or it could be also the case that they have difficulties in articulating such a contrast, which does not relate to the lack of mental representations of L2 categories. In this dissertation, these three stages of L2 processing are called "input", "mental representations" and "output" stage respectively, following Sakamoto (2010). This thesis investigates possible sources of foreign accented speech, testing each of these stages of L2 speech processing. In order to try to separate these stages, I conducted perception and production experiments varying memory load that is required for the task in order to manipulate to what extent the tasks involved phonetic and phonological processing. By using the same sound stimuli and by testing the same participants in perception and production experiments, I aimed at comparing the results of their speech processing in the "input", "mental representations" and "output" stage and at analysing the relationships between them.

To achieve these goals, I examined the L2 acquisition of Japanese prosody, in particular the acquisition of nonnative lexical consonant length contrasts and pitch contrasts by German learners and non-learners (= naïve speakers/listeners). These two prosodic phenomena are "nonnative" in the sense that they are *not lexically used* in German. Hereafter, the adjective "nonnative" is used to describe a prosodic phenomenon that does not share the same linguistic function or category in an L1 and an L2 or that does not exist in either an L1 or an L2 instead of using "L2"(, because the latter does not always mean that something is nonnative).

Japanese and German constitute a contrastive language pair in terms of their prosodic systems. For instance, in Japanese, pitch and consonant length have primarily lexical functions, while in German they are not lexically contrastive. These functional differences of the same prosodic properties in the two languages were exploited for the experimental design. I will also test two groups of German L1 speakers - German learners of Japanese and non-learners - in comparison with Japanese L1 speakers. Testing both learners and non-learners makes it possible to examine L2 learning effects.

1.2 Outline of the dissertation

The chapters of this thesis are organised as follows: Chapter 1 gives a description of the theoretical framework and motivates the experimental design of the thesis. Then, Chapter 2 describes the account of a production experiment in order to document foreign ac-

long-term memory (Baddeley and Logie, 1999) in the phonological short-term memory, see details Subsection 1.6.1.

cents in prosody by German L2 learners of Japanese. Then, there are three chapters covering the experiments in which the same groups of participants were tested in different tasks. Chapter 3 presents a perception experiment, in which the discrimination ability of nonnative consonant length contrasts was examined. The listeners' cognitive load was increased by using a longer duration of inter-stimulus interval (= ISI) (2500 milliseconds, henceforth ms vs. 300 ms) and the demand on attention control was increased by adding psychoacoustic complexity of the stimuli (trials with task-irrelevant pitch falls vs. with monotonous flat pitch). The two ISI conditions were expected to manipulate the memory systems involved (Burnham and Francis, 1997; Wayland and Guion, 2004; Werker and Tees, 1984a). The shorter ISI condition was expected to involve more language-general phonetic processing in sensory memory (Atkinson and Shiffrin, 1968) and the longer ISI condition more language-specific phonological one in working memory (e.g. Baddeley, 2003). Chapter 4 presents another perception experiment that examined the discrimination ability of pitch contrasts. The same participants took part in the experiment and the same methodological paradigm was used as in Experiment 2. In order to increase the demand on attention control, the stimuli pairs were presented in native and nonnative segmental length structures that were task-irrelevant for the discrimination of pitch contrasts.

Chapter 5 presents a production experiment, in which the same participants imitated stimuli with nonnative and native segmental length structures and pitch contours either immediately after the stimuli or after a pause of 2500 ms. The immediate imitation task tested the articulation ability that is supposed not to necessarily require to access long-term mental representations while the delayed imitation task tests the production ability involving the access to long-term mental representations. The series of experiments in Chapter 3 to Chapter 5 allowed us to examine whether or not the learners' problems in their speech productions related to their ability to discriminate the acoustic correlates of the target L2 prosodic contrasts (corresponding to the "input" stage) or to access their phonological categories (corresponding to the "mental representations" stage) or to articulate them (corresponding to the "output" stage).

Chapter 6 presents a summary and general discussion of the findings obtained in this study. Lastly, the chapter ends with some further issues for future studies and conclusions.

1.3 Languages in focus: Japanese and German

Japanese and German are investigated as examples of two languages whose prosodic systems build a contrastive pair. In this section, the functions and forms of pitch and segmental length in Japanese and German are described that are necessary to understand the following studies. In this thesis, Japanese refers to standard Tokyo Japanese and German to standard German. Prosodic variations in dialects are not addressed.

1.3.1 Pitch

In Japanese, pitch accents are used primarily for lexical contrasts. The presence/absence of a pitch accent is an inherent property of a word[3] and does not have any prominence-lending or discourse function (Beckman and Pierrehumbert, 1986). The meaning of a word changes depending on the position of a lexically specified pitch fall (e.g. /hàshi-ga/ = chopsticks$_{NOM}$, /hashì-ga/ = bridge$_{NOM}$, /hashi-ga/ = edge$_{NOM}$, the grave accent indicates the position of the pitch fall, if lexically specified). Phonetically, a Japanese pitch accent is realised as a sharp pitch fall from a high level occurring near the end of the accented mora to a low level in the following mora (Gussenhoven, 2004; Vance, 1987), which is not the case for a German falling pitch accent. If the first word in a phrase does not have an accent on the first mora, or if a word is spoken in isolation, then it starts with a low pitch, which then rises to high over subsequent morae, which is known as initial low (ibid.). Note that only 14 % of homophonic word-pairs are really distinguished by pitch accents in Japanese (Shibata and Shibata, 1990).

This small number of homophonic word-pairs distinguished by pitch accents in Japanese poses the question whether whether Japanese L1 speakers store lexical pitch information together with words and use pitch accent for an efficient word recognition. This is especially questionable, because the Japanese pitch accent is ranged between two syllables (= a polysyllabic phenomenon), so that it takes longer time for its processing in spoken word recognition (Walsh, 1993) (see details of the study in the next paragraph). Moreover, Japanese pitch accent patterns vary due to various post-lexical factors, such as word position in a phrase and in compounds (e.g. Hirose and Minematsu, 2004) and they vary across dialects

[3] Hereby, a word is meant as *jiritsugo* (= "an independent word") in Japanese, that contains lexical meaning as opposed to *fuzokugo* (= "an ancillary word") that carries a grammatical function, (Masuoka and Tabuchi, 1992).

Walsh (1993) posed Limiting-Domain Hypothesis claiming that the syllable is a unit of processing in lexical access. According to her hypothesis, tone is effectively used for lexical access, while pitch accent is not, although both are perturbations of F_0 and acoustically identical. This is because the former is a meaningful F_0 defined for every syllable, while the latter is defined only once per word (= defined as *jiritsugo*, see Subsection 1.3.1). To verify her hypothesis, she conducted a same-different judgement experiment in which Japanese listeners heard pairs of CVCV words or nonwords which were either same or different, either in pitch accent or in one of the four segments. *Different* judgements were significantly slower for pairs varying in pitch accent than for pairs which varied segmentally, irrespective of the position of the segmental difference. Thus even a difference in the final vowel (at which time the pitch accent pattern should also be unambiguous) led to significantly faster responses than the pitch accent difference. Further studies also support the view that pitch accent is not effectively used for word recognition. Otake et al. (1993) found no effects of pitch accent in a syllable-detection task with Japanese listeners: The first CV of a word was perceived equally rapidly and accurately irrespective of whether the word had HLL (e.g. monaka) or LHH (e.g. kinori) accent pattern. Also a study in neuroscience supports this view: Tamaoka et al. (2014) investigated whether L1 Japanese listeners necessarily use pitch accent in the processing of accent-contrasted homophonic pairs measuring electroencephalographic potentials. Electrophysiological evidence (i.e., N400) was obtained when a word was semantically incorrect for a given context but not for incorrectly accented homophones. Their finding suggests that pitch accent plays a minor role when understanding Japanese. In the case of Chinese, the N400 was consistently observed when an incorrectly accented word was embedded in a context (Li et al., 2008; Zhao et al., 2011) strongly supporting that tonal information is essential for the word recognition in Chinese.

However, Minematsu and Hirose (1995) reported opposite findings: They conducted gating experiments and found that detection of the pitch accent that has pitch fall on the second mora facilitates the word recognition. An early pitch fall in the F_0 contour for this type of accent makes it possible to identify the accent before the completion of the word recognition process. Accordingly, prosodic information should be utilised to facilitate the access to the mental lexicon by limiting the searching space. Also Cutler and Otake (1999) conducted a two-choice classification task, a gating task and a lexical decision task. In their experiments, words were successfully recognised exploiting F_0 and they conclude that accentual information constrains the activation and selection of candi-

dates for spoken-word recognition. However, their experimental method is questionable as they used only few speakers for the stimuli and they found a talker effect.

I regard the controversial results across the previous studies as empirical support for the claim that pitch accent in Japanese plays only a minor role in word recognition, because the studies testing Chinese L1 listeners report more consistent findings that they exploit pitch information for word recognition regardless of different task requirements. Braun et al. (2014) examined the ability to store lexical tone testing German, French and Japanese L1 listeners in comparison to Mandarin Chinese L1 listeners and showed that Mandarin Chinese controls had the highest sensitivity, followed by the German participants. The French and Japanese participants showed no sensitivity. Tonal information seems to be stored and processed differently by Chinese and Japanese L1 listeners, even though both languages employ pitch lexically. Further, these apparently different roles of lexical pitch in Japanese and Chinese bring to the discussion about the linguistic typology of tone languages and pitch accent languages. Japanese is classified to the former one, while Chinese or Thai to the latter, even though Japanese is sometimes classified as a restricted type of a tone language (Hyman, 2009). The Japanese pitch accent system is characterised as a mixture of various properties in prototypical stress vs. tone systems (Hyman, 2006, 2009; Hyman and Wilson, 1992).

In German, pitch is not used for a lexical distinction. Instead, the locations of metrically strong syllables are determined in a word and they contribute to a lexical distinction. The metrically strong syllables serve as docking sites to which pitch accents may be associated, for example to post-lexical information such as syntactic or pragmatic information (e.g. signalling statement vs. question sentence or a double contrast) (e.g. Braun, 2006; Féry, 1993) as well as to paralinguistic information such as attitude and emotion of a speaker (e.g. Chen, 2005; Gibbon, 1998; Liscombe, 2007; Scherer et al., 1984; Wichmann, 2000). Phonetically, a pitch fall (and also a pitch rise) in German is accompanied with a longer vowel duration and higher intensity because it takes place in a metrically strong syllables. In a stressed-timed language such as German or English, metrically strong (= stressed) and weak (= unstressed) syllables differ in duration, vowel quality, pitch and intensity (Ladd, 1996).

Moreover, the inventory of phonological accent types in German pitch accents (= six basic pitch accents types) (e.g., Grice et al., 1996) is richer than in Japanese (= only one pitch accent type, e.g. Venditti, 2000).

Since a Japanese pitch accent primarily has a lexical function and there is only one type of the accent, its use and variation for a post-lexical or paralinguistic purpose is lim-

ited compared to German. For example, Li et al. (2013) showed that Japanese L1 speakers expressed different emotional statuses by varying maximum and minimum as well as mean pitch without changing the phonological form of a pitch accent. German L1 speakers may additionally vary pitch accent types to convey such paralinguistic information (Bänziger and Scherer, 2005; Gibbon, 1998).

1.3.2 Segmental length and rhythm

Regarding another prosodic property under investigation in this thesis, segmental length, Japanese has more lexical restrictions than German. Japanese exhibits lexical vowel and consonant length contrasts (e.g. /kite/ = *come*, /kiːte/ = *listen*, /kitːe/ = *cut*, all verbs in the imperative form, the colon indicates a long segment). German, on the other hand, exhibits only lexical vowel length contrasts to a limited extent. That is, vowel length contrasts in German are accompanied with the vowel quality difference except for /a/ and /aː/ as in *Stadt* (= *city*) with a short vowel and *Staat* (= *state*) with a long vowel (Wiese, 2000). Consonant length contrasts are not used lexically in German (but in Swiss German, Kraehenmann, 2001 and Kraehenmann and Lahiri, 2008). Double consonants in German are used only in the spelling system and indicate the shortness of preceding vowels. True geminates, consonants containing a syllable boundary and potential word boundary occur only in sandhi in sequences like *Erbpacht*, *gut tun*, *Schiffahrt*, *viel leisten*, *hinnehmen* (Goblirsch, 1990, 18) or *Mitteilung*.

Speech rhythm, a more global combination of durations, also differs in both languages. Japanese is classified as a mora-timed language (Bloch, 1950; McCawley, 1968), while German as a stress-timed language, although the validity of the typological categories is critically discussed (e.g. Arvaniti, 2009; Warner and Arai, 2001). In Japanese, a vowel (V) or a consonant-vowel (CV) syllable takes up one timing unit (mora) and all morae have approximately the same perceptual duration (Bloch, 1950; McCawley, 1968, but also see Beckman, 1982; Han, 1962 for controversial findings). A Japanese pitch accent does not trigger a longer duration (Beckman, 1982; Homma, 1981). Hence, it does not affect the mora-timing. In German, stressed syllables occur at approximately regular intervals.

All these differences in the use of the same prosodic property in Japanese and German may become hurdles in L2 acquisition. The next section presents how transfer from one's L1 to an L2 is predicted in the influential models and theories of L2 acquisition.

1.4 Prosodic transfer in the models and theories of L2 acquisition

Transfer is an important issue in the L2 acquisition research and is probably the most investigated factor that is believed to influence the L2 acquisition not only in phonology, but also in other linguistic areas. In the following, some of the most influential theories and models are outlined: 1. Lado's Contrastive Analysis Hypothesis (CAH) (Lado, 1957), 2. Eckman's Markedness Differential Hypothesis (MDH) (Eckman, 1977, 2008), 3. Flege's Speech Learning Model (SLM) (Flege, 1999; Flege et al., 2002, 1995a), 4. Best's Perceptual Assimilation Model (PAM) (Best, 1995; Best et al., 2001; Best and Tyler, 2007) and 5. Kuhl's Native Language Magnet Model (NLMM) (Kuhl, 1991; Kuhl and Iverson, 1995). The first three theories predicted transfer both in speech perception and production, while the latter two were models on speech perception. As for linguistic areas, the first two did not specify a certain applicable linguistic area, whereas the latter three were proposed for the field of phonetics and phonology. Finally, the last model was originally proposed to account for the L1 acquisition, while other models and theories were originally considered to explain the L2 acquisition. As for the theories of phonological acquisition, phonology mostly at the segmental level was concerned. Therefore, I will discuss the possibilities and limitations to extend these theories and models to account and predict prosodic transfer.

1.4.1 Influential theories and models up to date

CAH

The first theory in L2 studies that put transfer from one's L1 to an L2 into the core was the CAH proposed by Lado (1957). The framework was embedded in behaviourist psychology and structural linguistics in which L2 learning was considered to be a matter of new habit formation while this being impeded by existing L1 habits. The CAH attempted to predict and describe all difficulties in L2 learning by systematically comparing the language to be learned with the L1 of the learner without taking learners' individual differences or their strategies into consideration that are actively applied by themselves.

The predicted degree of difficulties based on the CAH is shown in the hierarchy of difficulties proposed by Stockwell et al. (1965), see Figure 1.1. It is important to mention that this hierarchy of difficulties does not have theoretical or empirical basis. It is

based only on the conviction that the degree of linguistic difference predicts the degree of learning difficulties. For example, contrary to the CAH, an L2 feature may be new to the learner, and yet easy to acquire (Rasier and Hiligsmann, 2007, 42). Nowadays, a strong version of the CAH lent himself too much criticism due to its strict conviction to attempt to predict *all* kinds of difficulties in L2 acquisition (Wardhaugh, 1970). Still, the notion to compare various features of an L1 and an L2 in order to predict a *possible* difficulty or to understand the sources of error a posteriori is still useful.

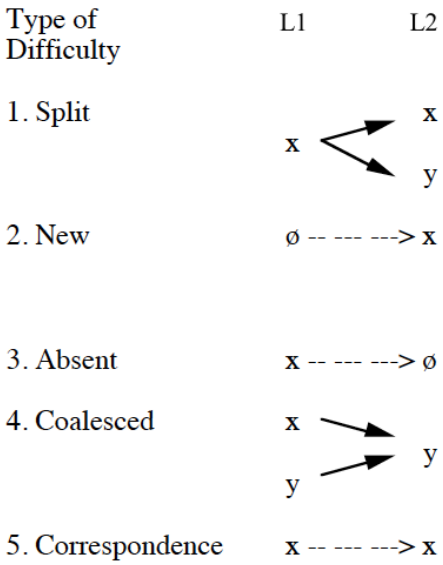

Figure 1.1 *Simplified version of the hierarchy of difficulty (based on information given in Stockwell et al., 1965) cited from Ellis (1994, 307)*

MDH

Whereas the CAH attempted to predict and explain L2 learning difficulties only on the basis of the differences gained from the theoretical comparison between an L1 and an L2, Eckman (1977) claimed that such a comparison between an L1 and an L2 is necessary, but is not sufficient to explain and predict L2 learning difficulties. He proposed to incorporate typological markedness into the explanation (Eckman, 1977, 2008). A phenomenon *A* in a language is more marked than *B* if the existence of *A* necessitates the existence of *B*, but not vice versa. Eckman predicted that a marked rule is more difficult to learn than an unmarked one in L1 acquisition and that marked L2 patterns that are less marked than in L1 should not be difficult in L2 acquisition. Unmarked patterns

can be easily transferred to L2 than the marked ones. His predictions were supported by numerous studies on the acquisition of L2 segments (e.g. Anderson, 1987; Major, 2008).

SLM

The SLM (Flege, 1999; Flege et al., 2002, 1995a) attempted to account for age-related limits on the ability to produce L2 segments in a native way and proposed four postulates and seven hypotheses (postulates and hypotheses are all shown in Flege et al., 1995a, 239). Core aspects of the model are summarised in the four postulates. They claim that adult L2 learners 1) keep the capacity to form new categories using the same processes and mechanisms used for their L1 acquisition, 2) use the same cognitive resources in L1 acquisition and L2 learning, 3) store phonetic information in the "common phonological space" while mutually influencing one another, and 4) can modify the mental representations in L2 learning. The seven hypotheses predict the conditions and stages of phoneme acquisition. For example, they state that the grade of the perceived (dis)similarities between L2 phones and L1 mental representations predict how it is likely that these L2 sounds are assimilated into the L1 representations: The greater the perceived dissimilarity of an L2 sound from the closest L1 sound, the more likely a new category will be formed for the L2 sound, but when an L2 sound is too similar to an L1 counterpart, the L1 and L2 categories will assimilate (Flege, 1995, 239). Flege et al. (1995a) themselves clearly formulate that the SLM primarily explains the ultimate attainment of L2 segmental acquisition and focuses on bilinguals or advanced learners, but not beginners.

PAM

The PAM proposed by Best and colleagues (Best, 1995; Best et al., 2001; Best and Tyler, 2007) can be discussed most clearly by first briefly reviewing the aspects of a direct realist view of speech perception (Best, 1994, 1995; Fowler, 1986, 1990a,b) and its philosophical foundations, on which PAM is based.

The central premise of direct realism is that a listener directly apprehends the perceptual object and does not sorely apprehend representative or abstract features from which the object must be inferred or constructed (Best, 1995, 173). As for speech perception, it is a direct mapping from acoustic qualities to the gestures that produced them. The acoustic waveform is regarded simply as an energy medium shaped by and therefore carrying information about distal vocal tract gestures.

When acquiring an L1, infants develop the ability to pick up the information that distinguishes segmental categories and that does not. The perceptual learning entails discovering which constellations of articulatory gestures are used in their L1; for example, the temporal phasing between alveolar closure, velar narrowing (Best, 1995, 177) and acquiring the critically distinctive features and the most telling differences among objects and events that are of importance to the perceiver (Best, 1995, 184).

When perceiving L2 sounds, the PAM predicts that L2 segments will be perceived according to their similarities to the L1 segmental representations that are the nearest from the L2 sounds in the L1 phonological space. The PAM underlays the direct realist view of speech perception and shares the view that the phonological space is universally defined by phonetic domain with the spatial layout of the vocal tract and the dynamic characteristics of articulatory gestures and that those distal properties provide the dimensions within which a similarity is defined. The PAM defines similarities between L2 segments and L1 gestural constellations and predicts and determines listeners' perceptual assimilation of the L2 phones to L1 categories (Best, 1996; Fowler and Rosenblum, 1989).

Based on this assumption, an L2 segment can be assimilated to an L1 segmental category either as a good exemplar of that category or as an acceptable, but not as an ideal exemplar of the category or as a notably deviant exemplar of the category (Best, 1995). Otherwise the L2 segment will be assimilated within L1 phonological space as a speech-like gestural constellation, but not as a clear exemplar of any particular L1 category or it will not be assimilated to speech and will be recognised as nonspeech sound. Importantly, the PAM provides a useful framework for a psycholinguistic study that conducts a discrimination task since it outlined the degree of perceptual differentiation and of discriminability for L2 segment contrasts as follows: In *Two-Category Assimilation*, each L2 segment is assimilated to a different L1 category and discrimination is expected to be excellent. In *Category-Goodness Difference*, both L2 sounds are assimilated to the same L1 category, but they differ in terms of the distance from the L1 "ideal" (e.g. one is acceptable, the other is deviant). Discrimination is expected to be moderate to very good. In *Single-Category Assimilation*, both L2 sounds are assimilated to the same L1 category and they are equally far from the native "ideal" (e.g. both are equally acceptable or equally deviant). Discrimination is expected to be poor. In *Both Uncategorisable*, both L2 sounds fall within phonetic space, but outside of any L1 category. Discrimination is expected to range from poor to very good. In *Uncategorised versus Categorised*, one L2 sound assimilated to an L1 category, the other falls uncategorised outside L1 categories. Discrimination is expected to be very good. In *Nonassimilable*, both L2 sounds fall outside of speech

domain being perceived as nonspeech sounds. Discrimination is expected to be good to very good. Differently from the CAH, the PAM does not account that a new or absent L2 category in comparison to an L1 category will automatically cause difficulties to acquire or to discriminate the sound. For example, the PAM predicts that an L2 sound which is very different from an L1 category will not be assimilated to any L1 category and can be perceived without any difficulty.

NLMM

The NLMM (e.g. Kuhl, 1991; Kuhl and Iverson, 1995) underlies the theory of exemplars (Goldinger, 1996, Hintzman, 1986, Johnson, 1997 or see summary in Pierrehumbert, 2001) and claims that each time when infants hear a phoneme, it is stored as an exemplar. Each exemplar builds a part of a cloud, with the best one becomes a prototype. The prototype acts as a phonetic reference of that category like a "perceptual magnet", attracting new exemplars towards the category centre falling within its zone of influence. In this way, a new exemplar will be assigned to the existing prototype categories. The model was originally intended to explain how infants tune their language-general perception to L1-specific perception abilities by the end of the first year of life (Werker and Tees, 1984a), but even then its possible application to L2 acquisition was stated (Kuhl, 1991, 1993). An L1 prototype attracts an L2 sound towards its centre when an L2 learner hears an L2 sound that is similar to an L1 sound, but not when the sound is not similar to the L1 sound. As for discrimination ability, Kuhl (1991) conducted a discrimination task testing English L1 listeners. In the *Prototype* condition, the prototype /i/ vowel served as the referent stimulus and its 32 surrounding variants served as the comparison stimuli, while in the *Non-Prototype* condition, the nonprototype /i/ vowel served as the referent stimulus and its 32 surrounding variants served as the comparison stimuli in the discrimination task, see Figure 1.2. The results showed that overall percent-correct scores were significantly lower in the *Prototype* condition, namely when a stimulus perceived as having high category goodness was used as the referent vowel in the discrimination task, indicating the difficulty in perceiving differences between the prototype and other members of the category. The opposite result was found for the *Non-Prototype* condition.

Further, the theory that underlies the NLMM, namely the exemplar theory, still leaves some questions to explain the acquisition of L2 sounds. The NLMM claims that L1 prototypes attract incoming L2 sounds and predicts that L1 phonological categories absorb L2 stimuli and block on the path to acquire the L2 sounds. This claim is difficult to combine with the assumption of the exemplar theory that new prototypes can be developed

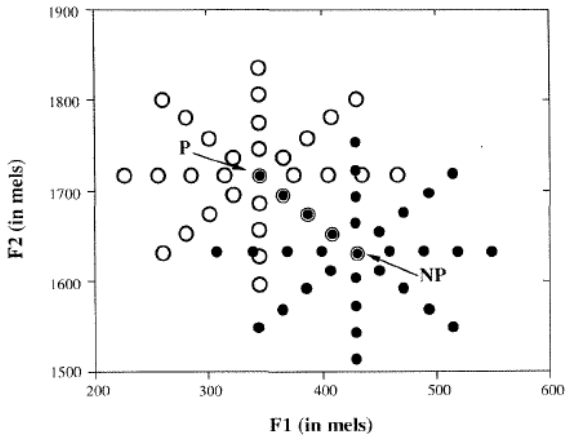

Figure 1.2 *The prototype /i/ vowel and variants on four orbits surrounding it (open circles) and the nonprototype /i/ vowel and variants on four orbits surrounding it (closed circles) in (Kuhl, 1991, 95).*

just by having a certain amount of exemplars. Following the exemplar theory, this reorganisation of the listener's perceptual space should occur straightforward correlating with the amount of experience with exemplars (Lacerda, 1995). If the statistical learning suggested by the exemplar theory is true not only for L1 acquisition by infants and small children (Werker and Tees, 1984a) before critical period (Lenneberg, 1967; Scovel, 1988), but also for L2 learning by adults who were exposed to the L2 after the period, L2 learners should be able to efficiently acquire novel sounds and phonological contrasts only with sufficient inputs. Additionally, as pointed out in Lacerda (1995), the "old" L1 prototypes must be relocated during the L2 acquisition process despite the magnet effect. Within the current exemplar-based model, re-tuning is a consequence of memory decay affecting "old" exemplars and fading out the representations of inactivated exemplars (Lacerda, 1995). However, as for L2 acquisition, the existing L1 prototypes and exemplars that are distributed around the prototypes should not be deleted, because the L1 prototypes are "old", but still required. Finally, if even adult L2 learners can develop new prototypes and re-tune prototypes as a result of the learning procedures, the NLMM does not explicitly state under which circumstances they are able to counteract the magnet effect.

Findings in previous studies indicate that it is difficult and time-consuming for adults to acquire novel phonemes only via statistical learning to which they were perfectly sensitive only during their earliest time of life. Thus, successful acquisition of novel phonological categories potentially requires explicit tutoring (cf. Menning et al., 2002). For ex-

ample, Dobel et al. (2009) investigated the acquisition of the voiceless, bilabial fricative /ɸ/ via a statistical learning paradigm testing German L1 speakers. Their behavioural data and the N400 component (= the index of lexical activation/semantic access) showed that participants were able to learn to associate the pseudowords with the pictures, but they could not discriminate them within the minimal pairs. Importantly, the novel words with the sound /ɸ/ showed smaller N400 amplitudes than those with L1 phonemes before learning, indicating their higher non-word status. After learning, it was shown that /ɸ/ had become integrated into the L1 category /f/, instead of establishing a novel category. Their results demonstrate that L1 phonemic categories attract novel sounds and this interferes the acquisition of novel L2 contrasts. Further, they discussed that the results do not exclude the possibility that adults learners can acquire novel phonological categories, possibly by means of a more intensive and active training (e.g. training with feedbacks and improvement) or larger differences between L1 and L2 sounds.

1.4.2 Possible extensions and limitations to L2 prosodic research

In this subsection, I will proceed the evaluation of the possible extensions and limitations of the presented models and theories to account for the learning of L2 prosodic contrasts.

CAH

The CAH, which was not specifically proposed to account for phonological phenomena, was applied for the research of L2 phonology. According to Ringbom (1994, 738), contrastive phonology is the area in which the predictions of a contrastive analysis work best. However, it seems that the comparison between L1 and L2 phenomena is not straightforward in all research areas in phonology, such as in prosody. For some of the categories of the hierarchy of difficulties proposed by Stockwell et al. (1965), it is difficult to find a suitable example in the prosodic domain (e.g. for the categories “split” or “coalesced”, see Figure 1.1). For other categories, it is difficult to define what does “completely new” or “completely absent” mean. For instance, Japanese lexical use of pitch accents for German learners of Japanese could be claimed to be “completely new”, however, the use of pitch accents itself is not new for German, because they are used not at the lexical level, but at the post-lexical or paralinguistic level. Therefore, it should be clarified what is “completely new” in this case. For the easiest category (= an L1 and an L2 category completely correspond to each other), an example would be that both German and Japanese apply a rising boundary tone for a polar question. Note that the phonetic realisation of

the rising boundary tone in Japanese (Fujisaki and Hirose, 1993) is different than the one in German (Michalsky, 2014), so that it is ultimately not clear, whether this prosodic phenomenon "completely correspond" to each other despite the phonetic differences. Some terms used in the predictions (such as "completely correspond" or "correspond new") are too vague to make predictions on cross-language prosodic transfer.

For all aforementioned five categories, it is more clear to find examples at the segmental level e.g. by comparing L1 and L2 vowels and consonants. What the contrastive analysis of L1 and L2 prosody makes more difficult is that the same prosodic property may be layered at different linguistic levels in an L1 and an L2. Moreover, differences can be manifested at the phonological level, but also at the phonetic level. Therefore, multiple aspects of the same prosodic phenomenon in an L1 and an L2 must be taken into account together. For this reason, it is difficult to define what is new or absent between the L1 and L2. The same phonological category (e.g. a rising boundary tone) may have cross-language differences in the phonetic realisation forms. Since there are phonetic variations within a prosodic category, it is crucial to determine whether an instance of an L2 category can be identified as a member of an L1 category. In order to compare prosodic phenomena cross-linguistically, it is important to take different dimensions of prosodic transfer into account.

MDH

The applicability of the MDH to the research on L2 prosodic transfer has two opposite views. Rasier and Hiligsmann (2007) support the MDH to best predict prosodic transfer and applied it in his study on L2 acquisition of pitch accent in Dutch and French. They claimed that structural constraints on accentuation outweigh pragmatic information in French, while it is the opposite in Dutch. Many other languages rely on both structural and pragmatic rules in their accent placement strategies, albeit in a different order of preference. But there seems to be no language where structural constraints are totally absent, see Figure 1.3.

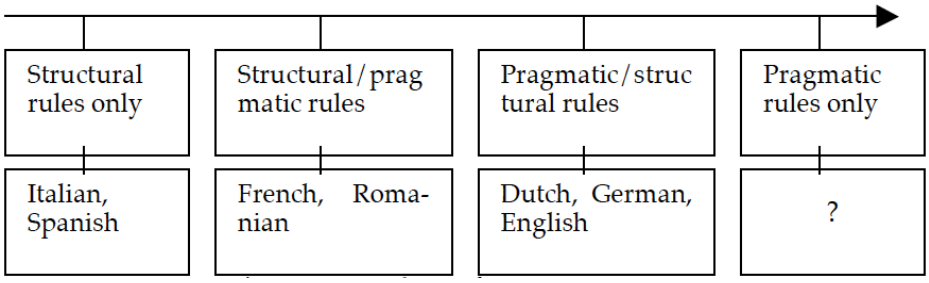

Figure 1.3 *Typology of accented system Rasier and Hiligsmann (2007, 53)*

They thus defined that structural accentuation rules constitute the unmarked case, whereas pragmatic ones the marked case and assumed that the German pragmatic constraint for French L2 learners is more difficult to acquire than the French structural constraint for German L2 learners. Their production experiment approved their assumption. On the other hand, He (2012) stated the limitation of the MDH for the research on acquisition of L2 prosody. He argued that the MDH presupposes the concept of linguistic universals (Greenberg, 1966) and its scope may be restricted to linguistic areas such as syllable structures or the frequency of segmental speech sounds for which linguistic universals have been proposed. Other phonological features and prosodic phenomena are difficult to be classified in terms of typological markedness, so the MDH may not be able to predict their acquisition (He, 2012, 21).

Let's take an example from a prosodic phenomenon investigated in my thesis; pitch accents in German and Japanese. While German exhibits the post-lexical and paralingusitic use of pitch accents, Japanese additionally employs the lexical use of pitch accents. Applying the logic made by Rasier and Hiligsmann (2007), the lexical use of pitch accents supposed to be marked and the post-lexical and paralinguistic use of them unmarked. Following the MDH, the acquisition of Japanese lexical pitch accents by German L2 learners should not be as difficult as the acquisition of German pitch accents by Japanese L2 learners. However, it is questionable whether the acquisition of pitch accents at the lexical level and at the post-lexical level are qualitatively comparable. Moreover, even though both languages use pitch accents at the post-lexical and paralinguistic level, it seems that the acquisition of post-lexical and paralinguistic use of pitch accents in Japanese as L2 and German as L2 seems to be qualitatively different, since the post-lexical and paralinguistic use of pitch accents in Japanese are restricted by their lexical use (Asano, 2015), whereas pitch accents in German do not have such a lexical restriction. These examples indicate the complexity of the notion of markedness. Since my thesis primarily investigates the processing of L2 prosody and does not mainly aim at discovering the hierarchy or grade of difficulties of certain prosodic phenomena in comparison to other prosodic phenomena, I will merely make use of the notion of comparing L1 and L2 prosodic phenomena to predict *possible* (but not all) difficulties in L2 acquisition.

SLM and PAM

SLM: Flege et al. (1995a) explicitly states that the model is proposed to account for the acquisition of L2 segments (vowels and consonants) and to make predictions based in the

phonetic systems used in the perception and production of vowels and consonants. The four postulates proposed in the SLM present the processes and mechanisms of the acquisition of L2 segments rather in general terms, which could be empirically investigated in the research of prosody. Most of the seven hypotheses proposed in the SLM require to empirically measure or to define perceived (dis)similarities, what is often difficult to do for prosodic phenomena. A distance between two vowels in the phonological space and the perceived (dis)similarity between the vowels can be measured by their F1, F2 and F3, thus it is relatively clear to define. Researchers measure a distance between an L1 and L2 consonant by comparing their place and manner of articulation and by examining whether a similar consonant to the L2 consonant exist in the L1. In this way, their perceived (dis)similarities are quantifiable. However, in the case of prosodic contrasts such as ours, it is difficult to define the perceived (dis)similarities. For instance, how can a perceived (dis)similarity between a Japanese lexical pitch accent and a German non-lexical pitch accent be defined? If ever, then the difference between their phonetic realisation may be compared. The function of a prosodic property in a language seems to be difficult to quantify. Although the same prosodic cues are employed in many languages, the relevant acoustic cues and their functions vary widely from language to language.

Moreover, it is important to mention that Flege et al. (1997) themselves admitted the difficulty to predict the perceived (dis)similarity between an L1 and an L2 category even at the segmental level and pointed out the importance of an adequate assessment of perceived L1-L2 (dis)similarities and the necessity of standardised measure of perceived L1-L2 phonetic distance. Although establishing (dis)similarities is crucially important for the SLM, to date there is no commonly accepted way of measuring cross-language (dis)similarities (Bohn, 2002) and our understanding of the exact nature of cross-language (dis)similarities is still rather limited (Strange et al., 2001). The same problem is even more acute to establish (dis)similarities of cross-language prosody as it interacts with other prosodic parameters at different linguistic levels. This owes the current situation that the focus of L2 speech models has been on segments rather than prosody, given that segments are relatively easy to describe, to analyse and to test compared to prosody (Mennen, 2015).

PAM: Similarly to the problem stated for the SLM, it is difficult to define perceived (dis)similarities for prosodic contrasts and to apply the PAM to predict prosodic contrasts such as the research objects of my thesis. Following the core aspects of the PAM, the perception of L2 segments crucially depends on the (dis)similarity of phonetic properties of the L2 segment and L1 categories.

Nevertheless, some attempts have been undertaken to apply the PAM for prosodic research (= the PAM-L2). So and Best (2008) and So and Best (2011) investigated whether L1 listeners of non-tone languages (Australian English and French) perceive L2 Mandarin tones according to their L1 prosodic categories. They asked participants to categorise the Mandarin tones into four categories "Flat pitch", "Question", "Statement", and "Exclamation". They reported that both English and French participants categorised non-native tones according to their L1 intonational categories, and that categorisations were based on the phonetic similarities of the pitch contours they perceived between Mandarin tones and their L1 intonational categories. Thus, they concluded that their findings would support an extension of the PAM to the suprasegmental domain and that L2 prosodic categories (e.g. lexical tones) would have been assimilated to the categories of listeners' L1 prosodic system (e.g. nuclear tunes). However, the assimilation discussed in the studies appear to be induced by their methodological design. In the experiment, participants were explicitly asked to select one of the given L1 categories when hearing the L2 tones. Therefore, it might be possible that the L2 listeners actually would have not associated the L2 tones with their L1 categories when hearing them, if no categories were given to select. Moreover, the studies tell us about how acoustic correlates of lexical tone assimilate to acoustic correlates of post-lexical tone in other languages, but they still do not address whether L1 prosodic categories used at one linguistic level (e.g. at the post-lexical level) can be utilised to acquire L2 prosodic categories used at another linguistic level (e.g. at the lexical level). Following So and Best (2008) and So and Best (2011), German learners of Japanese should not have difficulties in acquiring L2 Japanese lexical pitch accent, given that a falling pitch accent in German may correspond to the phonological form of the Japanese lexical pitch accent, which is the falling one as well. However, the falling pitch accents in both languages phonetically differ in details (see section 1.3.1). It is questionable whether the falling pitch accent in Japanese may still be assimilated to the German falling pitch accent despite the phonetic differences. Similarly to the aforementioned problem in applying the SLM to the prosodic domain, it seems to be difficult to define a perceived (dis)similarity between an L2 and an L1 prosodic category in the phonological space.

NLMM

Taking a prosodic contrast investigated in my thesis as an example; a consonant length contrast for German L2 listeners, Kuhl's statement suggests that German learners of Japanese may be able to easily discriminate short and long consonant contrasts, when

the durations of long consonants would be "sufficiently" different from those of short consonants, but if not, long consonants will be attracted to the L1 prototype of short consonants and thus they may not be able to discriminate the contrasts. This statement is nothing else but categorical perception of speech (e.g. Harnad, 1987). As for L2 lexical pitch contrasts, it is not clear whether the distance between the lexical use of a pitch accent in Japanese and the non-lexical use of a pitch accent in German can be compared simply on the same scale. As it was the case for the SLM and the PAM, it is not clear, whether the phonetic realisations of the Japanese and German pitch accents can be simply compared with each other and which phonetic cue of pitch accents is the most crucial to define a phonological differences between the Japanese and German pitch accents and how the "sufficient" differences between the categories can be defined and quantified. It seems to be less complex to define a distance between two vowels than a prosodic cue that is functionally different in two languages.

1.5 Stages of speech processing under investigation

As presented in Section 1.1, this thesis aims at discovering whether deviant productions of L2 prosody relate to the very initial stage of speech perception in sensory memory (="input" stage) or to the speaker's short-term memory while communicating with their long-term mental representations (= the path from "input" to "mental representations") or to the speech production that does and does not require the access to mental representations (= the path from "input", "mental representations" to "output" and in the path from "input" directly to "output", respectively). The experiments in chapter 3 and 4 test the processing in the "input" and the path from "input" to "mental representations". The experiment in chapter 5 examines the processing in the path from "input", "mental representations" to "output" and in the path from "input" directly to "output". This section presents these stages of L2 processing investigated in the experiments in this thesis and provides the methodological paradigm to differentiate the stages in the experimental conditions.

To this aim, the model of speech perception and production illustrated in Ramus (2001) and Szenkovits and Ramus (2005) is modified (in the specification of "sensory memory" and "working memory" based on the text in Ramus (2001)) and cited as a base to visualize the stages "input", "mental representations" and "output" within one figure. I am aware of the background that the ideas in Ramus (2001); Szenkovits and Ramus (2005) were inspired from the classic logogen model (Morton, 1969), whose main ideas with lo-

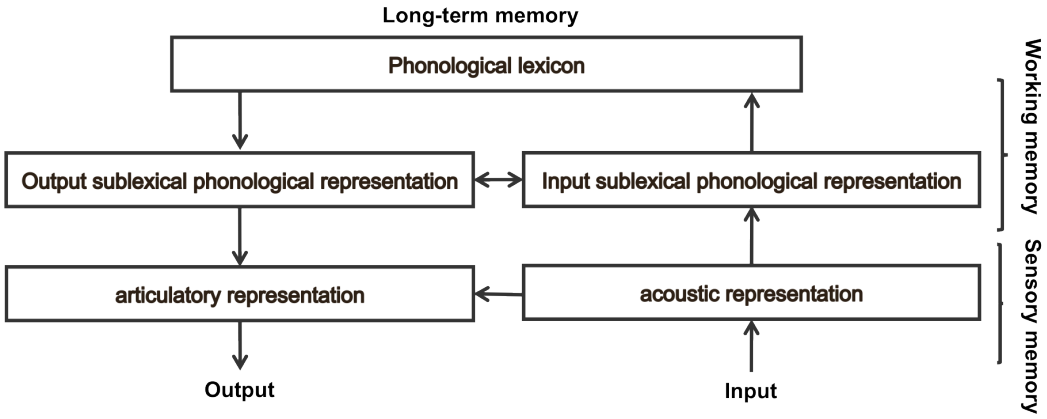

Figure 1.4 *Model of speech perception and production modified from Szenkovits and Ramus (2005, 255). The specification of "sensory memory" and "working memory" was added based on Ramus (2001).*

gogens does not show a direct relevance to my discussion in this chapter. The reasons for using the model by Szenkovits and Ramus are that 1) it visualises the relationships between speech perception, mental representations and speech production in a suitable way to explain the experimental paradigm and motivation for my study, and that 2) it presents both lexical and sublexical phonological representations whose distinction is relevant in my study as pseudowords were used, see Figure 1.4.

According to Ramus (2001), the phonological lexicon is a permanent storage for word forms, and word forms only; whereas the sublexical phonological representation is a short-term storage for whatever can be represented in a phonological format, that is, words, whole utterances and nonsense sequences of phonemes (pseudowords) (ibid., 201). Using pseudowords in the following experiments, the speech processing focused in this work does excludes the phonological lexicon in the model, even though the discussion that real words may be activated by hearing pseudowords in an experimental situation remains open (see e.g. Cibelli, 2012).

Acoustic representation and *input sublexical phonological representation* in Figure 1.4 on the right side of the model present speech perception, while *articulatory representation* and *output sublexical phonological representation* on the left side of the model speech production.

In my thesis, in the "input" stage, it was examined how L2 listeners perceive acoustic correlates of nonnative prosodic contrasts in sensory memory. This stage corresponds

to the path from *input* to *acoustic representations*. In *acoustic representation* (or *acoustic storage* in Gathercole, 1999), acoustic record of the most recent auditory speech item is stored in a sensory form (Gathercole, 1999, 413). In the path from "input" to "mental representations", speech perception that requires the listener to access mental representations was examined. In the model, this corresponds to the path from *input*, *acoustic representations* to *input sublexical phonological representation*. In the path from "input", "mental representations" to "output", speech production that requires the listener to access mental representations was examined. In the model, this corresponds to the path from *input*, *acoustic representations*, *input sublexical phonological representation*, *output sublexical phonological representation*, *articulatory representation* to *output*. Finally, in the "output" stage, speech production that does not necessarily require phonological representations was examined. In the model, this corresponds to the path from *input*, *acoustic representations*, *articulatory representation* to *output*.

In order to understand the speech perception that involves and that does not involve the access to long-term mental representations, the human memory types (a.o. Atkinson and Shiffrin, 1968) are briefly presented first, see Figure 1.5. Sensory memory, in which sensory information of the original stimuli is retained, is the shortest-term memory and decays or degrades very quickly, typically in the region of 200 - 500 ms (Cowan and Morse, 1986; Pisoni, 1973). The information that is perceived with attention reaches in short-term memory for temporary recall of the information. It holds a limited amount of information, typically (well-known) ± 7 items (Miller, 1956) for a limited period of time (up to 2000 ms with the rehearsal system and 15000 to 30000 ms without)(Atkinson and Shiffrin, 1971; Baddeley, 2003). Building on empirical evidence, Baddeley and Hitch (1974) developed an alternative model of short-term memory which they called working memory, providing different systems for different types of information, see details in Subsection 1.6.1.

Based on this different stages of speech processing and different memory systems involved, L2 learners could have problems in the earliest stage in speech perception; in perceiving acoustic correlates of sounds in sensory memory (Atkinson and Shiffrin, 1968). This problem relates to the so-called "input" stage in the current work. Previous studies generally agree, despite their methodological variability and thus their incomparability each other, that L1 listeners obtain higher discrimination ability of prosodic contrasts than L2 listeners (Altmann et al., 2012 testing L2 Italian - L1 German for consonant length contrasts; Hirano-Cook, 2011 testing L2 Japanese - L1 English for pitch accent contrasts; Qin and Mok, 2013 testing L2 Chinese - L1 English or French for tone contrasts;

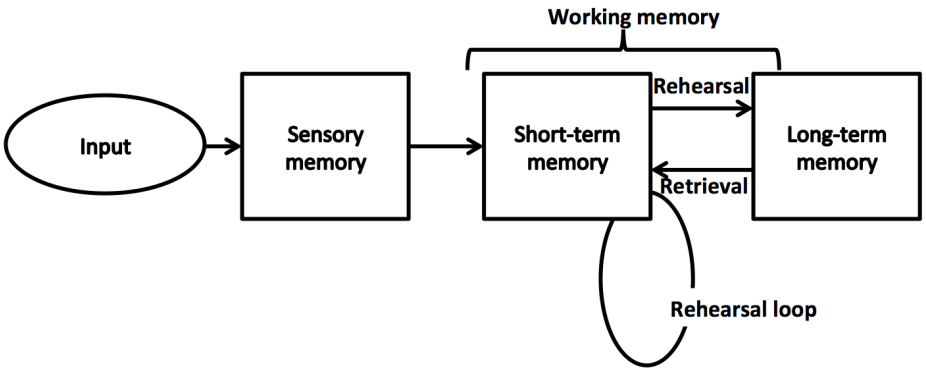

Figure 1.5 *Model of human memory developed based on Atkinson and Shiffrin (1968).*

Wayland and Guion, 2004 testing L2 Thai - L1 English or Chinese for tone contrasts). On the other hand, there is a good reason to believe that L2 listeners could have no difficulties in perceiving acoustic correlates of sounds and that they perform as well as L1 listeners in this stage. Even listeners who never had exposure to the L2 have sometimes been reported to discriminate nonnative prosodic contrasts as well as L1 listeners despite the lack of the L2 phonological categories in a listeners' L1. For example, Hayes-Harb and Masuda (2008) conducted a follow-up auditory discrimination experiment with two monolingual English listeners. They were asked to determine whether pairs of the test stimuli (e.g. *meso* and *messo*) that contrasted in consonant length were the "same" or "different". They discriminated the minimal pairs with 93% accuracy, indicating that the minimal pairs of consonant length were discriminable even to the non-learners. (However, note that the study did not provide the comparison with Japanese L1 listeners and only two non-learners were tested.) A reasonable explanation for this finding is the perceptual reliance on the durations of sensory memory (Pisoni, 1973).

Second, L2 learners could have difficulties in keeping the phonetic information in working-memory while accessing long-term memory. This problem relates to the path from "input" to "mental representations". In this view, L2 learners' deviant patterns in L2 productions compared to those of L1 speakers are related to their poor ability to store the target L2 contrasts phonologically and access them after phonetic memory decayed (see also section 1.6).

There is empirical evidence that L2 listeners have difficulties in perceiving nonnative prosodic contrasts when phonological representations are required for the task accomplishment, but not when the prosodic contrasts can be discriminated on the basis of their

acoustic correlates. For example, French L1 listeners, whose L1 does not employ lexical tone and stress contrasts, have been claimed to be "stress deaf", but this appears to be true for the processing at a phonological level rather than for the processing of acoustic correlates of stress contrasts (Dupoux et al., 2001, 2008; Dupoux and Mehler, 1997; Schwab and Llisterri, 2011). Moreover, French L1 listeners seem to be also "tone deaf". Hallé et al. (2004) reported that French L1 listeners had had difficulties in discriminating Taiwanese Mandarin tones in phonological processing, but not when the task required only phonetic processing. Furthermore, Sakamoto (2010) showed that English L2 learners of Japanese had difficulties in storing Japanese lexical pitch accent contrasts, but not in differentiating them at the acoustic level. These results show that L2 learners have difficulties in storing nonnative prosodic contrasts phonologically, but not in discriminating them based on their acoustic correlates.

In my thesis, I tested the "input" stage and the stage from "input" to "mental representations" using two different durations of ISIs, the time between the two stimuli presented. The duration of an ISI is claimed to influence the involved levels of speech processing (the longer an ISI is, the more phonological processing and less auditory and phonetic processing are activated and involved) (Cowan and Morse, 1986; Gerrits, 2001; Johnson, 2004; Pisoni, 1973; Schouten and Van Hessen, 1992; Werker and Logan, 1985; Werker and Tees, 1984b). This is because, as discussed before, acoustic information decays within several seconds (the discussion on human memory agrees on the time limit of two seconds, Baddeley, 1986; Crowder and Morton, 1969). After this time limit, the information will either decay or it is refreshed through an articulatory control process or rehearsal. In the latter case, phonetic information is held in working memory and then taps into long-term memory. During the rehearsal, the first echo back from long-term memory contains idiosyncrasies of the stimulus, but it is already somewhat abstract. If the echo in working memory is communicated to long-term memory again, the next echo will move closer to the central tendency of the stored category. After several seconds, the echo in working memory will be an abstracted prototype of the category. The feedback loop between working memory and long-term memory forces a regression toward the mean of the stored category. Thus, idiosyncratic details of the original acoustic information will be attenuated in the eventual echo (Goldinger, 1998, 256). Following this mechanism, idiosyncrasies of a stimulus will gradually decay in every rehearsal cycle if an L2 listener does not have the phonological category of the stimulus in her/his L1 mental representations. After a while, the listener can only hold the phonological information that (s)he already had in the L1 mental representations, but not the original

acoustic information. With a short duration of an ISI L2 listeners have a chance to discriminate nonnative prosodic contrasts based on the acoustic information of the stimuli. However, with a long duration of an ISI, they cannot rely on the acoustic information, but have to compare the first stimulus that is drifted toward more robust L1 phonological codes (Crowder, 1982; Pisoni, 1973; Werker and Logan, 1985; Werker and Tees, 1984b) with the second stimulus. Given the lack of an appropriate L2 prosodic representation in their L1 representations, L2 listeners' performance in discriminating nonnative prosodic contrasts is predicted to decrease in a long ISI condition.

Cross-language perception studies testing discrimination ability of nonnative segmental and prosodic contrasts reported that L2 listeners' discrimination ability was higher in a short ISI condition (= 500 ms as a standard) than in a long ISI condition (= 1500 ms as a standard), while the reverse was true for L1 listeners (e.g. Burnham and Francis, 1997; Werker and Logan, 1985; Werker and Tees, 1984b). L2 listeners had advantages when the contrasts could be discriminated on the basis of acoustic correlates. L1 listeners' advantage in the long ISI condition indicates that they filtered irrelevant phonetic information that was not beneficial for the discrimination of phonological categories (Burnham and Francis, 1997). The authors of these studies claimed that the short ISI condition triggered a phonetic level of processing, while the long ISI condition a phonological level. It has to be pointed out that they strictly differentiated "phonetic mode" and "phonological mode" and used the 500-ms vs. 1500-ms ISIs paradigm as an established standard upon which phonetic versus phonological modes of processing were distinguished. According to them, the phonetic mode of processing is a language-general mode of perception in which phones are discriminated without any influence of linguistic experience. However, this established standard should not be taken for granted. Wayland and Guion (2004) tested the discrimination of L2 Thai tone contrasts by Chinese and English L2 learners in comparison to Thai L1 listeners and did not find a clear ISI effect in the L2 listeners' groups. They argued that the short duration of an ISI (= 500 ms) did not necessarily preclude listeners from accessing phonological information stored in long-term memory. They further claimed that the discrimination performance may be affected by one or more of related factors such as L1 phonological system, salience of acoustic information.

The argument that phonological processing could be already involved in a short ISI condition (e.g. 500 ms) is also supported by previous results from semantic priming studies: Holcomb and Neville (1990) and Sabol and de Rosa (1976) showed that lexical access takes place around 200 ms after target onset. Based on this, listeners would be able to ac-

cess stored phonological category representations already in a short ISI condition, when ISIs are 500 ms long such as in the above-cited studies (e.g. Burnham and Francis, 1997 or Werker and Tees, 1984b). The short ISI of 500 ms might not necessarily prevent L1 (or even L2 listeners) from accessing their stored phonological information to aid their discrimination. As Wayland and Guion (2004) pointed out, the observed advantage in discriminating nonnative prosodic contrasts by L2 listeners in a short ISI was probably due to comparatively weaker demand placed on working memory. This effect, however, was not found for L1 listeners, because they were able to discriminate the contrasts by accessing their long-term memory representations in both short and long ISI conditions. No strict distinction between phonetic vs. phonological processing is also supported by Darcy et al. (2012) in their "direct mapping from acoustics to phonology". They claimed that phonological representations are co-activated and are mapped directly from acoustic speech signals (Darcy et al., 2012).

At this point, it should be mentioned that I will use the term "phonetic" (processing) including psychoacoustic and auditory processing without distinguishing them from each other as discussed by the Liberman's school and their associates (Fujisaki and Kawashima, 1969, 1970; Massaro, 1972; Pisoni, 1973, 1975; Pisoni and Tash, 1974; Studdert-Kennedy et al., 1970, 1972; Werker and Logan, 1985). This is because the difference between the auditory and phonetic information concerning the perception of prosodic contrasts is not crucial for the research aim of this thesis.

Thirdly, L2 learners could have problems in articulating L2 sounds independent from the problems relating to their mental representations. Figure 1.4 shows the arrow which connects *acoustic representations* and *articulatory representations* directly without mediating phonological representations. Speech production without requiring the access to phonological representations was investigated in the immediate imitation task in this thesis. Gestural theories, for example motor theory of speech perception (Liberman and Mattingly, 1985, 1989) and direct realist theory (Fowler, 1986, 1990a,b) claimed that speech perception is automatically mediated by an innate, specialised speech module to which listeners have no conscious access. In this view, an immediate imitation is nothing more than phonetic gestures which are automatically mediated during speech perception. According to the theories, "speech is perceived by processes by processes that are also involved in its production" (Liberman et al., 1967, 452) and they strongly support an underlying direct perception-production link. If the claim supported by the gestural theories is reliable, L2 speakers should be able to perceive acoustic correlates of an L2 sound without mediating their phonological representations, just in the same way as L1

listeners would do. Shortly after the acoustic stimulus is sensorially perceived, perceived gestures serves as a prime or goad for the speech production (Fowler, 1986). Therefore, L2 speakers should be able to immediately imitate the L2 sound without difficulties.

If, however, L2 speakers did not succeed in immediately imitating L2 sound correctly, they should have problems in articulatory processes. The claim that L2 learners could have problems in articulatory processes is supported by the assumption that the articulatory apparatus such as lips, jaws or tongues are adjusted in a language-specific way for one's L1(s) (Esling and Wong, 1983; Honikman, 1964; Laver, 1994; Mennen et al., 2010b). Also recent technological development of ultrasound images enables us to study actual language-specific phonetic settings such as tongue positions or lip protrusion (Gick et al., 2004; Wilson, 2006; Wilson et al., 2007).

L2 productions have been extensively documented in previous studies, but they are not always suitable to investigate articulations, because articulatory problems and problems relating to stored categories cannot be separated. The analysis of L2 production by means of an immediate imitation task attempts to separate the articulation problem from the problems relating to stored categories.

Finally, L2 learners could have problems in producing an L2 sound while accessing their stored phonological L2 information. Speech production that requires the access to phonological representations was investigated in the delayed imitation task in this thesis. The process is partially overlapped with the one in the path from "input" to "mental representation" that mediates phonological representations. While waiting to imitate, a speaker has to maintain acoustic information and store it phonologically. While continuous interactions occur between working memory and long-term memory, idiosyncratic details of the original imitation stimulus will be attenuated in the eventual echo, which makes it more demanding for the speaker to imitate the stimulus accurately.

The methodological paradigm of an immediate and a delayed imitation has been applied in previous studies. Different time intervals between the offset of a stimulus and the begin of an imitation were used to make phonetic information decay (Goldinger, 1998; Shockley et al., 2004) or to make speech processing access lexical representations (Balota and Chumbley, 1985). For example, Goldinger (1998) conducted an immediate and a delayed shadowing task. He tested how far idiosyncratic details of the original stimuli decay in the course of time. To this end, he varied the frequency of the words used as stimuli and talker variability (different vs. same voices of the stimuli) and examined whether these variables affected the immediate and delayed shadowing performance in an L1 (testing English L1 speakers). He found that the frequency and talker variability were normalised

in the delayed shadowing task. Balota and Chumbley (1985) tested an effect of word frequencies of stimuli on an imitation performance in an L1 (English). Their prediction was to find the effect of word frequencies when participants access their long-term memory. They varied the duration of delay intervals (400, 900 and 2900 ms) and found stronger effects of word frequencies in the longer delay interval conditions. Even though the studies vary with respect to task requirements and durations of a delay, they agree that the accuracy of an L1 speakers' performance decreases in a delayed imitation or shadowing task compared to an immediate one.

1.6 Cognitive load to understand L2 processing

The effect of a duration of an ISI on the extent of an information decay may be discussed in terms of the cognitive load placed on working memory. In Cognitive Psychology, cognitive load is understood as the total amount of mental effort being used in working memory (Sweller, 1988) whose capacity is inherently limited (Miller, 1956).

In this thesis, the cognitive load placed on working memory is regarded as one of the key aspects to detect the sources of foreign accents and to explain differences between L1 and L2 processing. For example, it gives an explanation why it is difficult to maintain nonnative sounds for a long time. The experimental conditions in this thesis were varied in the way that they systematically manipulated task demands and consequently cognitive load. By investigating the effects of task demands, our everyday speech processing may be better understood, because a task can be understood not only as an experimental task, but includes general activities in our everyday life. In this section, I will first present the most relevant aspects of working memory for the following experiments. Then I will proceed with discussing further factors that affect task demands in L2 processing.

1.6.1 Working memory

The initial model of working memory provided by Baddeley and Hitch (1974) comprises three components; *the central executive* and two slave systems *the phonological loop* and *the visuo-spatial sketchpad*. The central executive acts as a supervisory system and offers the mechanism for control processes in working memory, including the switching of attention and the mental manipulation of material held in the slave systems (Baddeley et al., 1998; Baddeley, 1986).

The phonological loop contributes to the storage of phonetic information over short periods of time. The system is fractionated into a store of the phonological short-term memory (Aliaga-Garcia et al., 2011; Gathercole et al., 1997; Kormos and Sáfár, 2008) and an active articulatory rehearsal component (Baddeley, 1986, 2003; Baddeley and Hitch, 1974). The phonetic input that reached the phonological short-term memory is subject to decay within approximately two seconds, unless refreshed through an active sub-vocal rehearsal process. In order to refresh the phonetic memory, this rehearsal receives feedbacks from mental phonological representations and links short-term memory and long-term memory. The rehearsal is not unidirectional (from short-term memory or working-memory to long-term memory), but it is a reciprocal one (Gábor and Mihály, 2008) and involves the same process with speech production (Baddeley and Hitch, 1974).

The other subsidiary system, the visuo-spatial sketchpad is responsible for holding visual and spatial information for short periods of time. In Baddeley (2000), he added a fourth component to the model; *the episodic buffer* that is responsible for integrating memory representations across different domains such as an auditory, a visual domain and possibly also with a smell and a taste (Baddeley, 2010), see Figure 1.6. Working memory plays a crucial role for complex cognitive activities such as speech perception and production involving multiple components of working memory.

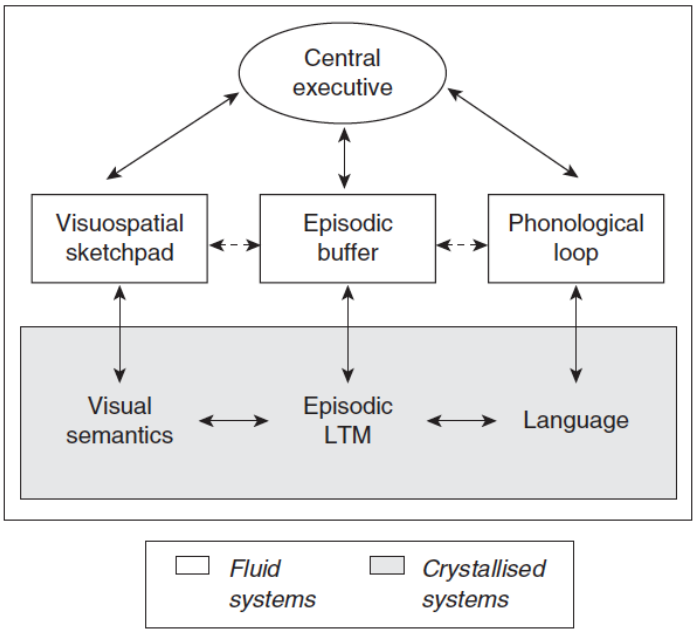

Figure 1.6 *The revised model of working memory (Baddeley, 2000, 421)*

Since the phonological loop and the visuo-spatial sketchpad present short-term memory mechanism in working memory, the distinction between working memory and short-term memory should be clear to this point. While working memory consists of a multi-component system that holds and manipulates information in short-term memory and applies attention to manage short-term memory, short-term memory merely refers to the short-term storage of information in a theory-neutral manner without entailing the manipulation or organisation of information in memory (Cowan, 2008). Complex cognitive activities including language processing require the moment-to-moment monitoring processing and maintenance of task-relevant information. Working memory perfectly presents the involvement of its multiple components and the dynamic coordination of activities among them that contribute to those activities (Baddeley and Logie, 1999).

In the experiments in my thesis, two aspects of cognitive load were systematically manipulated. One was memory load and the other was attention control. Memory load, operationalised as the capacity to hold decaying memory for a limited period of time (e.g. Baddeley and Wilson, 2002) in this work, was manipulated with different durations of ISIs (see section 1.5). Memory load has been frequently tested using the classical Sternberg memory task paradigm that involves presentation of a list of items to memorise, followed by a memory maintenance period during which the subject must maintain the list of items in memory (Sternberg, 1966, 1975). In my thesis, only the latter, temporal factor has been investigated while the former aspect, i.e., the amount of information to maintain, remained constant. Attention control (Baddeley and Hitch, 1974) (also called controlled-attention in Bialystock, 1992; Lavie and Hirst, 2004) is one of the most central functions of the central executive in working memory. This attention mechanism is supposed to control the limited cognitive resources in all forms of information processing through shifting efficient attention among foregrounding and backgrounding of task-relevant and -irrelevant information (Isaacs and Trofimovich, 2011; Rosen and Engle, 1998).

Attention control has been extensively investigated in the field of bilingual studies. Previous studies testing bilingual children (mostly 3 to 5 years old) show a higher performance in executive control tasks and cognitive advantages compared to monolingual children (Bialystock, 2005; Colzato et al., 2008; Costa et al., 2008; Prior and MacWhinney, 2010). Their advantages extend to their auditory processing (e.g. Kritzman et al., 2014). This is a consequence of the simultaneous activation and efficient control of two languages from early childhood (e.g. Blumenfeld and Marian, 2007; Dijkstra and van

Heuven, 1998; Green, 1998). They learned to efficiently switch from one language to the other in a context-appropriate manner (e.g. Costa and Santesteban, 2004; Rodriguez-Fornells et al., 2006). Also, the cognitive advantages shown by bilingual children hold true for adult bilinguals (Bialystok et al., 2012; Hilchey and Klein, 2011).

However, the picture becomes different when it comes to adult L2 learners (who are not considered bilinguals). Their L2 perception is affected more strongly by distracting background noise than L1 perception (Antoniou et al., 2013; Cutler et al., 2007; Lecumberri and Cooke, 2006; Nabelek and Donahue, 1984) due to fewer cognitive resources in L2 speech processing through reduced L2 proficiency in comparison to L1 listeners (Antoniou et al., 2013). Based on these findings, it can be assumed that L2 listeners are less successful in ignoring or shutting down task-irrelevant information. L2 listeners' difficulty in perceiving nonnative length contrasts with greater talker variability has also been reported (Sonu et al., 2013), suggesting that their L2 perception ability could not be applied to demanding speech processing situations. Based on these findings, L2 perception is expected to be more "vulnerable" than L1 perception under such demanding listening conditions with increased task demands. The reason for the differences found between the bilinguals and adult L2 learners may lie in the fact that such cognitive flexibility develops in early childhood (De Luca and Leventer, 2008) and inhibitory control is among the earliest executive functions to appear (with initial signs observed in infants, 7 to 12-months old in Anderson, 2002; De Luca and Leventer, 2008, then children display a spurt in performance on tasks of inhibition between the ages of 3 to 5 years as reported in Best et al., 2009; De Luca and Leventer, 2008).

In my experiments, the demand on attention control was manipulated by adding a task-irrelevant distracting prosodic dimension to the stimuli. Listeners were required to pay attention only to a task-relevant prosodic dimension, while ignoring the task-irrelevant one. This operationalising of attention control is important as it has been classically investigated using speech in noise (Hill and Miller, 2010), a multi-talker paradigm (Mesgarani and Chang, 2012; Rimmele et al., 2015) or vocoded speech by reducing the amount of speech information (Wild et al., 2012).

1.6.2 Factors affecting cognitive load in L2 processing

Besides the two aspects of cognitive load that were systematically manipulated and varied in my experiments, there are numerous further factors that influence task demands. Task is understood hereby as all kinds of speech activity including listening and speaking. The investigation on the influence of experimental task demands on speech process-

ing is also useful to understand the factors that influence our everyday speech processing. Bohn (1995) extensively illustrates the variables and their interactions that influence cross-language speech perception, see Figure 1.7.

The tetrahedron shows that L2 perception is influenced by numerous factors, influencing each other. Taking one of the tasks in the following experiments, discrimination task of nonnative consonant length contrast as an example, the task relates to numerous factors such as native and nonnative categories, subjects' linguistic experience and training, individual abilities, methodological differences in training procedures and testing procedures as well as stimulus materials used to assess perceptual abilities.

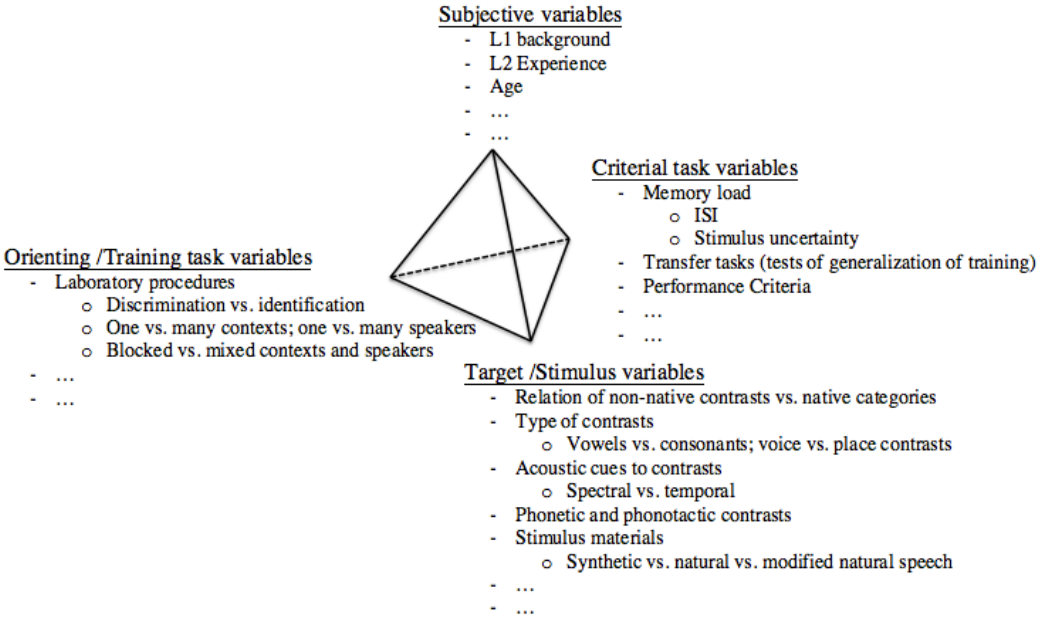

Figure 1.7 *Tetrahedron illustrating a cluster of variables and interactions in cross-language speech perception. The edges represent two-way interactions and the planes call attention to a three-way interaction, and the whole figure represents the four-way interaction of all the variables. (Figure and wording of the legend are from Bohn, 1995, 281)*

Some factors in Figure 1.7 have been more widely investigated than others in previous studies on cross-language speech perception. For instance, the studies that focus on nonnative contrasts in relation to learners' L2 proficiency investigate *subject variables* and *target/stimulus variables* and interactions between them. Such studies build the majority in the literature. Compared to them, the studies that focus on *orienting/training task variables* or *criterial task variables* constitute the minority (e.g. Antoniou et al., 2013; Cutler et al., 2007; Werker and Tees, 1984b) and this is all the more true when it comes to

the perception of cross-linguistic segmental *length* contrasts (, but some exceptions see e.g. Sonu et al., 2013; Tajima et al., 2008; Wilson et al., 2005, all about speech rate and perception of nonnative segmental length contrasts). Importantly, Figure 1.7 was developed for cross-language speech perception, though the factors shown in the figure can be also applied to speech production.

The current study therefore aims at investigating speech perception and production of nonnative segmental *length* contrasts and pitch contrasts in relation to *subject variables* (L1 background), *target/stimulus variables* (type of contrasts) and *critical task variables*, which were yet seldom studied in the field of research.

1.7 Summary

This chapter has provided the background and motivation for the following experiments in this thesis. First, I stated the importance of the investigation on L2 prosody. Deviant L2 prosody is known to impede successful communications and can lead to an undesired negative social and personal consequences. Despite this importance, L2 prosody is still understudied and undertaught in an L2 classroom. Second, I presented the most relevant aspects of the use and the form of F_0 and segmental length contrasts in Japanese and German. Japanese shows overall more lexical restrictions than German. The differences between Japanese and German prosodic systems are expected to cause difficulties in L2 learning. Third, I outlined five models and discussed their limitations and applicability to research on L2 prosody. The most crucial problem appears to define and quantify cross-language prosodic (dis)similarities, while the definition of cross-language segmental (dis)similarities (e.g. similarities between two vowels or consonants) appears to be more clear. This is because the same prosodic cue is used cross-linguistically at different linguistic levels. Fourth, I presented the stages of language processing under investigation each of which I tested in the experiments step by step. I further provided the experimental methods that tested these stages separately. For example, I exploited the established empirical notion that phonetic information decays after a while in order to test phonetic and phonological processing. The question remained open is however, whether the phonetic and phonological processing can be clearly separated only based on the factor of time. Finally, I proceeded the discussion on cognitive load to investigate L2 processing. Cognitive load is regarded as one of the key aspects to understand differences between L2 and L1 processing.

CHAPTER 2

COORDINATING LEXICAL AND PARALINGUISTIC USE OF F_0 IN L2 PRODUCTION

2.1 Introduction

The first study[1] documents semi-spontaneous L2 productions by German learners of Japanese (henceforth L2 speakers in this chapter). Documenting the L2 speakers' production data in comparison to those of L1 speakers is important as a starting point of this thesis. If L2 speakers show difficulties or deviant forms in their productions, it is worth investigating the sources of such difficulties. If, however, no difficulties are shown, the investigation would not be necessary.

As presented in chapter 1.3, the same prosodic property is used at different linguistic levels across languages. In this study, the lexical use of F_0 in Japanese and the non-lexical (post-lexical and paralinguistic) use of F_0 in German are of particular interest.

The study aims at testing how L2 speakers whose L1 and L2 exhibit differences in the use of F_0 (like those found between Japanese and German) manage to coordinate the lexical and paralinguistic use of F_0 in producing L2 utterances appropriately. To this end, I conducted a semi-spontaneous production experiment in which participants produced the same words three times in a row until they succeeded in getting the attention of an imaginary waiter in a crowded and noisy bar. In such a situation, an attitudinal change due to increasing frustration is expected, because the utterance was not heard by the interlocutor. I exploited the fact that F_0 of a Japanese pitch accent is lexically determined, as opposed to the German one and can be freely used for a paralinguistic purpose. In this

[1] A preliminary and shorter version of this chapter was published as a conference paper: Asano, Y. (2015). Integrating lexical and paralinguistic F_0 in L2 production. In *Proceedings of the 18th International Congress on Phonetic Sciences, Glasgow, UK*, paper number 577.

way, a situation where L2 speakers faced the uses of F_0 at different linguistic levels across languages was elicited. This prosodic adaptation in order to add paralinguistic meaning can be treated as increasing cognitive load (Sweller, 1988). I will analyse whether the deviation of L2 speakers' productions from the L1 is aggravated further when paralinguistic prosody comes into play and cognitive load increases.

To embed the current study into a psycholinguistic model for speech production, the phonological encoding in Levelt's blueprint of the speaker (Levelt, 1989, 1999) is most relevant. In the model, it was claimed that phonological representations of a word, such as its syllabic and metrical structure (feet), are firstly generated once a word is selected. The phonological segments are made available incrementally in a left to right manner. Then, phonologically syllabified words are combined to form a phonological phrase. A phonological phrase in Levelt's model is equivalent to an intermediate phrase (ip) in Metrical phonology (Nespor and Vogel, 1986). Phonological phrases are considered to build metrical units in utterance production (Bock, 1982; Garrett, 1982; Levelt, 1989, 1999; Van Wijk, 1987). Finally, metrical units of phonological phrases are concatenated to build an intonational phrase. Along with this process, Levelt postulated *The Prosody Generator* that requires input from three sources and outputs prosodic units. The three sources are 1) the output from the grammatical encoder that unfolds surface syntactic structure, 2) phonological form information (metrical structure of the words and their segmental content) and 3) intonational meaning, which provides the illocutionary force of the utterance, the emotions and attitudes of the speaker, see Figure 2.1. Such information has consequences for the global aspects of the intonation pattern (Wheeldon, 2000, 260). Figure 2.1 shows that lexical prosody is assigned to the metrical and segmental spellout, before a global prosody is generated in the prosody generator. This means that lexical prosody is processed before the post-lexical and paralinguistic uses of prosody are considered, and that the post-lexical and paralinguistic uses of prosody will be generated in a way allowing no interference to lexical prosody. The model therefore suggests that the lexical use of prosody outweighs the paralinguistic use of prosody in speech production. Translating this assumption into the current experiment, Japanese lexical use of F_0 should be maintained even while conveying paralinguistic information.

A prosodic adaptation for a paralinguistic purpose is the object of the current study. Prosodic adaptations have been mostly studied in terms of hyperarticulation (Lombard, 1911; Oviatt et al., 1996; Stent et al., 2008). The term hyperarticulation covers a wide range of articulatory adaptations under different intentional, interpersonal or environmental but also pragmatic factors. Despite different experimental situations that focused

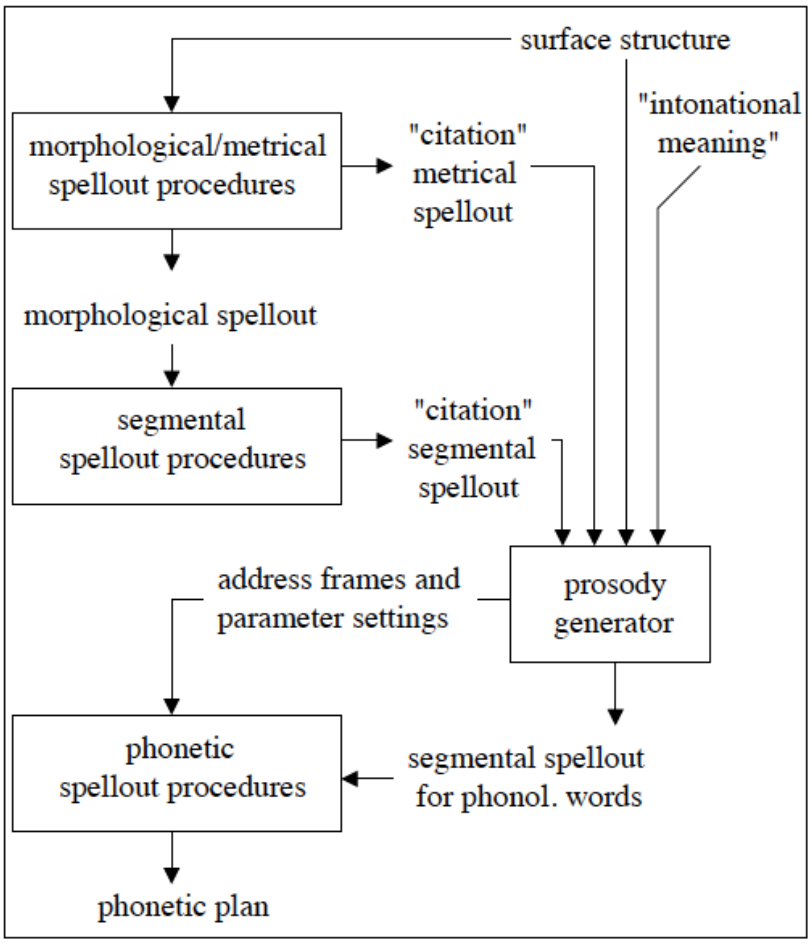

Figure 2.1 *The phonological encoding model for connected speech from Levelt (Chapter 10 1989, 366).*

on different languages, previous studies on hyperarticulated speech showed similar results such as slower speaking rate, greater pitch range or higher pitch (Moon and Lindblom, 1994; Oviatt et al., 1996; Stent et al., 2008). Regarding Japanese, the coordination of the lexical and paralinguistic uses of F_0 has been studied in the field of infant-directed speech (Ishihara, 2011; Kitamura and Burnham, 2003; Nagahara, 1994) or corrective focus (Maekawa, 2004). Previous studies agree that Japanese L1 speakers do not phonologically change a pitch accent type in these different speech conditions, because the lexical restriction of the Japanese lexical pitch accent prevents speakers from modifying local F_0 movement for a post-lexical or paralinguistic purpose. As for German, the lexical and

paralinguistic uses of F_0 do not compete with each other, because F_0 is not used lexically, but primarily to convey emotion or attitudinal states of a speaker (Baumann and Grice, 2006; Gibbon, 1998; Liscombe, 2007), syntactic structure of a sentence (Lingel et al., 2006; O'Brien et al., 2014), information structure such as topic vs. focus (Braun, 2006; Féry, 1993) and illocutionary force (question vs. statement, warning) (Petrone and Niebuhr, 2014). Based on these findings and the language differences between Japanese and German, the following hypotheses are stated: Japanese L1 speakers will not phonologically change a Japanese pitch accent in the repeated utterances, while German L2 speakers will phonologically vary F_0 contours, because in German F_0 conveys the speaker's attitude and emotion.

Regarding the performance in the L2 (=Japanese), the way in which F_0 is used in the speakers' L1 is expected to interfere with the production of F_0 contours in the L2. Perception studies show that L2 speakers often fail to pay attention to the *lexically* meaningful cue in their L2 because of the lack of the contrast of this cue in their L1, when L1 and L2 do not share the same mappings of lexical contrasts (Altmann et al., 2012; Braun and Johnson, 2011; Cutler et al., 2006; McAllister et al., 2002). Assuming that the findings in these perception studies hold true for L2 production, L2 speakers are expected to be insensitive in realising a prosodic property in the same way as L1 speakers do, when this property does not convey lexically meaningful information. I therefore expect that the German L2 speakers will fail to be faithful to the lexically fixed Japanese pitch accent, and that they will vary it.

Besides the analysis of F_0, the change of the speaking rate in the repeated utterances was analysed additionally by measuring the total durations of the utterances. As for the total durations, I hypothesise that both the L1 and L2 speakers' groups will show longer total durations in the repetitions, because longer utterance durations were found in the hyperarticulated speech (Moon and Lindblom, 1994; Oviatt et al., 1996; Stent et al., 2008). Additionally, I expect that the L1 speakers will produce longer durations to a greater extent than L2 speakers, because the Japanese L1 speakers will not vary F_0 and they thus have one prosodic property less than the German L2 speakers.

Finally, one of the Japanese target words contained a geminate consonant, which was nonnative for the German participants, because a consonant length contrast is not used lexically in German. Since an appropriate timing of a nonnative geminate is known to be difficult to produce even with a considerable amount of exposure to an L2 (Kabak et al., 2011), I hypothesise that the L2 speakers will produce the Japanese nasal geminate with

shorter durations than L1 speakers. I will examine this hypothesis by measuring relative segmental durations with respect to the total utterance durations.

2.2 Experiment

2.2.1 Methods

Participants

15 speakers of Tokyo-Japanese (8 females/ 7 males, aged between 19 and 36 years, mean age = 25.1 years) and 15 speakers of Standard German who were learners of Japanese (6 females/ 9 males, aged between 22 and 35 years, mean age = 28.9 years) participated in the experiment. None of the participants had any self-reported speech or hearing deficits. They were all unaware of the purpose of the experiment. Both learner groups were rated with respect to their proficiency in L2. Japanese L2 speakers completed a German C-Test (i.e. Coleman et al., 1994). On a scale from 0 (lowest) to 60 (highest) they ranged between 12 and 60, mean = 41.2, sd = 13.0. German L2 speakers performed a Japanese Simple Performance-Oriented Test (Spot-test) (Hatasa and Tohsaku, 1997) on a scale from 0 (lowest) to 50 (highest) ranging between 4 and 48, mean = 25.7, sd = 16.0. All L2 speakers had learned the L2 for at least 2 months (Japanese L2 speakers ranging between 2 and 216 months, mean = 50.5 months, sd = 58.7 months; German L2 speakers ranging between 2 – 420 months, mean = 111.3 months, sd = 115.8 months). Two of the Japanese L2 speakers had never stayed in a German-speaking country (ranging between 0 – 96 months, mean = 27 months, sd = 31.7 months) and all had learned English as an L2 since they were 12 years old. Seven of German L2 speakers had never stayed in Japan (ranging between 0 – 144 months, mean = 22.4 months, sd = 39.0 months) and none of them had learned another language which has lexical pitch-accent or tone languages. Based on the the analysis of collinearity (Belsley et al., 1980), only the L2 test scores were used to analyse an effect of L2 proficiency on the performance by L2 speakers in the experiment. Speaking ahead of results, L2 proficiency did not predict their performance, so that the report on the analysis with L2 proficiency will not be presented.

Materials

The target words were very frequent Japanese words, *sumimasen* ([sɯ̥.mɪ.ma.sẽ.n]) and *konnichiwa* ([kõ.n.ni.t͡ɕi.ɯɸ$^{\beta}$ä]). The first word means *excuse me* and the second one *hello*.

Both words can be used for calling someone's attention. *Sumimasen* is five-moraic and contains a lexically specified pitch fall associated with the penultimate mora in the word, "*se*". *Konnichiwa* is also five-moraic with a nasal geminate and contains a lexically specified pitch fall associated with the ultimate mora, "*wa*". However, the pitch accent of *konnichiwa* is not realised as a pitch fall when it is produced as an isolated word, which is the case for the current experimental setting. The stylised contours of the two target words are shown in Figure 2.2. The words may be realised with an initial low (Gussenhoven, 2004; Vance, 1987). These two words build an ideal contrastive pair to investigate the influence of the (perceivable) lexical pitch accent on L2 production.

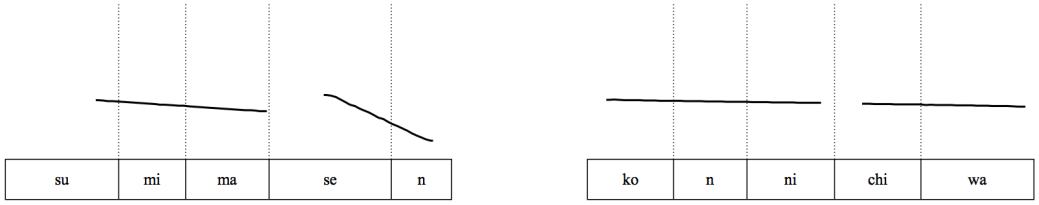

Figure 2.2 *Stylised contours for sumimasen (left) and konnichiwa (right). An initial low of the utterances may occur optionally.*

In order to analyse the influence from the L2 speakers' L1 to the L2, the German participants were additionally asked to produce a German target word *Entschuldigung* ([ɛn̥t.ˈʃʊl.dɪ.gʊŋ]), meaning *excuse me*. This word can be used in the same context as *sumimasen*.

Design and Procedure

Following the procedure in Prieto and Roseano (2010), materials were presented with descriptions of short scenes. The task was to produce the target words in a given context: For *sumimasen*, participants were asked to imagine to be in a crowded and noisy bar and were asked to attract a waiter's attention to order a drink. For *konnichiwa*, they were asked to imagine to enter a grocery store and call a grocer. In both situations, participants were asked to repeat the same words, because the utterance was not successfully conveyed. Additionally, the German participants were asked to produce and repeat *Entschuldigung* while imagining the same context as for *sumimasen*. Each trial consisted of four slides. The description below is an example of a trial. Descriptions were written in the participants' L1 with a picture of a typical situation, see the description below and

Figure 2.3. The experiment was designed with Microsoft PowerPoint 2011 and was presented on a Macintosh G3 laptop.

Slide 1 You are in a crowded noisy bar in Japan. Please call a waiter by saying <*sumimasen*>.

Slide 2 He did not hear you. You are a little bit frustrated. Please try it again by saying <*sumimasen*>.

Slide 3 He still did not notice it. You are very frustrated by now. Please try it the last time by saying <*sumimasen*>.

Slide 4 He finally heard you and is coming to you. Congratulations!

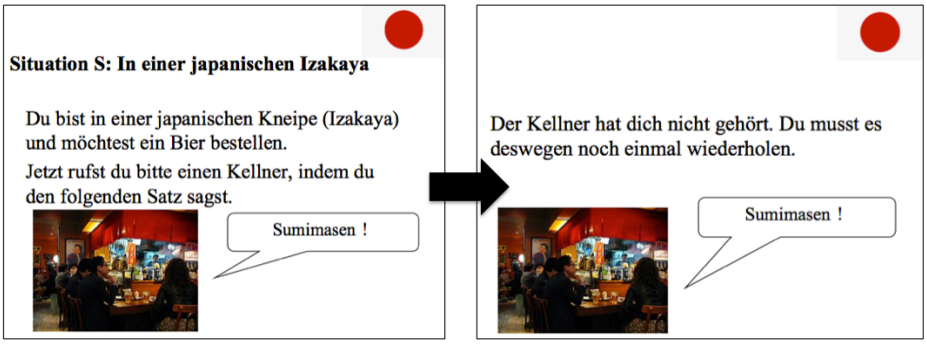

Figure 2.3 *Example slides (Slide 1, left and Slide 2 right) presented to German L2 speakers on the screen.*

Three attempts of each of the two Japanese words (and of the German word) were recorded for each speaker. Moreover, 8 filler tasks were provided for both Japanese and German participants, for which they produced 8 Japanese words or short sentences in various situations. Participants were tested individually in a quiet room. Japanese participants were recruited and tested at the Tokyo University of Foreign Studies in Japan and German participants at the Ruhr-University Bochum and Heinlich-Heine-University Düsseldorf in Germany. Their responses were digitally recorded onto a computer (44.1kHz, 16Bit) using a unidirectional short-range microphone. In total 225 utterances were used for the analysis (90 each for *sumimasen* and for *konnichiwa* and 45 for *Entschuldigung*). After the experiment, participants were asked whether they knew something about Japanese lexical pitch accents. All L2 speakers reported to have had heard about the Japanese lexical pitch accents in a Japanese course.

F_0 extraction and annotation

F_0 contours were computed using the F_0 tracking algorithm in the *Praat* toolkit (Boersma and Weenink, 2011). F_0 ranges were manually inspected and corrected if there were obvious errors such as octave jumps. A default range of 70-350 Hz for males and 100-500 Hz for females was used. Then, as a preparation for a manual annotation, segmental boundaries were marked using *Praat* applying standard segmentation criteria (Turk et al., 2005). The start of the nasal was the point when the amplitude in the waveform dropped and the waveform showed less high frequency components (drop in high frequency energy in spectrogram).

The manual annotation was carried out by two coders. One coder was a Japanese L1 speaker (the author) who was highly proficient in German as L2. The other coder was a German L1 speaker learning Japanese as L2. Japanese words were annotated in the following way: Each mora was associated to either an H (= high) or L(= low) pitch (if needed, upsteps with ˆ and downsteps with !), resulting in a sequence of five letters consisting of H or L (e.g. HHHHL). The annotation codes of the Japanese J-ToBI system (Venditti, 1997) were not used for the following reasons. First, H*+L is the only phonological category in the Japanese ToBI system and this would not have been sufficient to cover expected deviant variations produced by the German L2 speakers. Second, it was crucial to know where (in which mora) a fall or a rise occurred, because the L2 speakers were expected to produce a pitch fall or a rise in deviant mora from the appropriate one as shown in Figure 2.4. Using only the J-ToBI, the position of a deviant pitch accent would not have become obvious. The annotation that was associated with each mora was therefore more meaningful for the current study.

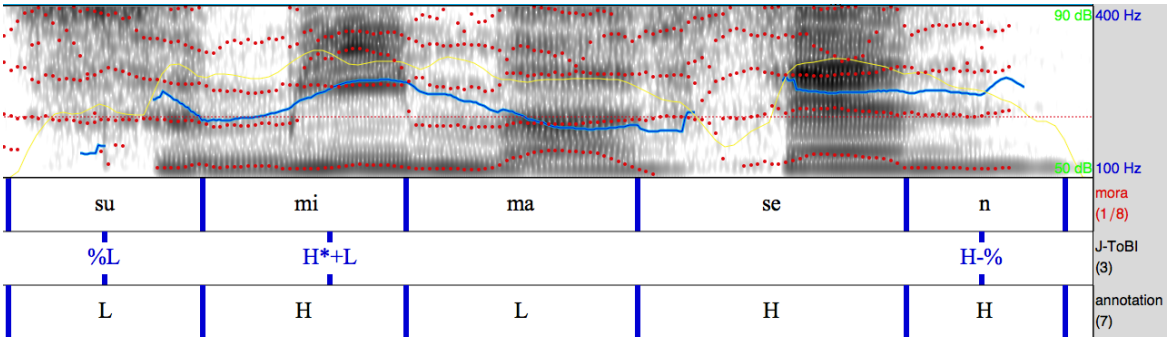

Figure 2.4 *An example of an annotated L2 production with a deviant pitch accent position.*

The German word was annotated using the pitch accent categories of the German ToBI system (Grice et al., 2005). The available accent types were six basic pitch accents

(H*, L*, L*+H, L+H*, H+L*, H+!H*) (Baumann et al., 2001). Additionally, the five of the basic accents containing H tones can be downstepped or upstepped, thus increasing the inventory from 6 to 11 accents (ibid.). The available boundary tones were L-%, L-H%, H-% and H-^H%.

Finally, the Kappa Coefficient of Agreement (Cohen, 1960), a common measure of inter-rater reliability was computed using the statistical program *MedCalc* by combining the annotations of the pitch accent and the final boundary tone.

2.2.2 Results

Results of F_0

***Sumimasen*:** The interrater reliability score for the utterances produced by the L1 speakers had a Kappa of 1.00 (SE = 0), and that produced by the L2 speakers 0.59 (SE = 0.08, 95% Confidence Interval, henceforth CI = [0.44, 0.74]). The former Kappa value signals an extraordinarily high level of agreement, while the latter shows a moderate level (Landis and Koch, 1977). In case of a disagreement between two coders, the results from the Japanese L1 coder are reported as the main source. Disagreements were found in the perception of downsteps and upsteps or in the perception of two different pitch accents or boundary tones.

All L1 speakers' utterances were coded as HHHHL or LHHHL with an initial low (Vance, 1987), a typical form of Japanese utterances. L1 speakers changed neither pitch accents nor boundary tones across repetitions. In the L2 speakers' data, twenty utterances were coded as HHHHL or LHHHL, the forms produced also by the L1 speakers. Further twelve utterances were coded as HHHHH, which were flat contours. Then, there were utterances in which the lexical pitch fall occurred before the mora *se*, and which ended with a falling boundary tone: Six utterances were coded as HHLLL, three as HHHLL and one as HLLLL. Other utterances had a rising end: One utterance was coded as HHHLH, two utterances as LLLLH, see Figure 2.5. Contrary to the Japanese L1 speakers' data, German L2 speakers showed variations both in pitch accents and final boundary tones. Noticeably, rising contours were found only in the first or second attempts. Considering the changes of a pitch accent or a final boundary tone implemented by the same speaker across repetitions, twelve L2 speakers varied either a pitch accent (seven of them) or a boundary tone (twelve of them).

***Konnichiwa*:** The inter-rater reliability score for the utterances produced by the L1 speakers had a Kappa of 1.00 (SE = 0) and that produced by the L2 speakers 0.64 (SE =

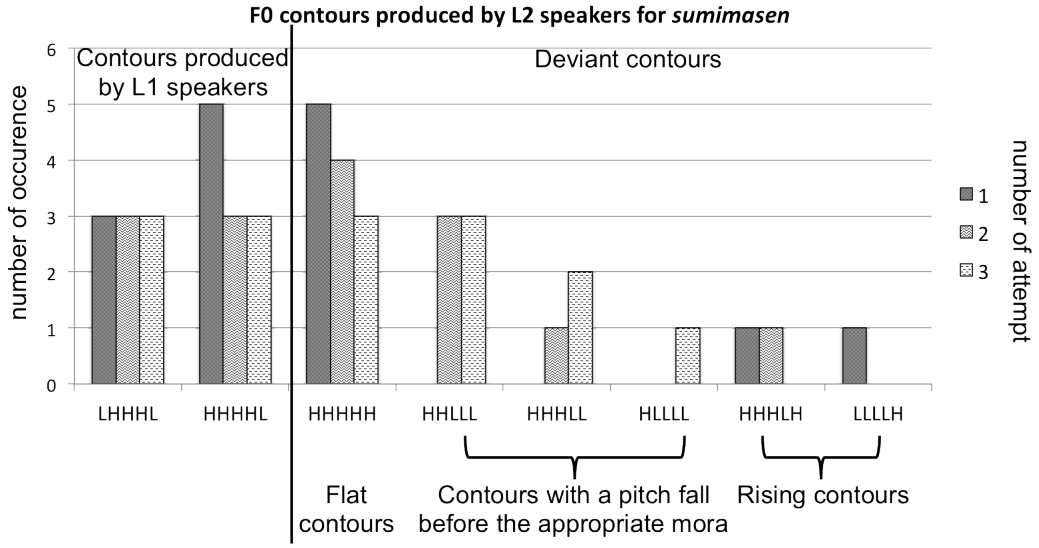

Figure 2.5 *Number of occurence of the contours for the word* sumimasen *in each attempt produced by the L2 speakers. The legend (1,2,3) refers to the number of attempt.*

0.09, 95% CI = [0.46, 0.82]). As it was the case for *sumimasen*, the former Kappa value shows an extraordinarily high level of agreement, while the latter indicates a moderate level (Landis and Koch, 1977).

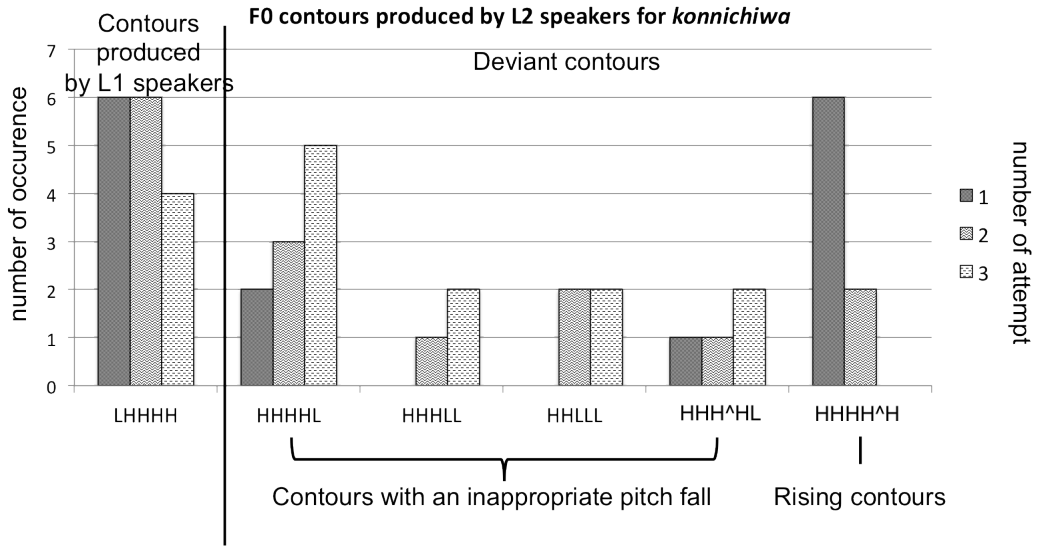

Figure 2.6 *Number of occurence of the contours for the word* konnichiwa *in each attempt produced by the L2 speakers. The legend (1,2,3) refers to the number of attempts.*

All L1 speakers' utterances were coded as either HHHHH or LHHHH with an initial low. As it was case for *sumimasen*, L1 speakers changed neither pitch accents nor boundary tones across repetitions. In the L2 speakers' data, sixteen utterances were coded as LHHHH, i.e. the form produced by the L1 speakers. Utterances with an incorrect pitch fall and a falling final boundary tone such as HHHHL (N=10), HHHLL (N=3), HHLLL (N=4) or HHH^HL (N=4) were detected as deviant forms. Furthermore, eight utterances showed a rising contour; HHHH^H, see Figure 2.6. Note that rising contours were found only in the first or second attempts. Both the standard form and the deviant forms occurred in all attempts.

***Entschuldigung*:** The inter-rater reliability score showed a Kappa of 0.86 (SE = 0.06, 95% CI = [0.73, 0.97]). Thirty-five utterances were coded as H* L-%. The other contours had rising final boundary tones; six with L* and seven with H*, see Figure 2.7. Rising contours occurred more frequently in the first attempt, followed by the second, and never occurred in the third attempt.

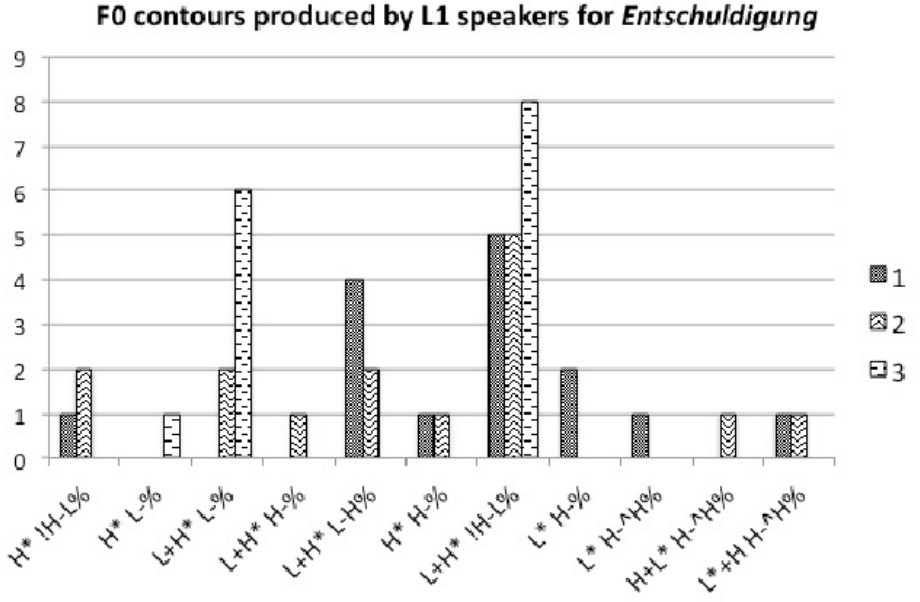

Figure 2.7 *Number of occurence of the contours for the word* Entschuldigung *in each attempt produced by the L1 speakers. The legend (1,2,3) refers to the number of attempts.*

Results of the total durations

Due to the small number of the samples, I will report descriptive mean values and 95% CI error bars instead of running statistical analyses that require a larger number of samples.

Additionally, I will also report the Cohen's *d* as a measurement of effect size (Cohen, 1969, 1992), since the amount of the data was small and an effect size would help more to interpret the data. By convention, an effect size is small when $d = 0.2$, medium when $d = 0.5$ and large when ($d \geq 0.8$). The plot with inferential error bars includes essentially all the information provided by a hypothesis-testing procedure plus a graphic signal of how much uncertainty is in the data. Regarding a between-subject variable, the CIs of two groups that do not overlap or just touch indicate a population difference, and p is approximately less than .01. If the CIs overlap by no more than half of the length of one whisker of the CI, there is a degree of evidence of a difference, and p is approximately less than .05. However, the data may be interpreted without invoking p-values (Cumming, 2011, 13).

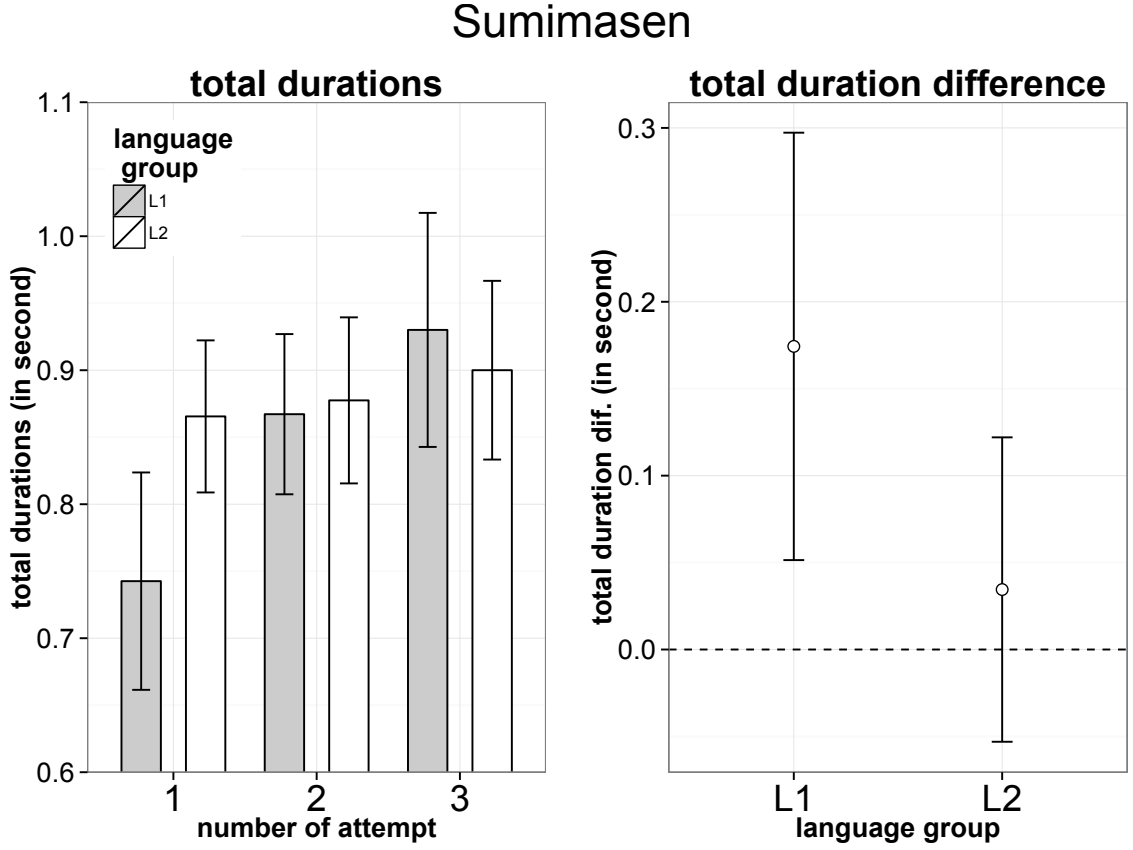

Figure 2.8 *Mean total durations for sumimasen with 95% CI bars for each attempt and language group (left) and mean total duration differences between the 1st and 3rd for the word sumimasen with 95% CI bars for each language group (right).*

***Sumimasen*:** In order to provide an overview of the data, the right plot in Figure 2.8 shows mean total durations and 95% CI bars for each attempt in each group. The result shows that there was no difference between the L1 and L2 speakers' utterance durations across the attempts (M = 0.88 s, 95% CI [0.82, 0.94] for the L1 speakers and M = 0.84 s, 95% CI [0.77, 0.92] for the L2 speakers, d = -0.17). At first sight, the plot seems to show the tendency that only the L1 speakers produced longer durations in the repetitions, so that an interaction between *language group* and *number of attempt* could be expected. Note that the overlap of the separate CIs is irrelevant with a repeated measure so that the CI on the difference needs to be calculated in order to interpret the difference (Cumming, 2011). To this end, the differences in total durations between the 1st and 3rd attempt (3rd - 1st) were calculated, see the right plot in Figure 2.8. The plot shows that only the L1 speakers produced longer durations in the repetitions, while the L2 speakers did not, because the 95% CIs for the L1 speakers do not include 0. Moreover, there was a tendency for an interaction between *language group* and *number of attempt* (1st vs. 3rd) (M = 0.17 s, 95% CI [0.05, 0.30] for the L1 speakers and M = 0.03 s, 95% CI [-0.05, 0.12] for the L2 speakers, d = 0.73). The relatively large effect size also supported that the two groups shown in the plot tended to differ.

***Konnichiwa*:** In the same way as for the analyses of *sumimasen*, the left plot in Figure 2.9 provides an overview of the data showing mean total durations and 95% CI bars for each attempt in each group. The plot indicates that only the L1 speakers produced longer durations in the repetitions, so that an interaction between *language group* and *number of attempt* is expected.

Analogue to *sumimasen*, the differences in the total durations between the 1st and 3rd attempt (3rd - 1st) were calculated, see the right plot in Figure 2.9. The plot shows that only the durational difference for the L1 speakers does not cross 0. This means that only the L1 speakers produced longer durations in the repetitions. Moreover, there was an interaction between *language group* and *number of attempt* (1st vs. 3rd), because about half of the length of the whisker of the CIs overlapped (M = 0.28 s, 95% CI [0.08, 0.48] for the L1 speakers and M = 0.10 s, 95% CI [-0.03, 0.21] for the L2 speakers, d = 0.73). The large effect size also shows that the two groups shown in the plot differed.

***Entschuldigung*:** The data set for *Entschuldigung* contained only the German participants' data. In order to analyse whether they produced longer utterances in the repetitions, the differences between the 1st and 3rd attempt were calculated. It was found that the differences were consistently larger than 0, suggesting that the German participants

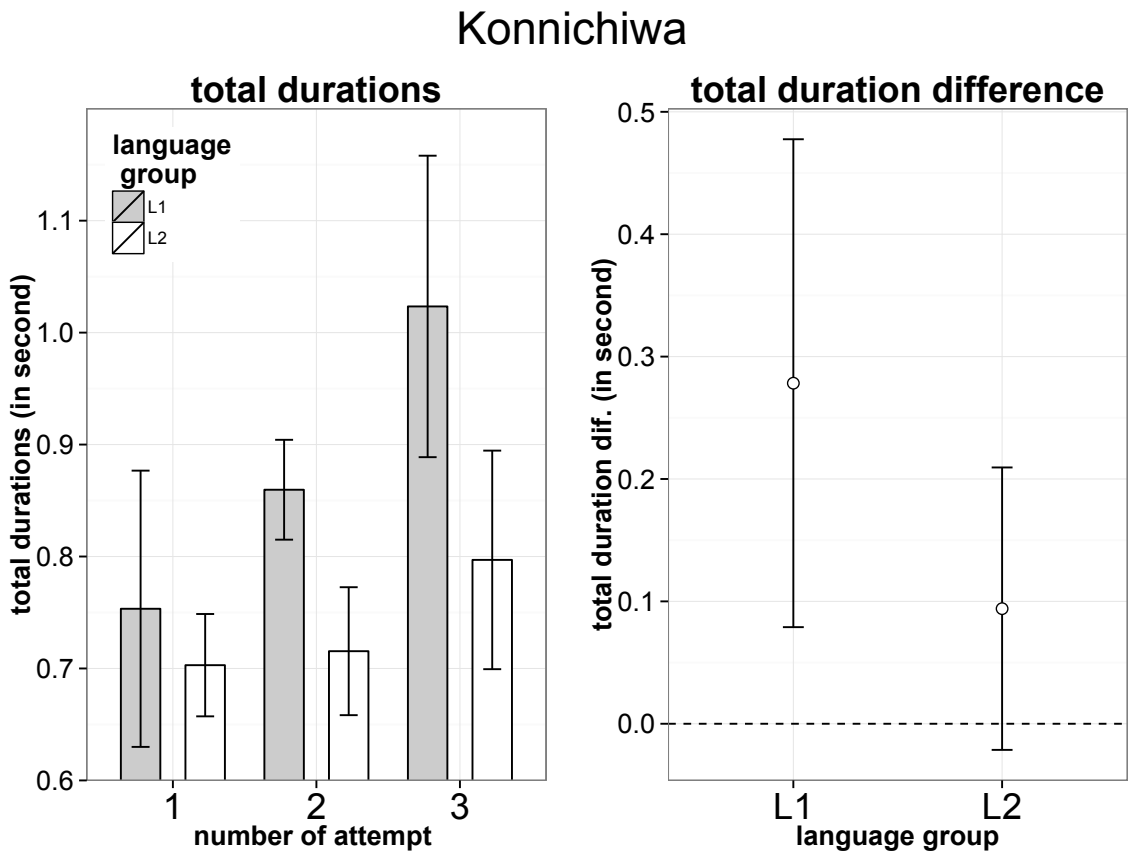

Figure 2.9 *Mean total durations for konnichiwa with 95% CI bars for each attempt and language group (left) and mean total duration differences between the 1st and 3rd for konnichiwa with 95% CI bars for each language group (right).*

produced longer utterances in the 3rd attempt than in the 1st attempt (M = 0.14 s, 95% CI [0.03, 0.24]).

Results of the production of a nonnative geminate

In order to investigate whether the L2 speakers produced the nasal geminate in the target word *konnichiwa* as long as the L1 speakers, the relative segmental durations of <ko>, <n:i>[2], <chi> and <wa> with respect to the total durations were analysed. The mea-

[2] The analysed segment <n:i> consists of two morae <n> and <ni> in Japanese. While the Japanese L1 speakers produced the boundary between <n> and <ni> clearly, the German L2 speakers did not. In

surement of relative durations is more suitable than the one of absolute durations in this study, because the utterances were lengthened in the repetitions and only relative durations can provide us rate-independent information (Idemaru and Guion-Anderson, 2010). Mean relative durations of <ko>, <n:i>, <chi> and <wa> with 95% CI bars for each language group are shown in Figure 2.10.

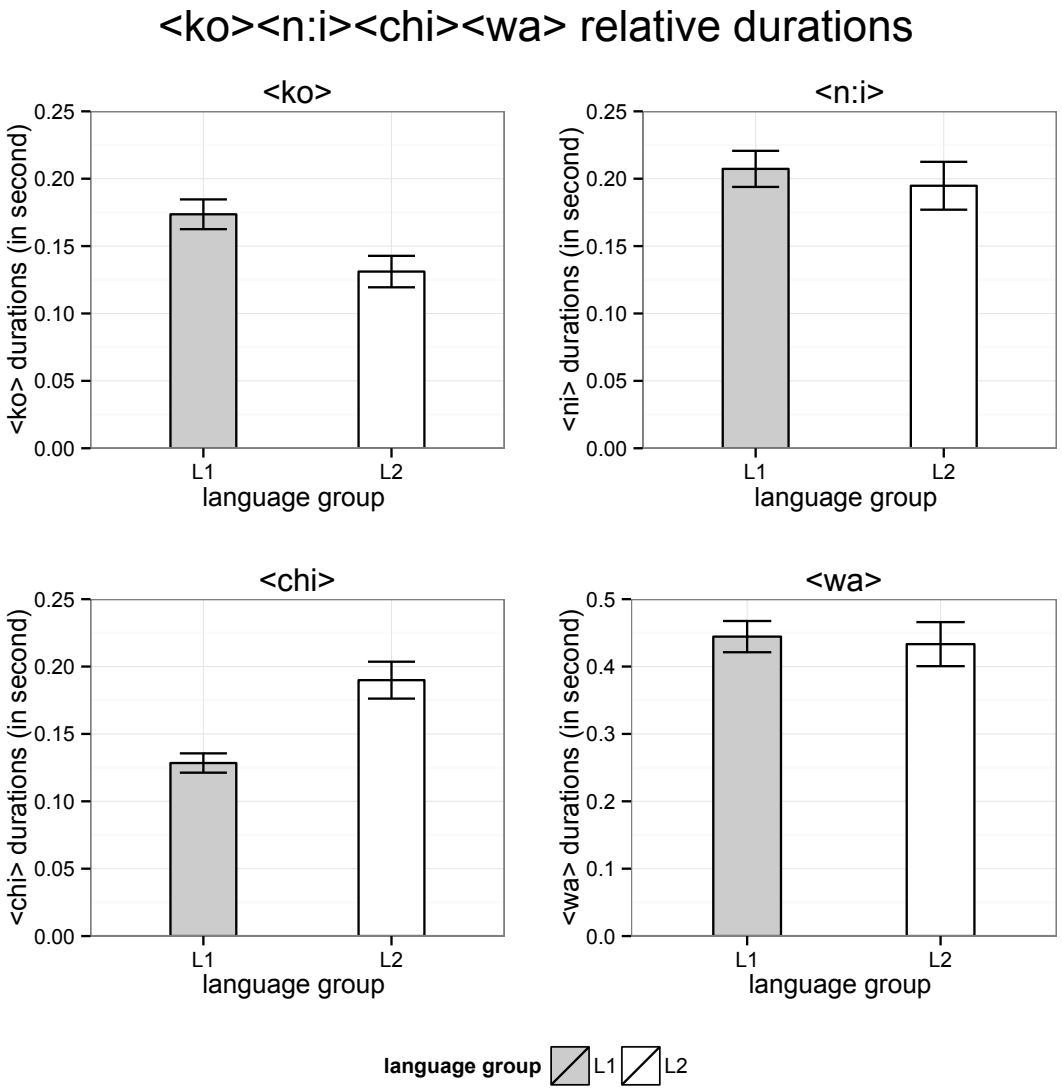

Figure 2.10 *Mean relative durations of <ko>, <n:i>, <chi> and <wa> with 95% CI bars for each language group.*

order to keep the consistency in the number of boundary of the annotation, I decided not to put a boundary between <n> and <ni>.

The plot shows that the L2 speakers produced the preceding mora (<ko>) (= one mora before the nasal geminate) shorter than the L1 speakers (M = 0.17 s, 95% CI [0.16, 0.18] for the L1 speakers and M = 0.13 s, 95% CI [0.12, 0.14] for the L2 speakers, d = -1.26). On the contrary, the following mora (<chi>) (= one mora after the geminate) was produced longer by the L2 the speakers than by the L1 speakers (M = 0.13 s, 95% CI [0.12, 0.14] for the L1 speakers and M = 0.19 s, 95% CI [0.18, 0.20] for the L2 speakers, d = 2.08). Both CI bars and d confirm a great difference between the L1 and L2 speakers' durations. It can be assumed that the L2 speakers produced "penultimate stress" as they are used to doing so in their L1, German. Notably, the relative durations of the nasal geminate itself <n:i> did not differ between the two language groups (M = 0.21 s, 95% CI [0.19, 0.22] for the L1 speakers and M = 0.19 s, 95% CI [0.18, 0.21] for the L2 speakers, d = -0.28). Also the relative durations of <wa> did not differ between the two language groups (M = 0.44 s, 95% CI [0.42, 0.47] for the L1 speakers and M = 0.43 s, 95% CI [0.40, 0.47] for the L2 speakers, d = -0.14). It is remarkable that the relative durations of the preceding and of the following mora between the L1 and L2 speakers' groups differed from each other, but those of the segment containing the nasal geminate itself did not.

2.3 Discussion

The study examined the coordination of the lexical and paralinguistic uses of F_0 in Japanese L1 and L2 productions. The changes in F_0 contours were analysed by manually annotating the productions. It was predicted that the Japanese L1 speakers would not modify F_0 contours in the repetitions, while the German L2 speakers would as they were expected to do so in their L1 (German) utterances. The results showed that the Japanese L1 speakers changed neither pitch accents nor boundary tones in the repetitions. The Japanese L1 speakers remained, irrespectively of the repetitions, faithful to the lexical formation of F_0. They constantly produced the falling lexical pitch accent in the target word *sumimasen* and flat pitch contours in *konnichiwa*. On the contrary, the German L2 speakers varied the type of pitch accents and boundary tones regardless of the number of attempt. The additional analysis of their German utterances (*Entschuldigung*) revealed that the German participants varied F_0 in their L1 in a similar way as they did in their L2 productions.

Since all L2 speakers reported to have had heard something about the Japanese lexical pitch accents in their Japanese courses, the following interpretations are possible: One interpretation is that they did not know the position of the pitch accent in the target

words even though they generally knew about the Japanese pitch accent. The second interpretation is that they knew the correct position of the pitch accent in the target words and they were willing to produce the pitch accent correctly, but they failed to apply their theoretical knowledge to their production. The third interpretation is that they ignored the lexical restriction in the use of F_0 in Japanese and modified F_0 contours, even though they knew that the lexical pitch accents exist.

Note that I predicted that the deviant variations in the L2 speakers' productions from those by the L1 speakers could be further aggravated when paralinguistic prosody came into play and cognitive load increased. The results revealed that the L2 speakers produced deviant forms already in the first attempt. This result lends itself to the interpretation that the L2 learners had difficulties in producing appropriate L2 prosody independently from conveying the attitudinal change (= increased frustration), in other words, that the difficulty in producing an L2 utterance with appropriate L2 prosodic forms was not exacerbated by adding paralinguistic information in the second and third attempt. However, it should be noted that even the first attempt already included a certain speech act (e.g. calling) (e.g. Searle, 1969), so that it was not the case that the participants did not have to convey any paralinguistic information in the first attempt.

There may be criticism that the participants might have had not felt stressed even in the second or third attempt, so that there was actually no "increased frustration" that had to be conveyed in the repetitions. I argue that it was not the case. First, the lengthening of the utterances were generally observed in the repetitions, which is known as a prosodic adaptation in hyperarticulated speech. This can be evidence for signalling the increased frustration in the repetitions. Second, rising pitch accents and boundary tones were mostly found in the first and second attempt both in the German participants' L1 and L2 utterances. Such rising contours are known to signal politeness and friendliness in English (Chun, 2002). Given that this hold true in German, it might be most probably the case that the German participants signalled politeness and friendliness in the first and second attempts by producing rising contours. In the repetitions, participants were instructed to become more frustrated due to the unsuccessful communication, and tried to pursue a response with falling F_0 contours which might sound less polite (ibid.). Finally, it has to be mentioned that the participants did not necessarily have to feel stressed in reality to accomplish the task appropriately, but it was sufficient to follow the task instruction and produce the words as if they had felt stressed.

In the introduction in this chapter, Levelt's prosody generator was cited in order to support the prediction that lexical prosody outweighs paralinguistic prosody and that

paralinguistic prosody is generated in a way that it does not interfere with lexical prosody, because lexical prosody is processed prior to paralinguistic prosody in the utterance formation. The Japanese L1 speakers' results confirm this prediction. Furthermore, the results of this study can be embedded into the discussion on psycholinguistic models for the acquisition of L2 segments into L2 prosody as follows. So and Best (2008) and So and Best (2011) showed that English L1 listeners assimilated Mandarin Chinese tones to their intonational categories. They claimed that the cross-language assimilation takes place across different linguistic levels (i.e. across the lexical and post-lexical levels), depending on the linguistic level at which F_0 conveys meaningful contrasts in a language. The current study showed interference across different linguistic levels in L2 production, which was reported by So and Best (2011) and So and Best (2008) in L2 perception.

The Japanese L1 speakers did not change F_0 in the repetitions, but they lengthened their utterances. The German participants also lengthened their utterances in their L1. Thus, the productions of the L1 speakers' groups replicated the findings in the previous studies on hyperarticulation, suggesting that the task in the current study was successful. However, this prosodic adaptation was not constantly observed in the L2 utterances. For *konnichiwa*, a clear interaction between *language group* and *number of attempt* was found; only the Japanese L1 speakers produced longer total durations in the repetitions, while the German L2 speakers did not. For *sumimasen*, the tendency of an interaction was found. As discussed in section 1.6.1, L2 processing is known to be more affected by increasing cognitive load because less cognitive resources are available in L2 processing due to reduced L2 proficiency in comparison to L1 processing (Antoniou et al., 2013). It might be possible that it was too difficult for the L2 speakers to concentrate simultaneously on the appropriate formation of both F_0 and segmental durations in the L2 utterances due to their limited automatised cognitive responses in L2 processing.

The further analysis of the segmental durations of *konnichiwa* revealed that the German L2 speakers produced the preceding mora much shorter and the following mora much longer than the L1 speakers. Previous studies report that multiple acoustic features covary with the short and long stop consonant distinction in Japanese. In disyllabic CV(C)CV words with single and geminate stops, preceding vowels were consistently reported to be longer, while following vowels were shorter in the context of geminates than singletons by about 2-20 ms (Campbell, 1999; Fukui, 1978; Han, 1994; Hirata, 2007; Homma, 1981; Idemaru and Guion, 2008; Kawahara, 2006; Ofuka, 2003; Ofuka et al., 2005), although this durational difference was not large enough to affect the total word duration and the mora counts. Interestingly, the L2 data in this study showed an opposite

covariation of the one found in the L1 data in the previous study. This deviant covariation found in the L2 utterances in this study might have contributed to the perceptual impression reported by the L1 annotator that the L2 speakers had failed to produce the geminate. Further research is needed to systematically analyse the opposite covariation found in this study.

Finally, the inter-rater reliability scores reported in this study should be discussed. The following order of the scores was found: The scores for the Japanese L1 speakers were greater than those for the German L1 speakers followed by those for the German L2 speakers. This order suggests that the greater the variation of pitch accents and boundary tones, the poorer the scores of the agreement, suggesting a general practical problem in annotating data with variations employing more than one coder. This practical problem became more serious in annotating L2 data that are characterised not only as the mixture of L1 and L2 prosody, which makes the data more complex, but they also show the dynamic character of an interlanguage (Selinker, 1972). In this study, one L1 and one L2 coder of the target words annotated the utterances. This different annotators' language backgrounds may also have influenced the annotation. It is not clear whether one should employ an L1 or an L2 coder in such a situation.

Taken together, the results clearly confirm the language-specific ways to modify F_0 for signalling paralinguistic information and the negative transfer from one's L1 to L2 across different linguistic levels (from the paralinguistic to lexical level). The findings are especially noteworthy as they were found in highly frequent words in Japanese and German that the L2 speakers should have encountered very often, suggesting that a rich amount of the input of the target language did not contribute to the formation of an appropriate L2 prosody. In chapter 1, I discussed the application of exemplar theories (Goldinger, 1996; Hintzman, 1986; Johnson, 1997) and statistical learning to explain the L2 learning. Given that L2 production and perception relate to each other, the result of this study does not seem to support the assumption of the pure exemplar theories that the input of L2 sounds alone is sufficient to reorganise perceptual space and to build novel L2 phonological contrasts. However, one might offer a counterargument that the amount of input that the L2 learners encountered was still not sufficient to reorganise their mental representations and to form new L2 representations.

Finally, the finding that the L2 speakers produced deviant F_0 contours regardless of the existence or absence of a lexical pitch fall and regardless of the number of attempt leads to the assumption that the L2 speakers might have either not stored the L2 prosodic information into the lexicon or they might have failed to articulate the utterance appro-

priately, irrespectively of what they had stored. In the former case, they might have not perceived prosodic information or/and have perceived it, but they have failed to store it. In the following experiments, the sources for these possible difficulties are investigated in more detail.

CHAPTER 3

DISCRIMINATION OF NONNATIVE SEGMENTAL LENGTH CONTRASTS

3.1 Introduction

Experiment 1 showed that L2 speakers' productions deviated from those of L1 speakers despite the high frequency of the uttered words ("Excuse me" and "Hello"). Since "[m]any L2 production errors have a perceptual basis" (Flege et al., 1995a, 238), the deviant productions could relate to incorrectly perceived sounds that were stored in long term memory not in an appropriate way. In order to determine the causes for the deviant forms in L2 productions, I will trace the speech process back to its basis and aim at examining one of the most basic perception abilities, the ability to discriminate nonnative prosodic contrasts. In this chapter[1], the discrimination ability of nonnative segmental length contrasts, in particular of nonnative consonant length contrast is investigated.

While a consonant length contrast is lexically used in Japanese, German does not employ this cue for a lexical distinction. German consonants containing a syllable boundary and a potential word boundary are produced longer only in sandhi, but they do not contribute to a lexical distinction.

Following the CAH (see e.g. Lado, 1957; Stockwell et al., 1965, and discussion in Chapter 1), nonnative consonant length contrasts are expected to be difficult to acquire for German L1 speakers, because the absence of an L2 contrast in one's L1 may cause difficulties in acquiring the L2 prosodic contrast. Additionally, as length in itself is different from laryngeal or supralaryngeal articulations, the acquisition of nonnative length is not

[1] A preliminary and shorter version of this chapter was published as a conference paper: Asano, Y. (2014). Stability in perceiving non-native segmental length contrasts. In *Proceedings of the 7th International Conference on Speech Prosody*. Dublin, Ireland, 321–325.

a simple matter of acquiring nonnative voicing, place of articulation or manner of articulation, which may make the acquisition of nonnative length contrasts more difficult. Following the PAM (Best, 1995; Best et al., 2001) and the SLM (Flege, 1995; Flege et al., 1995a), the absence of an L2 phonological contrast in listeners' L1 representations can either facilitate or impair the nonnative listeners' discrimination of a nonnative prosodic contrast. The facilitation and impairment depend on the perceived similarities of two sounds of the contrast. It may be that one L2 sound (= a short consonant) assimilated to an L1 category, the other (= a long consonant) falls uncategorised outside L1 categories (*Uncategorised versus Categorised*) (Best, 1995; Best et al., 2001; Best and Tyler, 2007). In this case, the discrimination performance is expected to be very good. What is more, this prediction presupposes a sufficient *perceived* distance between the perceived durations of short and long consonants. In other words, a long consonant must be perceived long enough to be different from a short consonant. In case an L2 listener perceives a long consonant not long enough to keep this distance and perceives it as a short consonant, *Category-Goodness Difference* predicts that the both L2 sounds are assimilated to the same L1 category, but they differ in terms of the distance from the L1 "ideal" (e.g. one is acceptable – singleton, the other –geminate is deviant). Also in this case, the discrimination performance is expected to vary from moderate to very good. In case an L2 listener perceives the two sounds equally far from the L1 "ideal" and they are therefore perceptually equally acceptable or equally deviant, *Single-Category Assimilation* predicts that the discrimination performance is expected to be poor.

In a similar way, the SLM predicts that the discrimination is facilitated, if each of two sounds of a pair is highly similar to a different L1 category or if only one of them is highly similar to an L1 category, but the other lies uncategorised outside the category. If two nonnative sounds are both assessed as highly similar to a single L1 category, their discrimination will be impaired (Flege, 1995; Flege et al., 1995a). In order to build such predictions based on the PAM or the SLM, it is necessary to be able to define perceived distance a priori. In other words, it is necessary to know how nonnative listeners perceive two sounds of the contrast.

Previous studies on the discrimination of nonnative segmental length contrasts show that nonnative listeners' discrimination performance varies from "poorer than L1 listeners" to "as good as L1 listeners". Unsurprisingly, L2 listeners' discrimination or identification abilities of nonnative segmental length contrasts have been reported to be lower than L1 listeners' ones (Altmann et al., 2012; McAllister et al., 2002). Altmann et al. (2012) conducted a discrimination task for nonnative vowel and consonant length con-

trasts testing Italian and German L1 listeners. The duration of ISI was 1600 ms. The results showed that the nonnative consonant length contrasts were perceived poorly compared to the vowel length contrasts by the German participants. The Italian participants showed higher sensitivity for consonant length contrasts than the German learners of Italian, followed by the German naïve listeners. McAllister et al. (2002) conducted an identification task on nonnative Swedish vowel and consonant length contrasts testing Estonian, English and Spanish L2 learners and Swedish L1 listeners as a control group. Their results showed that the Estonian L2 learners whose L1 employs consonant and vowel length contrasts performed as well as Swedish L1 listeners, followed by English and then by Spanish L2 learners, showing that subjects' success in learning the Swedish consonant and vowel length contrasts mirrors the lexical role of segmental durations in the L1 of an L2 listener.

At the same time however, some studies reported that even naïve (= inexperienced nonnative listeners) listeners without any exposure to an L2 could discriminate nonnative segmental length contrasts. Muroi (1995) and Hayes-Harb and Masuda (2008) tested the discrimination ability of nonnative Japanese consonant length contrasts by English L1 listeners. Muroi (1995) conducted an AX and an ABX discrimination task. In an AX discrimination task, two identical or different stimuli are presented within each trial, inviting "same" or "different" responses (details of this methods see section 3.2). In an ABX task, called also matching-to-sample task (Gerrits, 2001), listeners have to indicate whether the third stimulus (X) is identical to the stimulus in the first or second interval. She examined the discrimination ability of one consonant length and one vowel length contrast pair with 31 English learners of Japanese and reported that the participants were able to discriminate both consonant and vowel length contrasts well above chance, although the consonant length contrast was more difficult on average (consonant = 86% correct; vowel = 99%). Hayes-Harb and Masuda (2008) conducted a discrimination task (as a follow-up experiment after their main auditory word-picture matching task) with two monolingual English listeners. They were asked to determine whether pairs of the test stimuli that contrasted in consonant length were "same" or "different". They discriminated the minimal pairs with 93% accuracy. Note that both studies did not provide detailed information about procedures (e.g. the duration of ISIs) and about statistical analyses, but provided an indication that even naïve listeners can discriminate nonnative consonant length contrasts. In chapter 1, I argued that a reasonable explanation for a good discrimination ability by naïve listeners could relate to the the perceptual reliance on phonetic memory (Pisoni, 1973).

There are further investigations that strongly support this argument claiming that L2 learners rely on rate-independent durations of critical consonants when perceiving nonnative consonant length contrasts, while L1 listeners perceive geminate relatively with respect to word and sentence durations. Watabe and Hiratou (1985) showed that Japanese L1 listeners' perception of a geminate depends on the duration of the preceding vowel and it is not a matter of the durations of critical consonants. On the contrary, Uchida (1993) (testing Chinese L2 learners) and Toda (1998) and Toda (2003) (testing English L2 learners) revealed that the L2 perception of a geminate is not affected by the duration of the preceding vowel, even by advanced L2 learners. Furthermore, Hirata (1990) investigated the influence of the speech rate on the perception of geminates. She added "koto to iimashita (*Engl: it was said that*)" after the stimuli "ita (*Engl: to be in past indicative*)" – "it:a (*Engl: to go in past indicative*)" with different speech rate and found that only L1 listeners were affected by the speech rate of the following sentence, while L2 English L2 learners were not.

Further, Porretta and Tucker (2015) showed that naïve listeners can discriminate nonnative segmental length contrasts only by relying on phonetic differences of the contrasts. They tested English non-learners' ability to identify and discriminate nonnative Finnish consonant length contrast. Three groups (non-learners without instruction and with instruction, and L1 Finnish control) performed a speeded forced-choice identification task and a speeded AX discrimination task on Finnish non-words (e.g. *hupo – huppo*) which were manipulated for intervocalic consonant duration. The duration of ISIs was 500 ms. The identification task was expected to involve more phonological processing and the discrimination task more phonetic processing. The instructions consisted in explaining Finnish lexical consonant length contrasts that were given prior to the experiment. The results showed that non-learners could detect increasing differences between intervocalic consonants, but they did not perform like L1 Finnish speakers in both identification and discrimination tasks. The instruction enhanced the non-learners' discrimination performance to a much greater extent than their identification performance. This finding suggests that even naïve listeners can become more sensitive to phonetic differences of nonnative segmental length contrasts, but such attention to phonetic details could not influence their phonological processing of the contrasts.

Finally, when extending experimental methodology to forced-choice identification tasks to test the L1 and L2 listeners' categorical perception using the classical paradigm of Categorical Perception (henceforth CP, Repp, 1984), there are a handful of study on the perception of nonnative Japanese consonant length contrasts (Bin, 1993; Kin, 2005; Mi-

nagawa and Kiritani, 1997 testing Korean L2 learners, Enomoto, 1992; Hayes-Harb and Masuda, 2008; Hirata, 1990; Muroi, 1995; Toda, 1998; Toda, 2003 testing English L2 learners, Uchida, 1993; Nishihata, 1993 testing Chinese L2 learners, Masuko and Kiritani, 1992 testing Indonesian and Thai L2 learners and Minagawa, 1996 testing more than one L2 learners' group, namely Korean, Thai, Chinese, English and Spanish L2 learners). The studies vary with respect to the number of participants (e.g., 6 learners in Hirata, 1990 to 210 Korean, 36 Thai, 46 Chinese, 63 English and 122 Spanish learners in Minagawa, 1996, Minagawa, 1998) and to the L1 backgrounds of tested L2 learners, but they all reported that even L2 learners could detect short and long consonant differences, but did not show categorical thresholds as clear as those shown by Japanese L1 listeners. As for the L2 learning effect on the perception of geminates, the findings in these studies vary; on one hand, a positive L2 learning effect was found (Enomoto, 1992), on the other hand, some studies reported to have found no effect (Bin, 1993; Minagawa, 1998; Uchida, 1993).

The experimental methods and languages in focus differ across the aforementioned previous studies, but taken together, their findings all seem to suggest that even non-learners and learners can discriminate nonnative segmental length contrasts by perceptually relying on phonetic differences of the contrasts, but that their discrimination abilities decrease when they access to phonological representations. I will examine this issue by means of an experimental paradigm that manipulates the degree of phonological representations required to accomplish a task, that is, a discrimination task using two different durations of ISIs.

As discussed in chapter 1, the duration of an ISI is known to influence the degree of involvement of phonetic and phonological processing (Cowan and Morse, 1986; Gerrits, 2001; Johnson, 2004; Pisoni, 1973; Schouten and Van Hessen, 1992; Werker and Logan, 1985). In the long ISI condition, a listener has to try to keep phonetic information of the first stimulus in short-term memory by means of rehearsal, because otherwise it will decay and it will be difficult to compare the first stimulus with the second one. This process requires additional memory load. This decay of phonetic information in the course of time is also illustrated in working memory (a.o. Baddeley and Hitch, 1974). The longer an ISI, the higher the memory load required for the task and, as a consequence, the higher also the cognitive load placed on working memory. Based on these claims, the short ISI condition in the current experiment aimed at testing the ability in discriminating acoustic correlates of nonnative segmental length contrasts (corresponding to the "input" stage presented in Chapter 1) and the long ISI condition the ability in discriminating the contrasts requiring more phonological processing (corresponding to the

path from "input" to "mental representations"). I postulate that the phonological contrasts for which nonnative listeners do not have their L1 representations may decay much more easily in comparison to the contrasts for which they have their L1 mental representations. Applying this postulation to the current study, the German listeners may have greater difficulties in discriminating consonant length contrasts in the long ISI condition in comparison to vowel length contrasts.

In this study, the duration of ISIs in one condition was 300 ms and 2500 ms in the other for the following reasons: In order to eliminate the risk of backward masking, an ISI should be longer than 250 ms (Cowan and Morse, 1986; Imada et al., 1993; Sussman, 2005). Around 250 ms after the offset of a sound, information is recognised at the sensory level, but is not yet identified or categorised (Crowder and Morton, 1969). The discrimination ability increases rapidly between 100 and 500 ms and falls gradually as the ISI increases further (Cowan and Morse, 1986; Pisoni, 1973; Schouten and Van Hessen, 1992). The decrease after 500 ms may be interpreted as the effect of gradually decaying phonetic information in short-term memory. In order to test the phonetic comparison of successive stimuli without risking a backwards masking effect, I decided to use 300 ms in one condition (slightly longer than 250 ms) to ensure that the acoustic trace was available. Then, it is claimed that this uncategorised acoustic information is maintained for a period of time (approximately 2000 ms) in the short-term phonological storage as an auditory echo (Baddeley, 2000, 1986). To keep the experiment to a reasonable duration that would not cause the participants to diminish their concentration or motivation, but at the same time to make sure that the processing taps into the more phonological level after 2000 ms (Baddeley, 2000, 1986), I decided to use 2500 ms for the duration of the long ISI. Note that I do not share the view that the short ISI condition strictly tests only phonetic, language-general processing, while the long ISI condition exclusively phonological or language-specific processing, because the short ISI does not necessarily prevent L1 (and L2 listeners) from accessing their stored phonological information to aid their discrimination (Wayland and Guion, 2004). It appears appropriate to claim that the short ISI condition does not necessarily require participants to access mental representations due to comparatively weaker demands placed on working memory it activates mental representations to a lesser extent than the long ISI condition.

No study has been carried out so far that tested the discrimination of L2 segmental *length* contrasts using different ISI conditions within one study, although some studies investigated the discrimination of nonnative segmental length contrasts by not varying the durations of ISIs. For example, Altmann et al. (2012) used the duration of 1600 ms

for their AX discrimination task and Porretta and Tucker (2015) the duration of 500 ms to test nonnative consonant length contrasts. Further, no specification of the durations of ISIs is provided in the study by Hayes-Harb and Masuda (2008) and Muroi (1995). (Note that Hayes-Harb and Masuda conducted this experiment as an informal follow-up study without any methodological description in the paper). Therefore, I believe it valuable to investigate the L2 listeners' discrimination ability of nonnative segmental length contrasts by systematically varying the durations of ISIs.

Noticeably, the use of two different durations of ISIs can be embedded into a part of the investigation on the effect of *critical task variables* (task demands) on L2 perception, which have been seldom studied in the field of research (see discussion in Subsection 1.6.2). Besides memory load as the investigation on *critical task variables* (task demands), the current study examines an effect of another variable that enhances cognitive load; attention control by adding a task-irrelevant prosodic dimension to the stimuli.

Under challenging situations (= situations with distracting information such as noise in a background or task-irrelevant information), speech processing becomes more demanding due to the necessity of increased demand on attention control (Dallett, 1964; Luce et al., 1983; Rabbitt, 1966). This appears to be all the more true when it comes to L2 speech processing. Previous studies show that the difficulty of listening to speech in noise or with greater talker variability is more exacerbated in L2 perception when compared to L1 perception (Antoniou et al., 2013; Cutler et al., 2007; Lecumberri and Cooke, 2006; Nabelek and Donahue, 1984). Given that there are less cognitive resources in L2 speech processing due to reduced L2 proficiency in comparison to L1 listeners (Antoniou et al., 2013), L2 perception is expected to be more affected by distracting information in demanding listening conditions than L1 perception.

In this study, attention control was examined by assessing whether listeners could ignore a task-irrelevant distracting pitch movement on the consonant length contrast in order to efficiently accomplish the task (to discriminate consonant length contrasts). Two stimulus conditions were built, one with a pitch fall that occurred simultaneously with the durationally contrasted critical consonants and the other with a flat (monotonous) pitch. The demand on attention control in the former condition was expected to be higher, because it required participants to ignore the task-irrelevant pitch movement. In the flat pitch condition, listeners had one prosodic cue less to process (only duration). In case of a limited attention control, the trials with a falling pitch were expected to be more difficult to discriminate than those with a flat pitch. I hypothesise that the increased demand on attention control becomes an impediment to a greater extent for L2 listeners

than for L1 listeners, because they are less likely to ignore the task-irrelevant pitch movement. It is to note that the task-irrelevant pitch fall phonetically represented the phonetic form of a Japanese lexical pitch accent. Therefore, the study can test whether the difficulty in perceiving nonnative consonantal length contrasts in Japanese may be exacerbated when it simultaneously occurs with the Japanese pitch accent. Strikingly, Japanese geminate are normally accompanied with a pitch movement (Kubozono, 2011a,b).

By manipulating the task demands (memory load and attention control), the study enables us to examine how unstable L2 perception becomes with increased task demands in comparison to L1 perception, and how task demands can affect results, and possibly to a different extent between L1 and L2 listeners. In case that task demands affect L2 perception, results would help to develop a more differentiated view of L2 perception. For example, instead of claiming that "German L2 learners can or cannot discriminate nonnative consonant length contrasts", one can formulate a result that "they can discriminate them under the condition of...." and such a formulation of results would help to compare findings across different studies. Moreover, the vulnerability of L2 processing also explains why discrimination exercises with minimal pairs contrasting in segmental length in a class room situation (with less distracting factors, thus with less task demands) appears to be easier for learners than length perception in real life situations (with more distracting factors, consequently with more increasing demands). The investigation of the influence of experimental task demands on speech processing is therefore useful for the understanding of the factors influencing everyday speech processing.

Not least, it should be emphasised that an AX discrimination task such as the one discussed here is a suitable task to investigate the effect of task demands on nonnative speech perception, because it requires the lowest task demands under certain conditions compared to other types of tasks with more stimulus such as ABX discrimination tasks (e.g. Wood, 1976). As a consequence, it is easy to vary or enhance task demands in other conditions. Such a task that provides the lowest task demands is also useful when testing non-learners and may aid to avoid a floor effect. It also needs to be mentioned that the AX discrimination task with the long duration of ISIs was preferred over an ABX task that requires both stimuli (A and B) necessarily being stored in memory and therefore could also test the phonological processing done by the AX discrimination task with the long ISI duration. However, ABX tasks are claimed to make the A stimulus more abstract than the B one due to the assumed decay in detail of these phonetic memories (McGuire, 2010), which leads to a bias towards the B token as it is more current in memory. Indeed, previous studies have shown that listeners reduce uncertainty in an ABX task by compar-

ing only B and X (Pastore, 1987) suggesting that three stimuli can be actually reduced to two. This bias can be accommodated by a strict balancing of AB ordering, but this also means that this design does not have a speeded analogue (except for possibly very short stimuli)(McGuire, 2010).

3.2 Experiment

The experiment tested the stability of L1 and L2 listeners' ability to discriminate consonant length contrasts that do not lexically exist in the L2 listeners' L1. To this aim I conducted a speeded AX discrimination tasks and analysed 1) whether the L1 listeners' (= Japanese native listeners') discrimination ability was equally high in all conditions, and 2) under which task conditions L2 listeners (German learners and non-learners) differed from the L1 listeners, and 3) under which conditions the learners differed from the non-learners. In order to provide a holistic picture, the discrimination of both vowel and consonant length contrasts was tested in this study. Since a vowel length contrast exists at the same linguistic level in Japanese and German, the discrimination performance of a vowel length contrast served as a baseline. This procedure allowed us to compare the cross-linguistic ability to perceive nonnative segmental length contrasts compared to native segmental length contrasts.

3.2.1 Methods

Participants

Three groups of participants attended the experiment for a small fee; twenty-four Japanese L1 speakers (10 male, 14 female, 20-31 years old with an average age of 22.1), 24 German native speakers with no training in Japanese (= non-learners) (8 male, 16 female, 19-30 years old with an average age of 22.8) and 48 German learners of Japanese (30 male, 18 female, 20-34 years old with an average age of 25). All of them learned English as second language at school and none of them were bilingual. None of them had have learned another L2 with lexical consonant length contrasts or with lexical pitch/tone contrasts. All Japanese L1 speakers spoke Standard Tokyo Japanese and all German learners and non-learners claimed to speak mainly Standard German in their daily life. None of the German participants was from Baden-Württemberg because the High Alemannic dialects partly spoken in Baden-Württemberg are known to exhibit lexical consonant length contrasts (Kraehenmann, 2003; Seiler, 2009; Willi, 1996). Most of them were from

North Rhine-Westphalia. Participants were all unaware of the purpose of the experiment. They had no history of speech or hearing disorders and no prior training in intonational phonology. None of the participants were raised in a bilingual environment. Detailed information regarding the participants' demographics is provided in Appendix B. Speaking ahead of results, participants' L2 proficiency that was used as an independent variable in the statistical analysis did not predict their performance, so that the report on the analysis on L2 proficiency will not be presented throughout Experiments 2 to 4.

Materials

First, 21 disyllabic CVCV pseudoword triplets were developed which differed segmentally only in the length of the first vowel or in the length of the second consonant (e.g., /pɯnɯ/,/pɯːnɯ/, /pɯnːɯ/). Pseudowords were used to eliminate lexical effects that would disadvantage only the nonnative participant groups. The triplets differed in the initial and medial consonants (/p, b, g, t, k, z, s, n/) and vowels (/a, u/). Since the current investigation should not be restricted to one specific type of consonants, but rather tests consonant length perception as a more general phenomenon, the triplets that differed in manner of articulation and voicing of consonants were constructed.

The 21 triplets were evaluated in a pretest with 24 Japanese and 24 German L1 speakers (different from those of the main experiment) to select only stimuli that showed lowest association strengths in both Japanese and German groups. The aim of the pretest was to ensure that stimuli did not associate words in both the Japanese and German speakers' lexicon or if they did, to the same extent for both groups. The pretest was conducted online. Participants were presented with one stimulus at a time and wrote down the first word that came to their mind. The association strength was defined as the highest occurrence frequency of the same word (including grammatical and semantic derivatives). For example, the association strengths of the stimulus *baana* was 54.2%, because the most frequent associated word (Banane, meaning *Banana*) occurred 13 times and this number was divided by the total number of response (= 24).

The association strengths were calculated separately for the Japanese and the German participant groups. From the tested 21 triplets, 6 triplets with the lowest association strength in both groups were selected (the word association rates of the selected six triplets ranged between 29.8% and 45.3%, mean = 34.5%, while those of all 21 triplets ranged between 29.8% and 100.0%, mean = 52.3%). The selected triplets included two nasals (/n/), two plosives (/p/, /b/) and two fricatives (/s/, /f/). Half of these consonants were phonologically voiced, the other half voiceless. They were all combined with the

vowel /u/ (because all triplets with the vowel /a/ showed higher association strength), see all stimuli in Table 3.1.

reference pseudoword	singleton	long-vowel	geminate
punu	/pɯnɯ/	/pɯːnɯ/	/pɯnːɯ/
gunu	/gunɯ/	/guːnɯ/	/gunːɯ/
gupu	/gupu/	/guːpu/	/gupːu/
gubu	/gubu/	/guːbu/	/gubːu/
zusu	/zusɯ/	/zuːsɯ/	/zusːɯ/
sufu	/sɯfɯ/	/sɯːfɯ/	/sɯfːɯ/

Table 3.1 *Stimuli in overview*

The materials were recorded by a female native speaker of Tokyo Japanese in two pitch conditions; high flat pitch and falling pitch (with a pitch fall during the medial consonant pitch tracks see Figure 3.1). Each stimulus was read in isolation and recorded six times in order to be able to use different tokens of the same type (N = 216). The single speaker recorded all the stimuli to avoid the issue that listeners' judgements could be influenced by voice characteristics of different speakers. The recording took place in a sound-attenuated studio at the University of Konstanz. Data were directly digitised using a DAT recorder with a sampling rate of 44.1 kHz and a resolution of 16 Bit.

Since the experimental stimulus pairs had to be discriminated based on segmental length contrasts, F_0 movement of the two stimuli presented as a pair (= A and X) had to be the same. To this aim, F_0 of the stimuli was manipulated using a method which is based on the representation of F_0 contours with B-splines (De Boor, 2001) and on a smooth warping of the time axis that allowed us to put F_0 points to desired positions with respect to selected time boundaries (see Gubian et al., 2013, and Appendix A). While the software tools for the manipulation of speech like PSOLA (Pitch Synchronous Overlap Add Method) re-synthesis tool available in Praat (Boersma and Weenink, 2011) manipulates the perceptual effect of those features in isolation, e.g. by keeping the original F_0 contour and varying segment duration or vice versa, the manipulation method applied in this study varies these parameters in combination and quickly generates large numbers of stimuli (Gubian et al., 2013). This method is advantageous, because changing prosodic parameters in isolation can give rise to stimuli that sound unnatural if the covariation of the parameters is violated. More specifically, the 18 tokens of each triplet of the same reference pseudowords (e.g. *punu*, *pu:nu* and *pun:u*, each of them contained 6 tokens, thus in total 18 for each triplet) realised in the same pitch pattern (flat or falling pitch) were aligned on the average pitch across tokens (in the flat pitch: average = 1.3 semitones, range = 1.0 – 1.6 semitones; in the falling pitch: average = 13.0 semitones,

range = 10.5 – 16.4 semitones), see Table 3.2. The table shows durations and contrast ratios for each *different* pair. The ratios show that that the acoustic criteria for the length distinction in vowels and consonants were mostly met. The ratio for Japanese vowels was approximately 2.4 for a flat pitch, 3.2 for a falling pitch, Akaba, 2008 and for consonants 2.0 to 2.99 for disyllabic words Han, 1994; Homma, 1981).

reference pseudoword	**mean pitch range (in semitones)**	
	flat	**falling**
gubu	2.2	13.0
gunu	1.0	10.5
gupu	1.3	16.4
punu	1.4	13.9
sufu	1.6	12.7
zusu	3.3	13.5

Table 3.2 *Average pitch values in semitones for each reference pseudoword, separately for the flat and falling pitch condition.*

Finally, a female native speaker of Japanese and a male native speaker of German selected the most naturally sounding tokens as experimental items. The number of the selected items was 3 for the long-vowel and the geminate condition, and 4 for the singleton condition due to all possible combinations of the experimental item pairs (singleton vs. singleton, long-vowel vs. long-vowel, geminate vs. geminate for *same* pairs, singleton vs. long-vowel, singleton vs. geminate for *different* pairs). There was no disagreement on the decisions. Further, to compare the spectral quality for long and short vowels /uː/ and /u/, the first and second formants at the midpoint of the vowel were automatically extracted, see Table 3.4.

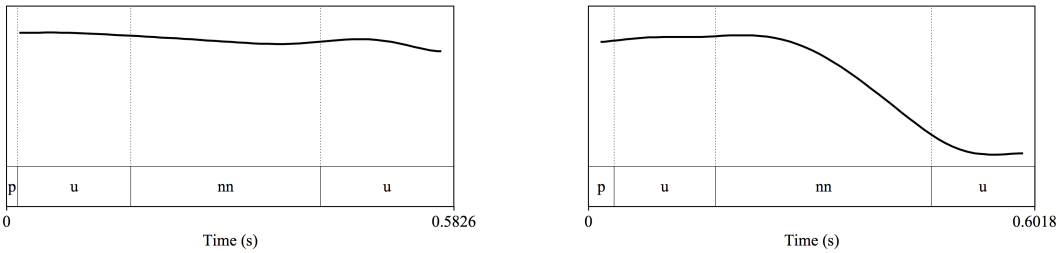

Figure 3.1 *Smoothed pitch track of geminate stimuli in the flat and falling pitch condition. F_0 range is shown between 100 and 350 Hz.*

reference pseudoword	length contrast	pitch condition	short segment	long segment	duration of short segment (in ms)	duration of long segment (in ms)	durational ratio
/punu/	v	flat	/u/	/u:/	99.4	429.5	2.3
/gunu/	v	flat	/u/	/u:/	123.5	377.4	3.3
/gupu/	v	flat	/u/	/u:/	102.7	351.3	2.9
/gubu/	v	flat	/u/	/u:/	130.2	372.0	3.5
/sufu/	v	flat	/u/	/u:/	115.4	376.2	3.1
/zusu/	v	flat	/u/	/u:/	108.9	357.8	3
/punu/	v	falling	/u/	/u:/	118.5	398.4	3
/gunu/	v	falling	/u/	/u:/	125.0	376.3	3.3
/gupu/	v	falling	/u/	/u:/	101.0	334.7	3
/gubu/	v	falling	/u/	/u:/	130.3	376.2	3.5
/sufu/	v	falling	/u/	/u:/	130.3	358.9	3.6
/zusu/	v	falling	/u/	/u:/	115.3	335.6	3.4
/punu/	c	flat	/n/	/n:/	66.2	248.8	2.7
/gunu/	c	flat	/n/	/n:/	72.5	258.7	2.8
/gupu/	c	flat	/p/	/p:/	148.5	382.1	3.9
/gubu/	c	flat	/b/	/b:/	79.0	344.9	2.3
/sufu/	c	flat	/f/	/f:/	101.8	368.5	2.8
/zusu/	c	flat	/s/	/s:/	132.8	330.6	4
/punu/	c	falling	/n/	/n:/	72.5	280.6	2.6
/gunu/	c	falling	/n/	/n:/	79.9	315.3	2.5
/gupu/	c	falling	/p/	/p:/	163.8	388.9	4.2
/gubu/	c	falling	/b/	/b:/	100.4	357.8	2.8
/sufu/	c	falling	/f/	/f:/	117.1	386.6	3
/zusu/	c	falling	/s/	/s:/	126.5	341.3	3.7

Table 3.3 *Critical segment durations and ratios for each reference pseudoword separately for the flat and falling pitch condition. "v" in the length condition stands for the vowel length contrast and "c" for the consonant length contrast.*

Procedure

A speeded AX-task was used to test the subjects' sensitivity for the consonant and vowel length contrasts. Half of the trials (N = 48) were trials without a contrast, the other half with a contrast. For the pairs with a contrast, there were eight possible combinations (e.g. *punu-pun:u, pun:u-punu, punu-pu:nu, pu:nu-punu*, each in the flat and falling pitch condition). One base list was assembled by presenting all possible pairings of the stimuli (N = 96). The experiment was programmed in *Presentation* (Neurobehavioral Systems) and the order of presentation was automatically randomised for each participant using the program.

Auditory stimuli were presented via headphones (Sony MDR-CD570). Each trial began with a sinusoid beep of 44100 Hz (= 500 ms) followed by 500 ms of silence. After this start signal, the first auditory stimulus was presented simultaneously with a visual presentation of the letter "A" on the screen (font size = 40, and for all letters presented in the experiment). After the offset of the first stimulus, a silence of 300 ms in the short ISI condition and of 2500 ms in the long ISI condition was inserted. After the silence, the second

vowel condition	pitch condition	F1	F2
short	flat	541 (101)	1570 (187)
short	falling	454 (68)	1466 (162)
long	flat	511 (37)	1511 (163)
long	falling	474 (93)	1512 (178)

Table 3.4 *Average F1 and F2 values for the short and long vowel /u/ and standard deviations in parentheses, separately for the flat and falling pitch conditions.*

auditory stimulus was presented. As in the first stimulus, a letter "B" appeared on the screen simultaneously with the onset of the stimulus. Immediately after the offset of the second stimulus, the visual information "ANSWER" (presented in English) signalled the time for answering. Participants were then given a maximum of 2500 ms before timeout. The intertrial-interval after the timeout was 1000 ms. No feedback was provided during the experiment. First, participants took part in the experiment with the short ISI condition (=first experimental session). After a pause that participants could take for how long they needed (ranging between 0 to about 3 minutes), they participated in the experiment with the long ISI condition (= second experimental session).

Before starting, participants were given a short description of the experiment and the procedure on a piece of paper which was written in their L1. It was described that they would hear two words either one after the other or with a long pause in between and their task was to decide whether the two stimuli were same or different in terms of their durations. It was further indicated that a training session would precede the experimental session and that the whole experiment would last 10 to 15 minutes. The aim of the study was not communicated to the participants, but it was indicated that they should pay attention to durational differences. This instruction was given as they also participated in Experiment 3 (in Chapter 4) in which they discriminated pitch contrasts. After reading the description, they sat in front of the computer that displayed "Welcome! The experiment will start soon. Press mouse button to start.". Participants then clicked a mouse button to start the training session.

Each experimental session began with the training session with 10 training trials using the pseudowords that were not used as experimental ones (*guna* and *puna*). The training started with the visual presentation of "Training session. Please pay attention to DURATION and answer soon after ANSWER is shown. Press a button on the box to start.". After the training, there was a pause (1 minute) before the experimental part started. The

experimental part started with the visual presentation of "Experimental session". The rest of the visual instruction was the same as for the training session.

In the middle of the list in each session (after 48 trials), there was an optional pause that participants could take for how long they needed. Then, they pressed a button to continue. After the first session with the short ISI condition, there was a pause. Then the second session with the long ISIs followed presenting the same trials, but in a different randomised order from that in the first session. The two sessions lasted approximately 20 minutes. All answers and reaction times (henceforth RTs) were recorded using a button box. Participants used their dominant hand for a "same" response and their non-dominant hand for a "different" response. RTs were measured in relation to the offset of the second stimulus.

Japanese participants were recruited and tested at the Tokyo University of Foreign Studies, and German learners at the Ruhr-University of Bochum and at the University of Düsseldorf and German non-learners at the University of Konstanz. All written instructions were given in English in the same way to all three participants groups.

3.2.2 Results

In total, 18432 data points were recorded (96 participants x 192 trials). From these, 66 data points had to be discarded due to timeout (9 from the Japanese listeners' data, 40 from the learners' ones and 17 from the non-learners' ones). In the following subsection, I will first present the analyses of d' scores (Macmillan and Creelman, 2005), a measurement of sensitivity to the kinds of contrasts tested, and then the analyses of RTs to examine a task difficulty. The response accuracy is shown in Table C.1 in Appendix B.

Sensitivity to contrast: d' score analyses

Participants' sensitivity to the contrasts was calculated using d' scores (Macmillan and Creelman, 2005). They are based on signal-detection theory and are measured by combining information about the likelihood that a participant successfully detects differences ("hits"), with information about the likelihood that the subject incorrectly indicates differences that are not there ("false alarms"). The analysis of d' scores is important when conducting an AX task, because listeners are known to tend to favour one response (e.g. the response "same" if the acoustic differences between stimuli are rather small) over the other, which leads to an unwanted response bias (Gerrits, 2001, 42). To eliminate the bias analytically, a signal-detection analysis of performance is needed.

For this experiment, d' scores were calculated for each participant and for each consonant and vowel length contrast, flat and falling pitch, as well as for short and long ISI (8 d' scores for each participant, resulting in total N = 768). A hit rate of 1 was corrected by subtracting 1/2N (N = 12), and a false alarm rate of 0 was corrected by adding 1/2N following the method (Macmillan and Kaplan, 1985). The theoretically possible highest d' in this data set reached 5.43. Average d' scores in each pitch and ISI condition in the three participant groups are shown in Table C.2 in Appendix B.

In the following analyses, I will first report statistical results from linear mixed-effects regression (LMER) models and then additionally show descriptive mean values and 95% CI error bars in plots as a complementary data analysis, because the use of a statistical significance referring only p-values has been critically discussed recently due to the low reliability of obtained p-values (e.g. Cohen, 1994; Cumming, 2011, 2013; Loftus, 1993; Simmons et al., 2011). Cumming (2013) showed *dance of the p-values* (p. 6) to point out an enormous variation in the p-values from less than .001 to 0.75 when replicating 25 times an experiment with two independent groups, each group having an N of 32. CIs are informative because they indicate the possible variations of p-values shown in the dance of the p-values, while a single p-value gives virtually no information about the infinite sequence of p-values.

In order to analyse performance differences between the vowel and consonant length contrast, d' scores were normalised by subtracting those of the vowel length contrast from those of the consonant length contrast (normalised $d'_{\text{consonant-vowel}}$). Negative values mean that participants had difficulties in discriminating consonant length contrasts compared to vowel length contrasts.

A set of LMER models were built with *d' scores* as dependent measure and *language group* (Japanese vs. learners vs. non-learners), *pitch* (flat vs. falling), *ISI* (short vs. long) as fixed factors and *participant* as a random factor including random slopes for the fixed factors (Barr et al., 2013; Cunnings, 2012). In case the model did not converge, the random slope for items was removed. P-values were calculated using the Satterthwaite approximation in the R-package lmerTest[2]. In the following, statistical results are reported on the most parsimonious model obtained by eliminating factors that were not significant if this did not deteriorate the fit of the model, and proceeding with backward elimination based on log likelihood ratio tests. Model specifications are shown in Appendix C.

[2] `http://cran.r-project.org/web/packages/lmerTest/ lmerTest.pdf`

The analysis using LMER showed a significant three-way interaction between *language group*, *ISI* and *pitch* ($p < 0.001$). To investigate the nature of this interaction, the data were split by *pitch*. In the flat pitch condition, there was an interaction between *language group* and *ISI* (the d' scores by the non-learners decreased in the long ISI condition in comparison to those by the Japanese, ß= -0.70, SE = 0.20, t = -3.5, $p < 0.001$, and to those by the learners, ß= -0.72, SE = 0.17, t = -4.1, $p < 0.001$). In the falling pitch condition, there was a main effect of *language group* (the d' scores by the non-learners were lower than those by the Japanese, ß= -0.93, SE = 0.20, t = -4.6, $p < 0.001$, and by the learners, ß= -0.63, SE = 0.17, t = -3.6, $p < 0.001$, and the d' scores by the learners tended to be lower than those by the Japanese, ß= -0.30, SE = -1.7, $p = 0.09$). Further, there was an interaction between *language group* and *ISI* (the d' scores by the non-learners became higher in the long ISI condition in comparison to those by the Japanese, ß= 0.42, SE = 0.16, t = 2.4, $p < 0.03$, and to those by the learners, ß= 0.40, SE = 0.15, t = 2.6, $p < 0.03$).

Figure 3.2 shows mean normalised $d'_{\text{consonant-vowel}}$ scores and 95%CI bars for the flat pitch condition (left) and for the falling pitch condition (right) for each participant group and ISI condition. The comparison between the gray bars (for the short ISI condition) in the left plot shows that all participant groups performed equally well. Their d' scores did not differ from the value of 0, suggesting that the d' scores for the consonant and the vowel length contrast were the same. The comparison between the white bars (for the long ISI condition) in the left plot shows that the Japanese and the learners performed equally well for the vowel and the consonant length contrast, but the non-learners' discrimination ability for the consonant length contrast was lower than the one for the vowel length contrast. The normalised d' scores for the long ISI and the flat pitch condition differed between the three participant groups (Japanese > learners > non-learners). As for the right plot, the gray bars (of the short ISI condition) show that the Japanese performed equally well for the vowel and the consonant length contrast, but both nonnative participant groups performed better for vowel length contrast than for consonant length contrast. This held true also in the long ISI condition (white bars). Normalised d' scores in both ISI conditions differed between the three participant groups (Japanese > learners > non-learners).

In order to corroborate the interaction found in the LMER-analysis and to better analyse the plots visually by removing the within-subject variable *ISI*, the normalised $d'_{\text{consonant-vowel}}$ scores for the short ISI condition were subtracted from those for the long ISI condition (= normalised $d'_{\text{long-short ISI}}$ scores). Figure 3.3 shows mean normalised

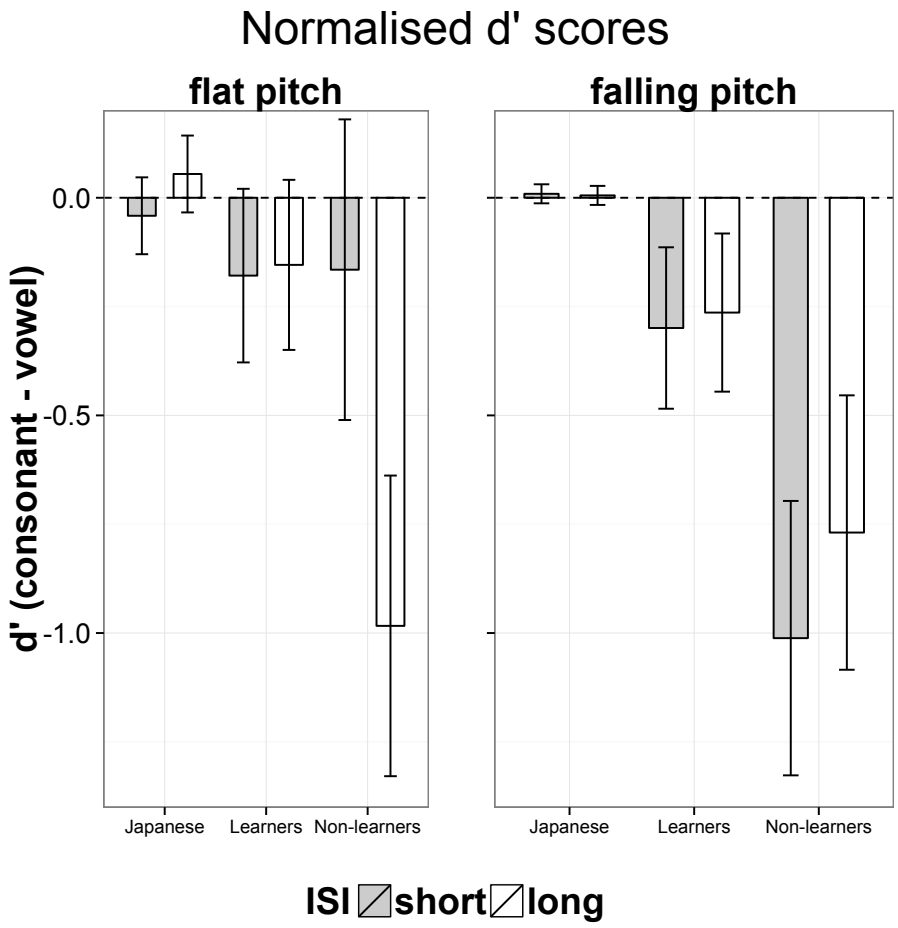

Figure 3.2 *Mean normalised d'$_{consonant\text{-}vowel}$ scores and 95%CI bars for the flat pitch condition (left) and for the falling pitch condition (right) for each participant group and ISI condition*

d'$_{\text{long-short ISI}}$ scores and 95%CI bars for the flat pitch condition (left) and for the falling pitch condition (right) for each participant group.

The left plot shows that only the CI bars of the non-learners are placed in the negative area, meaning that the non-learners' d'$_{\text{consonant-vowel}}$ scores in the long ISI condition were lower than those in the short ISI condition. The Japanese and the learners' d'$_{\text{consonant-vowel}}$ scores in the long and short ISI condition did not differ. The right plot confirms that the d'$_{\text{consonant-vowel}}$ scores in the long and short ISI condition did not differ in all three participant groups. Figure 3.2 and Figure 3.3 visually confirmed the interactions found in the LMER-analysis.

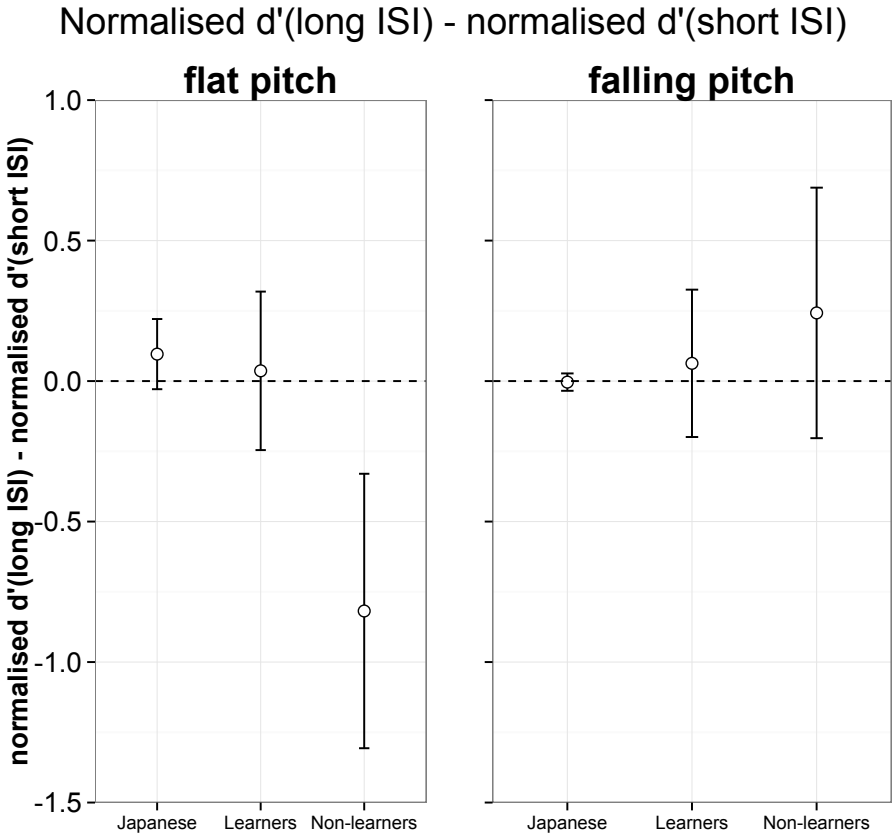

Figure 3.3 *Mean normalised d'$_{long\text{-}short\ ISI}$ scores and 95%CI bars for the flat pitch condition (left) and for the falling pitch condition (right) for each participant group*

Processing difficulty: RT analyses

The RTs were analysed to investigate the task difficulty of a decision (Pisoni and Tash, 1974 conducting discrimination tasks, Borràs-Comes et al., 2010; Chen, 2003; Jiang, 2012; Schneider et al., 2011 and Tomaschek et al., 2011 using the classical perceptual categorisation paradigm). The analysis of RTs can provide a useful insight of the data. In general, the longer the RT, the more difficult it was to make the decision. In my experiments (Experiments 2 and 3), only the RTs of trials with *different* pairs were analysed for the following reasons: First, Nosofsky (1992) and Shepard et al. (1975) reported that only RTs for "different" pairs became shorter as stimuli became more dissimilar, but this did not hold true for "same" pairs. Following their findings, only the longer responses to "different"

pairs can be taken as an indication that it was relatively difficult to hear the difference between the stimuli. Second, Pisoni and Tash (1974) showed that the nature of the required processing for *same* and *different* pairs and the level at which the comparisons were made seemed to be different. RT analyses in a discrimination task showed that RTs for *same* pairs with two different within-categorical stimuli (A-a) were longer than those for *same* pairs with two acoustically identical stimuli (A-A) and for *different* pairs (A-B) (Pisoni and Tash, 1974). Since the current study did not use two acoustically identical stimuli for *same* pairs, RTs for *same* pairs could be longer than for *different* pairs following the results of Pisoni and Tash (1974). Since the main interest of this study was to explore differences between the three participant groups for *different* pairs, these potentially longer RTs for *same* pairs could have made the RT analysis more difficult to interpret. Raw average RTs are shown in Table C.6 in Appendix B. The raw RTs show that the Japanese were generally faster than the learners and the non-learners. As the next step, in order to account for participant-specific RT-differences, the raw RT data were normalised in the following way: I discarded RTs longer than 2000 ms and aggregated the data for each participant, for each pitch, for each segmental length and for each ISI condition. Then, the averaged RTs for the vowel length contrast were subtracted from those for the consonant length contrast (as baseline) (= normalised $RT_{consonant\text{-}vowel}$).

The analysis using LMER showed a significant main effect of *language group* (the normalised $RTs_{consonant\text{-}vowel}$ by the learners and by the non-learners were longer than those by the Japanese, ß= 89, SE = 33, t = 2.7, p < 0.01, ß= 88, SE = 39, t = 2.3, p 0.03, respectively, but the learners' and the non-learners' RTs did not differ from each other, ß= -2, SE = 33, t = 0, p = 1.0) and a significant interaction between *language group* and *pitch* (the normalised $RTs_{consonant\text{-}vowel}$ by the learners became longer in the falling pitch condition in comparison to those by the Japanese, ß= 159, SE= 55 t = 2.9, p < 0.01).

Figure 3.4 shows mean normalised $RTs_{consonant\text{-}vowel}$ for each pitch condition, for each ISI condition and for each participant group. The value of 0 in the plot means that the RTs for the consonant and the vowel length contrast were almost the same. The right plot for the flat pitch condition shows that the Japanese listeners were faster in the consonant length contrast than learners both in the short and long ISI condition. The Non-learners responded as fast as the Japanese in the short ISI condition. However, their native-like performance did not last in the long ISI condition. In the long ISI condition, the Japanese listeners were faster than the non-learners and the learners. The learners were slower than the Japanese in both ISI conditions. Long CI bars show that the two nonnative participant groups contained a lot of variations and did not differ from each other in both ISI

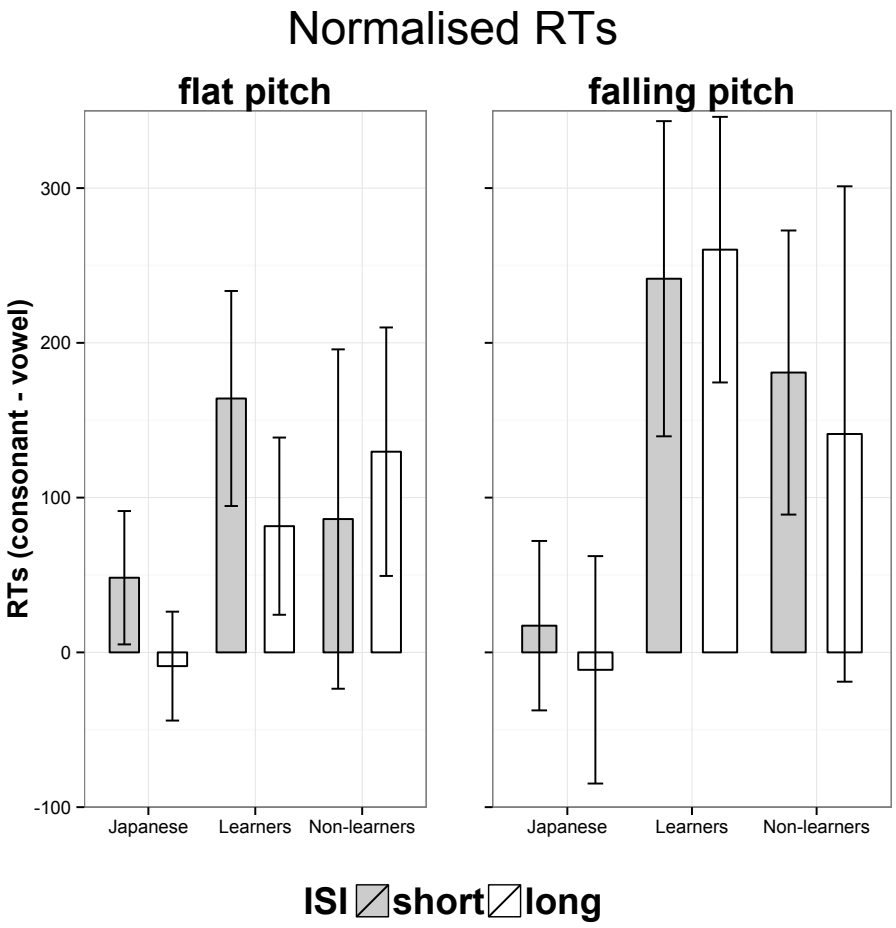

Figure 3.4 *Mean normalised* $RTs_{consonant\text{-}vowel}$ *with 95% CI bars*

conditions. The right plot for the falling pitch condition shows that the Japanese listeners were faster in the consonant length contrast than the learners and the non-learners. The mean $RTs_{consonant\text{-}vowel}$ for the learners seem to be longer than those for the non-learners, but the CI bars indicate that they did not differ from each other. The interaction found in the LMER-analysis was also approved in the plots.

Summary

The d' score analyses showed that the Japanese participants' discrimination sensitivity did not differ between the vowel and consonant contrast. However, their sensitivity was generally higher in the flat pitch condition in comparison to that in the falling pitch con-

dition. This was also true for the learners' and non-learners' d' scores. Their sensitivity was generally higher in the flat pitch condition than in the falling pitch condition. As for the effect of ISI, only the non-learners' sensitivity was affected by the ISI condition, and this was observed only in the flat pitch condition.

The raw RTs showed that the Japanese participants were overall much faster than the learners and the non-learners. The normalised RTs$_{consonant\text{-}vowel}$ showed that the Japanese participants were equally fast in discriminating vowel and consonant length contrasts. On the contrary, the learners and the non-learners were faster in discriminating the vowel length contrasts than the consonant length contrasts. Moreover, the learners generally took longer to respond in the falling pitch condition than in the flat pitch condition. Finally, the non-learners showed RTs as short as the Japanese only in the flat pitch and in the short ISI condition.

3.3 Discussion

The current experiment aimed at testing the ability in discriminating acoustic correlates of nonnative segmental length contrasts (corresponding to the "input" stage in Chapter 1) and in discriminating the contrasts requiring more phonological processing (corresponding to the path from "input" to "mental representations"). In order to differentiate these two stages, the durations of ISIs were varied. In the long ISI condition, more phonological processing was expected to be involved than in the short ISI condition. Moreover, the task-irrelevant falling pitch movement enhanced the demand on attention control compared to the flat pitch, because the pitch fall occurred simultaneously with the length contrast.

First of all, the raw d' scores showed that all three participant groups obtained higher d' scores in the flat pitch condition than in the falling pitch condition, and that the Japanese participants generally performed better than both groups of nonnative participants. The general effect of the task-irrelevant pitch on the discrimination of segmental length contrasts, regardless of the vowel or consonant length contrast, and of the participant groups, suggests that the effect of increased demand on attention control was more dominant than the effect of increased memory load.

In the flat pitch condition, normalised d' scores$_{consonant\text{-}vowel}$ did not differ in the short ISI condition across the three participant groups. The nonnative listeners' performance was equally good as the Japanese listeners' one. However, in the long ISI condition, the normalised d' scores$_{consonant\text{-}vowel}$ by the Japanese listeners were higher than those by the

learners followed by those by the non-learners. The further analysis revealed an effect of ISI condition only in the non-learners' normalised d' scores$_{consonant\text{-}vowel}$. The normalised d' scores$_{consonant\text{-}vowel}$ by the non-learners decreased in the long ISI condition. This decrease was caused by the low d' scores for the consonant length contrast in the long ISI condition. The native-like good performance shown by the learners and even by the non-learners was observed only in the condition with the lowest task demand (= in the short ISI and in the flat pitch condition). Such decreased performance indicates the nonnative listeners' decreased sensitivity in discriminating nonnative consonant length contrasts, when memory load increased and more phonological processing was required. In the falling pitch condition, the Japanese listeners showed higher normalised d' scores$_{consonant\text{-}vowel}$ than the learners followed by the non-learners in the short ISI condition and this did not change in the long ISI condition.

In the flat pitch condition, the non-learners were affected by the increased memory load to a greater extent than the learners. The learners obtained higher d' scores than the non-learners in the falling pitch condition. The learners' higher sensitivity scores for the consonant length contrast compared to the non-learners' ones can be regarded as evidence that the learners were forming new phonological categories (= short vs. long consonant length contrast) through the exposure to the L2. Flege's prediction that L2 learning reorganises L1 categories and forms new categories in the course of the learning process (Flege, 1995) does not appear to be restricted to segmental features, but was confirmed for segmental *length* contrasts. Also previous studies on the perception of nonnative segmental length contrasts reported the reorganisation of L1 categories and the formation of new L2 categories in prosodic features (e.g. Altmann et al., 2012; Hayes-Harb and Masuda, 2008).

As for the results from the RT analyses, the Japanese listeners' raw RTs were generally shorter than those by the learners and the non-learners. Also the normalised RTs$_{consonant\text{-}vowel}$ by the Japanese differed from those by the learners and the non-learners. Both learners and non-learners took longer in discriminating consonant length contrasts than vowel length contrasts. Additionally, the learners took much longer in discriminating consonant length contrasts than vowel length contrast in the falling pitch condition.

Additionally, the RT analyses revealed shorter raw RTs of the non-learners than the learners, although the two groups did not differ statistically, as indicated by strongly overlapping CI bars. Similar to the findings obtained in this study, Altmann et al. (2012) that investigated the discrimination ability of nonnative Italian consonant length contrasts

by German L2 learners reported higher d' scores by learners than by non-learners, but no significant difference between the learners' and non-learners' RTs. They explained that the learners were indecisive, because they had a degree of access to a newly forming category and thus needed to decide between two possible mapping representations (i.e. /C/ vs. /C:/). I interpret this finding as evidence that our learners and non-learners applied different strategies for the task. In the learners' phonological representations, two systems competed with each other (the L1 and L2 one) and the learners might have had faced the task of selecting one of them. For this selection, they would have had required more time. In the non-learners' phonological representations on the contrary, no language system with lexical consonant length contrasts existed (it was controlled that they had have not learned another language with lexical consonant length contrasts). Without the "selection" , shorter time was required for the decision (same-different). This explanation, however, presupposes that the competition between L1 and L2 systems already started in the short ISI condition because phonological representations were activated already in the short ISI condition. This finding could be a small piece of evidence that acoustic speech signals directly co-activate phonological representations (Darcy et al., 2012). Another possible interpretation for the tendency of longer RTs by the learners than by the non-learners is that the learners were more careful with giving responses and listening to Japanese-like stimuli in the experiments, perhaps because they felt being judged with respect to their performance in Japanese, while the non-learners felt less under performance pressure. The underlying mechanisms for the non-learners and the learners need further attention in future experiments.

The native-like good performance was found for the learners and the non-learners only in the flat pitch and short ISI condition, in which task demands were the lowest. The non-learners' normalised d' scores $_{consonant\text{-}vowel}$ were as high as the Japanese ones and their normalised RTs$_{consonant\text{-}vowel}$ were as short as the Japanese ones in this condition.

Moreover, the effect of ISI condition was found only in the flat pitch condition, in which the non-learners' performance decreased in the long ISI condition. In the falling pitch condition, the nonnative listeners' performance differed from that by the Japanese listeners already in the short ISI condition. In other words, the effect of ISI condition was found only in the flat pitch condition, namely when the processing only dealt with a length contrast without any differences in pitch. Once pitch came into play, the effect of ISI condition vanished. In other words, the effect of memory load surfaced only in a so-called "greenhouse condition" for the consonant length contrast, when the processing was not *disturbed* by another prosodic cue, pitch.

The results suggest that the effect of pitch condition (with increased psychophonetic complexity) had a stronger impact than the effect of the ISI condition, which can be discussed on the basis of the hierarchical relationships between the main components in working memory (e.g. Baddeley, 2010; Baddeley and Hitch, 1974) (see Figure 1.6 in Section 1.6.1). The demand on attention control is governed in *the central executive*, while the increased memory load governed in *the phonological loop*, which is subordinated under *the central executive*. Therefore, the more dominant effect of the pitch condition than of the ISI condition means a more dominant influence of *the central executive* than *the phonological loop* while processing speech stimuli and indicates that successful processing in the subordinated system is required to succeed in the native-like processing in *the central executive*, but not vice versa. Further research is needed to corroborate this non-reciprocal dependency between these systems in working memory.

Furthermore, the findings suggest that different processing mechanisms may underlie the three participant groups in discriminating the consonant length contrast already in the short ISI condition. When a difference between the acoustic correlates of the segmental length contrast was salient without the distracting pitch movement, nonnative listeners could rely on the acoustic correlates to discriminate the nonnative length contrast. However, once memory load increased in the long ISI condition or once the task-irrelevant pitch movement came into play, the comparison of the acoustic correlates became more daunting. In such situations, the nonnative listeners started to rely more on phonological processing, that might be already co-activated once acoustic input was perceived. As a consequence, the nonnative listeners who did not have appropriate phonological representations showed decreased performance, because they could not fully rely on the acoustic correlates. The advantage that the Japanese listeners showed already in the short ISI condition suggests that they could rely on both acoustic correlates and phonological representations, and that phonological representations aided the discrimination processing. As discussed in Chapter 1, it is not a minor view to assume that phonological processing is involved simultaneously with phonetic processing even while acoustic information is still "alive" (before it decays), and that the short ISI condition does not necessarily preclude listeners from accessing phonological information stored in long-term memory. Darcy et al. (2012) postulate a direct mapping from acoustic information to phonological representations and Wayland and Guion (2004) showed empirical findings to support this view, testing the discrimination ability of L2 Thai tone contrasts by Chinese and English L2 learners in comparison to Thai L1 listeners. The Chinese L2 learners outperformed English L2 learners in discriminating Thai tone contrasts

already in the 500-ms ISI condition. The advantage that the Chinese L2 learners showed suggests that they activated their L1 phonological representations of lexical tones already in the short ISI condition and their L1 phonological representations helped them to discriminate the L2 tone contrasts in Thai. Together with the previous studies, I argue that the result found in the current experiment suggests that the phonological representations were already activated in the short ISI condition, possibly to a lesser extent than in the long ISI condition.

The task demands controlled in the present study may be replaced with distracting factors in natural speech perception (e.g. L2 perception in noise or in different speech rate). Therefore, the performance decreases found in this study suggest that L2 speech perception can become unstable under various task demands in daily life. Moreover, the task-irrelevant pitch fall mirrors the phonetic form of a Japanese lexical pitch accent. The result therefore indicates enormous difficulties in perceiving consonantal length contrasts with pitch movement that takes place simultaneously. Strikingly, Japanese special morae (long vowel, syllabic nasal and geminate) are normally accompanied with a pitch movement (either with an initial low or a pitch accent) (Kubozono, 2011a,b). The finding in the current experiment provides an indication of how difficult it is for the nonnative listeners (whose L1 does not exhibit consonant length contrasts) to perceive short vs. long consonants of real Japanese words.

Finally, the findings are important as the vulnerability of L2 perception under increased task demands and under phonetic and phonological processing has been rarely taken into account in the discussion about the current L2 perception models. A notable exception was Best et al. (2001) that proposed in their recent version of the PAM to substantiate differences in the perception of auditory, phonetic and phonological information in nonnative speech, but so far their proposal has been restricted to segmental features. Moreover, the findings of this study show that perceived similarities for nonnative listeners cannot be simply defined by means of a cross-linguistic comparison of phonological categories. Perceived similarities of two nonnative sounds can depend on task demands under which two sounds are presented. This is because nonnative listeners do not always rely on stable phonological representations to discriminate two stimuli, but they often rely on the stimuli's phonetic differences. Task demands should be taken into account to define perceived similarities. It still seems difficult to predict cross-linguistic perceived similarities of prosodic contrasts a priori. In the current status of the research on L2 prosody, the perceived distance should be analysed a posteriori based on empirical findings.

Let us go back to the first and foremost research question whether nonnative listeners would have difficulties in discriminating acoustic correlates of nonnative segmental length contrasts (corresponding to the "input" stage) and in discriminating the contrasts requiring more phonological processing (corresponding to the path from "input" to "mental representations"). Even the nonnative listeners without any exposure to the L2 could discriminate acoustic correlates of nonnative segmental length contrasts under the condition with the lowest task demands in this experiment. Once the required memory load increased, the non-learners without exposure to the L2 decreased their sensitivity in the path from "input" to "mental representations". There, the learners with exposure to the L2 could keep their sensitivity to the consonant length contrast in both "input" and "mental representations" stages. However, once higher demands on attention control were required, the learners' and the non-learners' performance differed already in the short ISI condition from the native listeners' one, suggesting a nonnative listeners' disadvantage already in the "input" stage. Even the learners who were not distracted by the increasing memory load could not overcome the performance decrease due to the increased psychophonetic complexity of the stimuli. It was difficult for both learners and non-learners to ignore the task-irrelevant pitch and to focus their attention only on the task-relevant information. The finding indicates the difficulty in automatising L2 processing even after establishing or being exposed to the L2 categories. Tracing the speech process back to its basis, I found the vulnerability of the processing of L2 prosody.

In our natural listening situations, there are numerous distracting factors that might impair the L2 perception. The overall decreases indicate why the L2 perception still remains difficult in daily life situations, despite the successful exercise in L2 class room situations.

CHAPTER 4

DISCRIMINATION OF PITCH CONTRASTS

4.1 Introduction

Experiment 1 in Chapter 2 showed that the German learners of Japanese produced F_0 contours that the Japanese L1 speakers never produced in the experiment. Such deviant F_0 contours were found regardless of whether or not a lexical pitch accent fall was specified in the Japanese words. This finding leads to the assumption that the L2 speakers either did not store the pitch information into the lexicon or they had difficulties in their articulation. In the former case, it may well be that the L2 speakers either did not perceive pitch information, and had problems in the "input" stage (see Chapter 1), or they perceived it, but failed to store it and had problem in the path from "input" to "mental representations". In this chapter, these two possible causes in speech perception are investigated testing the same Japanese and the German learners and the non-learners as in Experiment 2. Analogue to Experiment 2, speeded AX discrimination tasks were conducted by varying task demands. Memory load was manipulated using the same two durations of ISIs as in Experiment 2. Attention control was increased by presenting stimulus pairs in task-irrelevant native and nonnative segmental length structures (singletons and long-vowels as native ones while geminates as nonnative one for the German participants). The investigation of the effect of the task-irrelevant segmental length structures on the discrimination of pitch contrasts is a contrary effect investigated in Experiment 2 (in which the effect of the task-irrelevant pitch movement on the discrimination of segmental length contrasts was investigated).

The investigation on the perception of F_0 is especially interesting as this prosodic cue is used in all languages, but carries different linguistic functions across languages (e.g. a lexical function in Chinese or Japanese, while a post-lexical or paralinguistic function in

German or English). In other words, listeners are familiar with F_0 contrasts, but in different ways depending on their L1. In Japanese, F_0 is used primarily lexically (Kubozono, 1999; Vance, 1987), but not in German. The question is whether German listeners are able to discriminate the acoustic correlates of pitch accent contrasts and to store them phonologically as well as Japanese L1 listeners when F_0 also has a linguistic function in their L1, albeit not at the lexical level, but at the post-lexical and paralinguistic one.

There are a handful of previous studies on the perception of Japanese pitch accents testing L2 listeners (Ayusawa et al., 1995; Ayusawa and Odaka, 1998; Cutler and Otake, 1999; Hayashi, 1996; Hirano-Cook, 2011; Honda, 2007; Nishinuma et al., 1996; Sakamoto, 2010; Toda, 2001). If anything, most of the studies on nonnative perception regarding Japanese pitch accents concerned the identification of the position of the accent in words (Ayusawa et al., 1995; Ayusawa and Odaka, 1998; Cutler and Otake, 1999; Hayashi, 1996; Hirano-Cook, 2011; Honda, 2007; Nishinuma et al., 1996; Toda, 2001). That is, there are only few studies investigating the discrimination and categorisation ability of Japanese pitch accent testing L1 and L2 listeners (e.g. Hirano-Cook, 2011; Hirata et al., 1997; Honda, 2007; Sakamoto, 2010).

Hirano-Cook (2011) conducted an AX-task with Japanese pitch accent contrasts using real Japanese words and non-words. She tested advanced English learners of Japanese and beginners, and reported that the accuracy rate shown by the advanced learners was statistically higher (90%) than that by the beginners (84%). The difference between the two groups still weakly supports the claim that nonnative prosodic contrasts were difficult to perceive. There are however points in the study that raise criticism against its experimental methodology. First, no comparison groups (Japanese native listeners and English naïve listeners) were tested. Therefore, it is not possible to compare the learners' performance with the Japanese native listeners' one. Second, no detailed information about the procedure of the AX task was provided e.g. the duration of an ISI. Third, the reported result was based on the response accuracy rate, but not on the d' scores that measure listeners' sensitivity to the contrasts (Macmillan and Creelman, 2005). Honda (2007) examined the role of pitch accent in the perception of spoken Japanese by L1 Japanese listeners with and without pitch accent in their variety in Japanese and and Norwegian and English learners of Japanese (with and without pitch accent in their L1 respectively). They were required to make a two-alternative forced choice response each time they heard a stimulus *hashi* (meaning "chopsticks" when it is realised with a high-low pitch); *hashi* ("bridge" with a low-high pitch) and *ame* ("rain" with a high-low pitch); *ame* ("sweets" with a low-high pitch). The pitch heights of the stimuli were manipulated.

The results revealed that the non-pitch accent background of the Japanese speakers impeded their perception of pitch accent in standard Japanese. Moreover, highly significant differences were observed between the L1 and L2 groups. Contrary to expectations, the Norwegian group demonstrated tendencies similar to those of the English group, meaning that the lexical pitch of the L2 learners' L1 did not facilitate the acquisition of L2 lexical pitch accent. Among the previous studies, the most relevant study for the current experiment is Sakamoto (2010) that investigated the perception of Japanese pitch accents testing English L2 learners of Japanese and Japanese L1 listeners. She conducted ABX tasks with an ISI of 700 ms and a three-alternative forced identification task. Her findings showed that the English learners' ability to discriminate F_0 contours in a nonspeech context was equal to that of the L1 listeners, but the learners' ability to identify them was poorer than the L1 listeners' one. Her finding strengthens the claim that nonnative listeners can discriminate nonnative prosodic contrasts based on phonetic memory, but they fail to do so when requiring their phonological knowledge. Summarising the findings in the previous studies, L2 listeners were reported to be able to discriminate nonnative pitch contrasts by exploiting acoustic correlates of the contrasts, but they become less successful in doing so once the phonetic information decays and more phonological information is required to accomplish the task. This claim is in line with the one made for the discrimination ability of nonnative segmental length contrasts by L2 listeners in Chapter 3.

Compared to the studies investigating the perception of pitch accent contrasts testing L1 Japanese and L2 listeners, the perception of nonnative pitch contrasts has been extensively studied in terms of discrimination or categorisation of tone contrasts testing L1 speakers of tone languages and of non-tone languages e.g. Mandarin Chinese or Thai L1 listeners vs. English L1 listeners. The findings from the studies on the perception of L2 tone contrasts are useful for my research question, because the difference between "tone" and "pitch accent" (e.g. Hyman, 2001, 2007) does not affect the hypotheses in this experiment. Therefore, I will review some studies on L2 tone perception. Unsurprisingly, behavioural studies generally report that L1 listeners of tone languages generally show advantages in discriminating tone contrasts or categorising tones than L1 listeners of non-tone languages in different types of tasks (e.g. Braun et al., 2014; Francis et al., 2008; Hallé et al., 2004; Krishnan et al., 2009; Pfordresher and Brown, 2009; Qin and Mok, 2013; So and Best, 2011; Wayland and Guion, 2004).

Further, there are two general kinds of accounts that predict the L2 performance in perceiving tonal contrasts. One account argues that a nonnative prosodic category can-

not be assimilated to a category in the listeners' L1 across different linguistic levels (= "levels of representations" account). It is crucial that a prosodic cue exhibits a linguistically meaningful contrast at the same linguistic level in one's L1 and L2. This account is supported by Wayland and Guion (2004). The other account claims that a nonnative prosodic category can be assimilated to a category of listeners' L1 prosodic system across different linguistic levels ("category assimilation" account). This account is an extension of the PAM to prosodic domains (= PAM-L2) (So, 2010; So and Best, 2011, 2014), and is supported by the findings in Braun et al. (2014), Francis et al. (2008), Hallé et al. (2004), So (2010), So and Best (2011) and So and Best (2014).

As for the first account, Wayland and Guion (2004) investigated the discrimination ability of Thai tone contrasts testing English, Chinese and Thai L1 listeners presenting stimuli with two durations of ISI (500 ms vs. 1500 ms). They reported that the Chinese L1 listeners outperformed the English L1 listeners in their ability to discriminate the two Thai tones under the short ISI condition before training. After the training, the Chinese participants outperformed the English participants under both ISI conditions. The L1 Thai listeners outperformed the two nonnative listener groups. The authors argued that the better performance shown by the L1 Chinese listeners and their improvement after the training was evidence for the fact that the Chinese listeners mapped the Thai lexical tone contrasts onto their L1 lexical tone categories while the English L1 listeners failed to do so, because their L1 intonational categories were not used at the same level as in Thai. This account is surprising as F_0 is linguistically (but not always lexically) used in most of the languages and thus it should not be unfamiliar to listeners of non-tone languages. Despite this fact, the L1 listeners of tone languages showed the advantage in processing tonal stimuli over the L1 listeners of non-tone languages. The findings suggest that the L1 listener of a language that uses a prosodic contrast lexically is more sensitive to the contrast of that prosodic cue than the L1 listeners of a language that does not use such a contrast lexically, but at different linguistic levels (such as post-lexically or paralinguistically), and that the lexical use of prosody outweighs the use of prosody at the other linguistic levels in processing.

As for the second account, there are studies claiming that the prosodic cue used at different linguistic levels between an L1 and L2 can be transferred to the L2 perception (Francis et al., 2008; Hallé et al., 2004; So, 2010; So and Best, 2011). These studies apply the account of the PAM to prosodic domains to predict learners' difficulties in perceiving L2 prosodic contrasts through (dis)similarities between L1 and L2 prosodic categories regardless of the linguistic levels on which the prosodic cue is used. For example, So and

Best (2011) examined L1 listeners of non-tone languages (Australian English and French) who were able to perceive nonnative Mandarin lexical tones according to L1 native intonational categories. The categorisations were based on the contextual phonetic similarities of the pitch contours they perceived between Mandarin tones and their L1 intonational categories. Also Francis et al. (2008) adopted the PAM to explain their predictions and outcomes. They investigated the identification ability of nonnative Cantonese tones testing English (a non-tone language) and Mandarin (a tone-language) listeners. Stimuli tones were embedded into a sentence and participants were given six choices to respond. The proportions of correct responses of the two L2 listener groups did not significantly differ both in a pre-test (before a training) and a post-test (after the training). This observation that the L1 speakers of a tone language and a non-tone language showed qualitatively similar patterns of perception for tone categories supports the conclusions of previous researchers that the mere presence or absence of lexical tone contrasts in the L1 are not in themselves sufficient to determine the performance in perceving L2 lexical tones. Instead, it seems that it is necessary to take the F_0 patterns into account that listeners have been exposed to in their L1 irrespective of their function as cues to tone as opposed to intonational categories. Although the two groups' proportions of correct responses were comparable, the two groups differed in terms of the kinds of errors they made before the training and the ways in which they improved as a result of the training. For example, Mandarin listeners were good at identifying the Cantonese high-level and high-rising tones, presumably because these tones were mapped onto the Mandarin Tone 1 (a high-level tone) and Tone 2 (a rising tone), respectively. English listeners were also good at identifying high-rising tones, possibly due to its assimilation to English question intonation (although this contour would be unexpected in the middle of an English sentence). However, performance on low-rising tones was not as good as might have been expected given that this contour form exists in English intonation. The authors claimed that the difference between a standard yes-no intonation pattern (L*+H H% in ToBI notation) and one aimed more at soliciting agreement (H*+H H%) may be less than categorical. They further discussed that the high-rising tone was too different from any L1 English intonational category to be successfully assimilated, or alternatively, that the low-rising tone was sufficiently similar to be interfered by the presence of the L1 category. Their conclusions might depend on the assumed degree of similarity between the L1 and L2 F_0 and raises the question of a valid measurement of cross-linguistic (dis)similarities of sounds.

Finally, there is empirical evidence that even the listeners whose L1 does not use F_0 lexically may be sensitive to nonnative tonal contrasts sorely by perceptually relying on acoustic correlates of the contrasts. So and Best (2011) showed that the listeners whose L1 does not use F_0 to convey linguistically meaningful contrasts at any linguistic levels are particularly sensitive to acoustic correlates of stimuli contrasting in F_0. They found that French speakers, but not English speakers, were sensitive to the fine-detailed phonetic feature differences between Tone 3 and Tone 4 (low/falling tone vs. high/falling tone). (Note that the Tone 3 in their experiment was produced as a low level or a low falling tone in connected speech, rather than the dipping pattern often found in citation-form productions.) They argued that the French L1 listeners failed to perceive lexical tones in a categorical manner not because they were unable to process tones per se, but they failed to map the Mandarin tones onto any particular native French phonological (intonational) categories and, as a consequence, they were more sensitive to the phonetic features of the stimuli than to phonological ones. Further, Hallé et al. (2004) investigated the discrimination ability of three Mandarin tone continua testing French L1 listeners and Taiwanese L1 listeners. They conducted an AXB discrimination task with an ISI of one second. They found that the French L1 listeners were sensitive to F_0 differences between same-category pairs across the continua, while Taiwanese L1 listeners were more sensitive to differences between tokens across different tone categories than to same-category pairs. Hallé et al. (2004) attributed the difficulty of the French listeners in identifying between-categorical tone stimuli to a difficulty in mapping nonnative tone categories onto their L1 ones, whether intonational or lexical insofar as adopting the PAM (Best, 1995; Best et al., 2001).

The phonetic sensitivity to a tonal contrast does not seem to be an advantage exclusively for the listeners whose L1 does not use F_0 to convey linguistically meaningful contrasts at any linguistic levels, such as for French L1 listeners. Qin and Mok (2013) showed that even L1 listeners of non-tone languages have been reported to show high sensitivity to within-categorical tonal contrasts as L1 listeners of tone languages even if their L1 F_0 conveys a meaningful contrast at the intonational level. They found that French and English L1 listeners outperformed Mandarin L1 speakers in discriminating tone contrasts and concluded that the prosodic systems without the constraint of tonal categorisation allowed the English and French listeners to detect the minor within-categorical difference between the level tones better than Mandarin Chinese L1 listeners did. There was no significant difference between English and French groups in tonal discrimination. Following the empirical findings of Qin and Mok (2013), listeners without the use of lex-

ical tones in their L1 perceived tones in a similar way regardless of whether their L1 has post-lexical use of F_0 or not.

Based on the findings in the previous studies, the following outcomes may be predicted for the current investigation: First, the Japanese listeners' and the German listeners' performance will not differ both in the short and long ISI condition, because the German listeners are familiar with the pitch contrast from their L1 intonation system. In this case, the outcome will support the account that the same phonological categories in the L1 and L2 alone are sufficient for a successful mapping regardless of the the phonetic difference between the Japanese and German falling pitch accent. Second, it could be possible that the German listeners' performance will not differ in the short ISI condition from that of the Japanese listeners, but will decrease in the long ISI condition compared to that in the short ISI condition. The German listeners may be able to discriminate the pitch contrast as well as the Japanese listeners in the short ISI condition by perceptually relying on the acoustic correlates of the stimulus contrast. However, in the long ISI condition, the German listeners may show difficulties in discriminating the pitch contrast, because they may fail to map the stimulus contrast into their L1 intonational pitch contrast even though they are familiar with a falling pitch accent from their L1 intonation system. Finally, it may be also the case that the German listeners' and the Japanese listeners' performance will differ already in the short ISI condition, because the listeners whose L1 uses F_0 lexically established sensitivity to pitch contours by associating lexical meaning which presumably lead to a generally higher ability in pitch processing (cf. Deutsch et al., 2006; Pfordresher and Brown, 2009) than the listeners whose L1 does not use F_0 lexically, but post-lexically or only paralinguistically. The last two possible outcomes presuppose that phonological representations of German intonational pitch contrasts were not successfully used to discriminate the Japanese pitch contrast, either because a successful mapping requires a prosodic contrast employed at the same linguistic level in one's L1 and L2 or because phonetic differences between the Japanese and German falling pitch accent impedes a successful mapping.

Recall that German L1 listeners achieved higher scores than Japanese L1 listeners in memorising Mandarin Chinese tone categories in a word-learning task in Braun et al. (2014). Since the present experiment does not test the discrimination ability of a variety of tonal patterns exhibited in Chinese or German, but is restricted to test the Japanese pitch contrast, it is reasonable to predict that Japanese L1 listeners will not underperform German L1 listeners.

Finally, in the same way as in Experiment 2, a task-irrelevant prosodic dimension was added to the pitch contrasts by presenting the stimuli in native and nonnative segmental length structures. Two stimuli in a pair were presented either in the singletons, the geminates or the long-vowels. The pairs presented in the geminates were considered to be nonnative for the German participants requiring higher demand on attention control, while those presented in the singletons and the long-vowels were more nativelike requiring lower demand on attention control. The expected outcome regarding the effect of attention control is that the German listeners' discrimination ability will be impeded by the task-irrelevant nonnative the segmental length structure, namely by the geminate pairs, but not by the singleton and long-vowel pairs. Whether this effect can be observed in both ISI conditions is an open question.

To conclude this section, the novelties of the present study are stated. It adds the following new aspects to Sakamoto (2010), the most relevant study for the current experiment, because it is the only study that systematically examined the discrimination of Japanese pitch accent contrasts and the categorisation of Japanese pitch accent by L2 learners. First, the current study applies the methodological paradigm that makes the two stages of speech processing ("input" and "from input to mental representations") more comparable than her study. Sakamoto (2010) conducted an ABX task with an ISI of 700 ms using non-speech stimuli (i.e., F_0 contours which were extracted from canonical tokens of Japanese lexical pitch accent contrasts) to test the discrimination ability of pitch accent contrasts, and a step-wise ABX task with an ISI of 700 ms using stimuli on a continuum to test a categorisation ability that requires more phonological processing. She predicted that learners could only categorise the membership of the representative, but not that of the stimuli in the boundary areas if they failed to categorise the pitch members. The task validity of an ABX task in order to test listeners' categorisation ability is questionable, since an ABX task with short ISI like hers has the risk to test discrimination ability while comparing solely *B* and *X* without taking *A* into consideration. In order to test L2 processing that requires more phonological processing, I will vary the durations of ISI, instead of conducting an ABX task. This experiment paradigm has been frequently used for studies on nonnative segmental contrasts, so that the methodological validity appears to be stable (Burnham and Francis, 1997; Wayland and Guion, 2004; Werker and Tees, 1984a). Second, I used the same stimuli for the short and long ISI condition, so that the findings become comparable with each other. The stimuli across Sakamoto's three experiments differed from each other, so that it became difficult to compare the results. Note that the experiments presented in this chapter were conducted as part of

a series of several experiments in this dissertation that used the same stimuli, testing the same participants, but varying the conditions. Therefore, the comparison between the experiments in this thesis is possible as well. Third, the present study investigates German L2 listeners, while her study tested English L2 listeners. Since previous studies on the perception and production of Japanese pitch accents investigated mainly English L1 listeners, the present study testing German L1 listeners contributes to generalise the findings in the previous studies. Fourth, I tested L2 learners as well as non-learners, while Sakamoto (2010) tested only learners in comparison to L1 listeners. The comparison between the two nonnative listeners' groups enables us to examine the effect of exposure to the L2. Fifth, psychoacoustic complexity was added by presenting the stimuli in the native and nonnative segmental length structures. The effect of pitch on the discrimination of segmental length contrasts found in Experiment 2 motivated us to examine whether native and nonnative segmental length structures similarly affect the discrimination of pitch.

4.2 Experiment

4.2.1 Methods

Participants

The same participants as in Experiment 2 attended this experiment.

Materials

The same recordings of the disyllabic triplets were used as in Experiment 2, but they were manipulated in a different way. The current experiment was designed to measure the discrimination of a pitch contrast while keeping the segmental durations of the two stimuli in a pair constant. The 6 tokens of each triplet realised in a flat pitch and a falling pitch (for example 6 /punu/ tokens realised as a flat pitch contour and 6 /punu/ tokens realised as a falling pitch contour, in total 12 tokens) were aligned on the average time point of the phon boundaries across realisations.

F_0 padding was used in order to avoid gaps due to voiceless segments (see details of the manipulation procedure Gubian et al., 2013, and Appendix A). Finally, a female native speaker of Japanese and a male native speaker of German selected the two most naturally sounding tokens for each *same* stimulus pair (N=72, for example /punu/-flat and

/punu/-flat) and two tokens from the same triplet, segmental variations, but from two different pitch conditions for the *different* stimulus pairs (N = 72, for example /punu/-flat and /punu/-falling) as experimental items. There was no disagreement among the judges.

Table 4.2 shows mean pitch ranges in semitones and 95% CIs for each pitch and segmental length condition. CI bars of geminate and long-vowel pairs in the falling pitch condition overlap by half the length of one arm of the CI, suggesting a degree of evidence for a difference. Mean pitch ranges of geminate pairs in the falling pitch condition were greater than those of long-vowel pairs.

pitch condition	**segmental condition**	**mean pitch range (in semitones)**	**95% CI**
flat	singleton	1.1	0.9–1.3
flat	geminate	1.5	0.7–2.4
flat	long vowel	1.3	1.0–1.7
falling	singleton	16.8	16.1–17.5
falling	geminate	16.5	15.4–17.5
falling	long vowel	15.5	15.0–16.0

Table 4.1 *Average mean pitch ranges and 95% CI for each pitch and segmental length condition.*

Furthermore, the slope of a pitch fall was calculated by dividing the pitch range in semitones by the duration of the pitch fall. The begin of a pitch fall started at the beginning of the second (voiced) consonant in singleton and long-vowel pairs, while the one in geminates in the middle of the (voiced) consonant.

pitch condition	**segmental condition**	**mean slope**	**95%CI**
falling	singleton	0.09	0.08–0.10
falling	geminate	0.17	0.12–0.22
falling	long vowel	0.08	0.06–0.10

Table 4.2 *Average mean pitch slopes and 95% CI for each segmental length condition in the falling pitch condition.*

The analysis revealed that the slope of the geminate pairs was steeper than those of the long-vowel and the singleton pairs. The slope of the long-vowel and of the singleton pairs seem to approach a significant difference. The slope of the long-vowel pairs was steeper than that of the singleton pairs.

Procedure

A speeded same-different (=AX) task was conducted in the same way as in Experiment 2 (using the same durations of ISI and of intertrial-intervals etc.). For pairs without a contrast (i.e. different tokens of the *same* type), there were six possible combinations for each triplet (e.g. punu/flat vs. punu/flat, punu/falling vs. punu/falling, pun:u/flat vs. pun:u/flat, pun:u/falling vs. pun:u/falling, pu:nu/flat vs. pu:nu/flat, pu:nu/falling vs. pu:nu/falling). For pairs with a contrast (= *different* type), there were six possible combinations (e.g. punu/flat vs. punu/falling, punu/falling vs. punu/flat, pun:u/flat vs. pun:u/falling, pun:u/falling vs. pun:u/flat, pu:nu/flat vs. pu:nu/falling, pu:nu/falling vs. pu:nu/flat). One base list was assembled by presenting all possible pairings of the stimuli (36 trials each for the *same* and the *different* pairs). The order of presentation was automatically randomised in each run using *Presentation* (Neurobehavioral Systems).

Analogue to Experiment 2, participants were given a short description of the experiment and the procedure on a piece of paper. The only difference between the procedure in Experiment 2 and 3 was that the participants were asked to pay attention to pitch differences, including the visual presentation on the computer (e.g. "Training session. Please pay attention to PITCH and answer soon after ANSWER is shown." etc.).

The experiment was conducted in the same room after Experiment 2. Participants were allowed to take a break between the two experiments as long as they wanted (ranging from 0 to 3 minutes).

4.2.2 Results

In total, 13824 data points were recorded (96 participants x 144 trials). From these, 100 data points had to be discarded due to timeout (11 from the Japanese listeners' data, 72 from the learners', 17 from the non-learners'). The response accuracy is shown in Table C.4 in Appendix B. In the following section, I will first present the analyses of d' scores, and then the analyses of RTs.

In the same way as in Experiment 2, I will report statistical results from LMER- models showing descriptive mean values and 95% CI error bars in plots. A set of LMER models were calculated with *d' scores* or *RTs* as dependent measure and *language group* (Japanese vs. learners vs. non-learners), *segmental length* (singleton vs. geminate vs. long-vowel), *ISI* (short vs. long) as fixed factors and *participant* as a random factor including random slopes for the fixed factors (Barr et al., 2013; Cunnings, 2012). In case the model did not converge, or the number of observations was smaller than the num-

ber of random effects, the random slope for items was removed. The rest of the analysis procedure was the same as in Experiment 2. Model specifications are shown in Appendix C.

Sensitivity to contrast: d' score analyses

As in Experiment 2, the participants' sensitivity to the contrasts was calculated using d' (Macmillan and Creelman, 2005). I calculated d' scores for each participant for each segmental length condition (singleton, geminate and long-vowel pairs) as well as short and long ISI (N=594) (for average d' scores see Table C.5 in Appendix B). Different from Experiment 2, normalised d' scores were not calculated for this experiment, because there was no condition that served as a baseline like the vowel length contrast to the consonant length contrast in Experiment 2.

The LMER shows a significant main effect of *language group* (the d' scores by the Japanese listeners were higher than those by the learners, ß= 0.86, SE= 0.28, t = 3.1, $p < 0.001$, followed by those by the non-learners, ß= 1.24, SE = 0.32, t = 3.9, $p < 0.001$, the latter two groups did not differ from each other, ß= -0.38, SE = 0.27, t = -1.4, p = 0.16). No other effects or interaction were found.

The plots in Figure 4.1 show that the Japanese listeners obtained generally higher d' scores than the learners followed by the non-learners. The CI bars show that the two nonnative listener groups did not differ from each other.

Processing difficulty: RT analyses

The RT analyses were performed to investigate task difficulty for the discrimination of the pitch contrast presented in different segmental length structures. Only the RTs in trials with different pairs were analysed for the same reason as in Experiment 2. Table C.6 in Appendix B shows average RTs for each segmental length condition and for each ISI condition and in each participant group.

After discarding RTs longer than 2000 ms and timeout data, raw RTs were aggregated for each participant for each segmental length condition and for ISI condition, see Figure 4.2 for mean aggregated RTs and 95%CI bars.

The LMER-analysis showed a significant main effect of *language group* (the Japanese RTs were shorter than those by the non-learners, ß= -152.0, SE = 57.4, t = -2.6, $p < 0.01$, followed by those by the learners, ß= -340.0, SE = 50.0, t = -6.8, $p < 0.001$, and the non-learners' RTs were shorter than those by the learners, ß= -187.8, SE = 48.5, t = -3.9, p

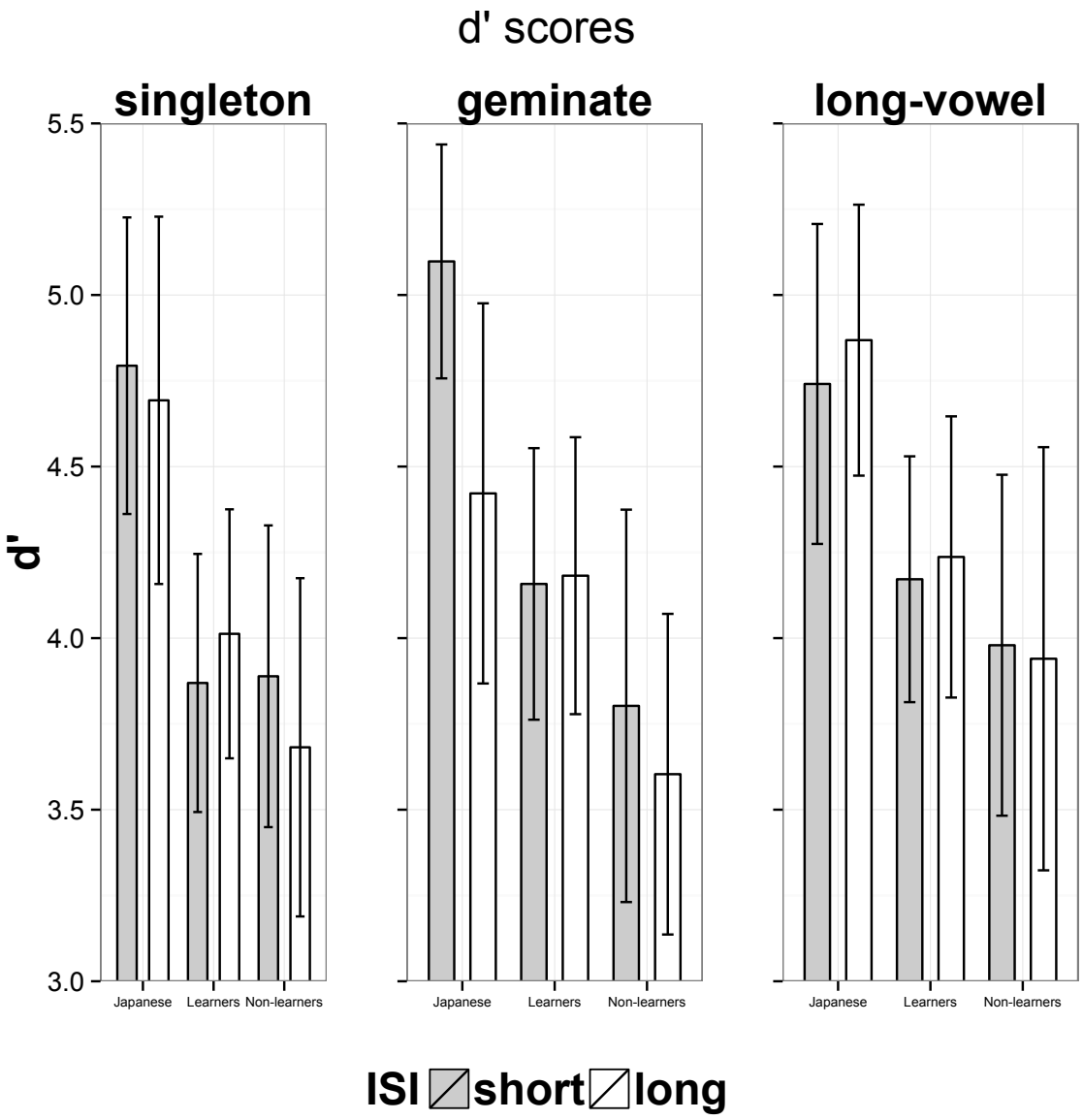

Figure 4.1 *Mean d' scores and 95%CI bars for each segment condition*

< 0.001). Moreover, there were an interaction between *segmental length* and *language group* (the learners' RTs were shorter in the geminate and singleton conditions than in the long-vowel condition compared to the Japanese listeners, ß= -61.8, SE = 27.1, t = -2.3, $p < 0.03$ for the singleton pairs, ß= -59.4, SE = 23.2, t = -2.6, $p < 0.001$ for the geminate pairs, further, the learners' RTs were shorter in the geminate and singleton conditions than in the long-vowel condition compared to the non-learners' extent, ß= -60.5, SE = 26.3, t = -2.3, $p < 0.03$ for the singleton pairs, ß= -74.4, SE = 22.6, t = -3.3, $p < 0.01$ for the geminate

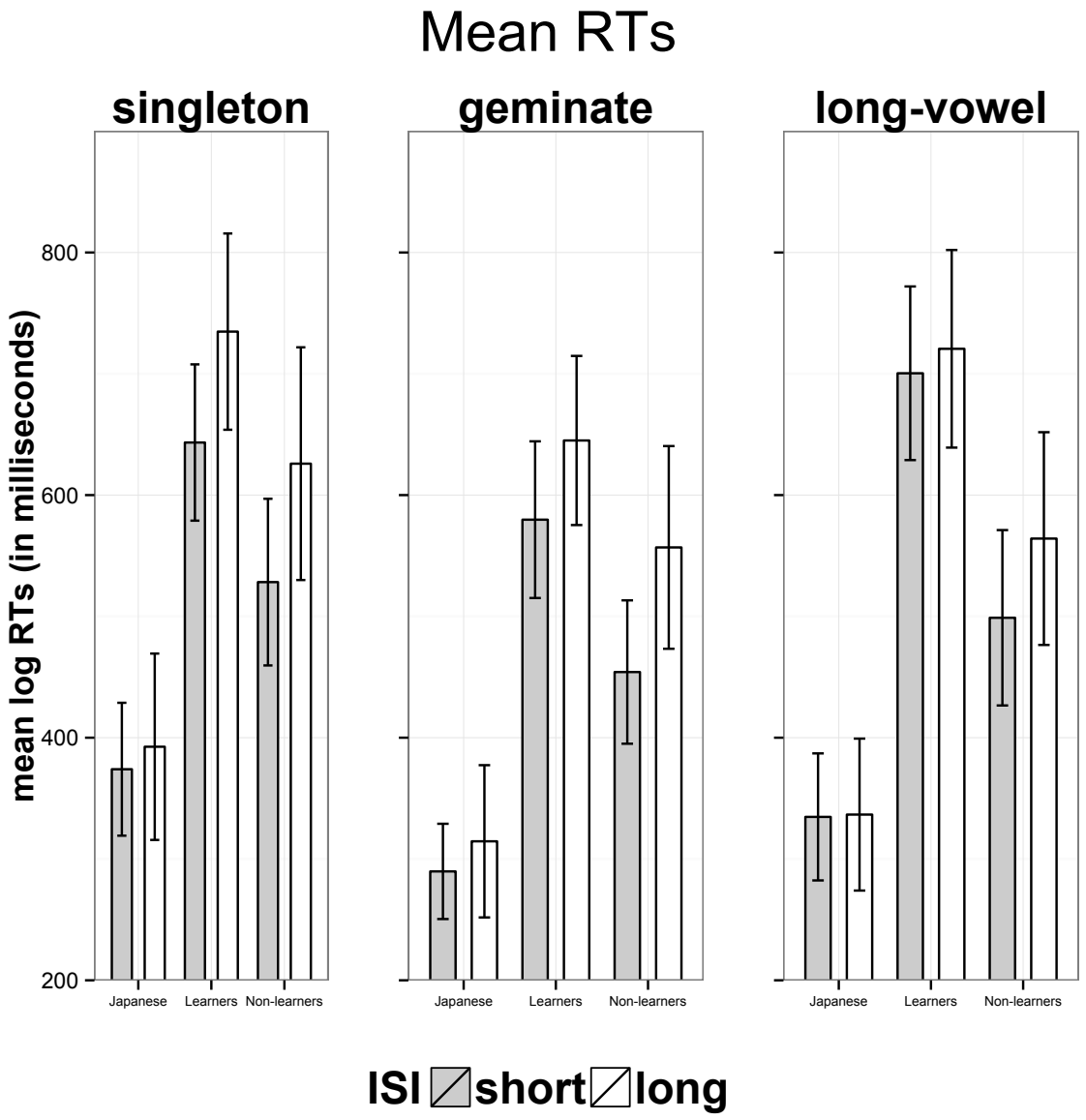

Figure 4.2 *Mean RTs and 95%CI bars for each segment condition*

pairs) and an interaction between *segmental length* and *ISI* (the RTs in the geminate and singleton conditions became longer in the long ISI condition in comparison to those in the long-vowel condition (ß= 41.8, SE = 16.1, t = 2.6, $p < 0.01$ for the geminate pairs, ß= 47.9, SE = 16.0, t = 3.0, $p < 0.01$ for the sigleton pairs).

The plots in Figure 4.2 confirm the main effect of *language* found in the LMER-analysis. Additionally, they show that the non-learners were faster than the learners in

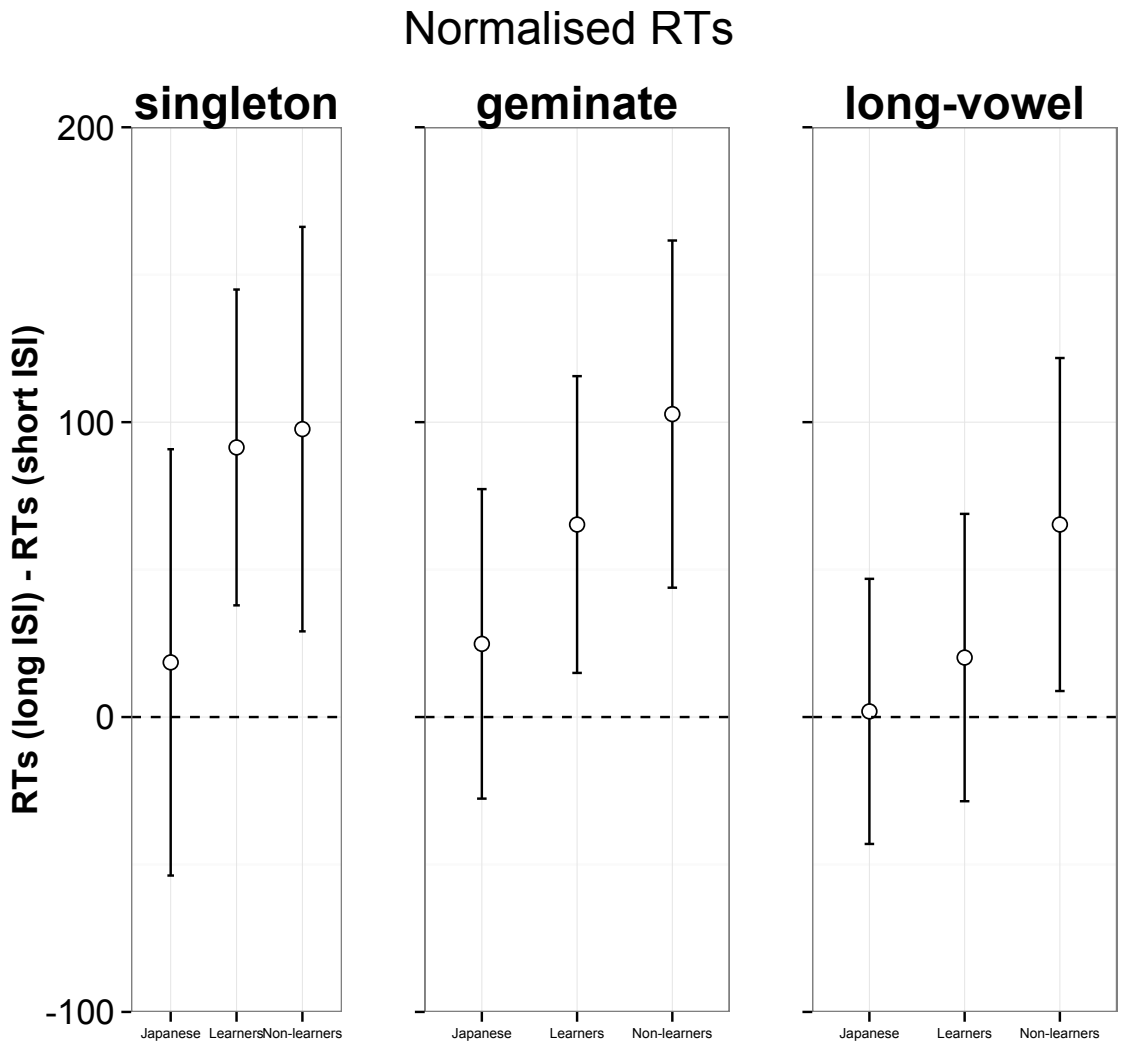

Figure 4.3 *Normalised mean $RTs_{long\text{-}short\ ISI}$ and 95% CI bars for each language group and segmental length condition*

the short ISI condition in all segmental length conditions. In the long ISI condition, the non-learners were faster than the learners only in the long-vowel condition.

In order to understand the nature of these interactions better visually, the RTs of the short ISI condition were subtracted from those of the long ISI condition, so that the ISI condition, which was a within-subject variable, was removed from the plots in Figure 4.2.

Figure 4.3 shows that both the learners' and the non-learners' normalised RTs were above 0 in the singleton and geminate condition. In the long-vowel condition, only the non-learners' normalised RTs were above 0, while those of the learners crossed 0. The Japanese listeners' normalised RTs crossed 0 regardless of the segmental conditions. The plots therefore show that the Japanese listeners were equally fast in the short and long ISI condition in all segmental length conditions, showing a minimal advantage in the long-vowel condition, in which their mean normalised $RT_{\text{long-short ISI}}$ was slightly nearer to the value of 0 and the error bars were shorter than in the singleton and the geminate conditions. Figure 4.3 also shows that no three-way interaction between *ISI*, *language group* and *segmental length* was found in the mean RTs due to the overall long error bars that overlap between the participant groups and segmental length conditions in Figure 4.3. Taken together, only the nonnative listeners' RTs became longer in the long ISI condition, especially when the stimuli were presented with the more nonnativelike segmental length structure.

4.3 Discussion

The present experiment aimed at testing the ability to discriminate acoustic correlates of a pitch contrast (corresponding to the "input" stage in Section 1), and to store the contrast into mental representations and to access them (corresponding to the path from "input" to "mental representations"). Additionally, the effect of the task-irrelevant segmental length structures on the discrimination of the pitch contrast was investigated in order to observe potential effects of attention control.

As for the results of the d' score analyses, the Japanese listeners showed generally higher d' scores than the learners, followed by the non-learners. These language group differences were found already in the short ISI condition. The two nonnative listener groups did not *statistically* differ from each other, although the raw d' scores of the learners were generally higher than those of the non-learners, which could be regarded as a tendency that the learners were at an advantage in discriminating the pitch contrasts compared to the non-learners who were not exposed to the L2. However, even the learners did not reach a sensitivity as high as the Japanese L1 listeners.

The results of the RT analyses were also in line with the results of d' scores showing an advantage for the L1 listeners already in the short ISI condition. The Japanese listeners were generally faster than the non-learners, followed by the learners. These language group differences were found already in the short ISI condition. The German listeners

were not as good as the Japanese listeners in the condition in which memory load was low and the phonetic memory of the two stimuli was still available.

Taken together, both d' scores and RT analysis show that the Japanese listeners generally performed better than the two nonnative listeners' groups. The phonological representations of pitch contrasts in the German intonational system apparently did not help German participants to achieve a sensitivity to the contrasts as high as the one shown by the Japanese listeners. There are some possible interpretations to this finding.

First, one might argue that these differences between the Japanese and German listeners can be best explained by phonetic and phonological forms of the stimuli, which might have been more similar to Japanese than to German. The pitch falls presented in the current experiment phonetically replicated the Japanese lexical pitch accent and the stimuli had Japanese mora structures (CVCV, CVC:V, CV:CV) and were recorded by a Japanese L1 speaker. As a result, it might be the case that the stimuli sounded more Japanese-like than German-like despite the use of nonsense words. However, there is a counter-argument for this interpretation: Experiment 2 in Chapter 3 that used the same stimuli showed that the nonnative listeners' d' scores were as high as the Japanese listeners' ones when the task demands were the lowest. If the Japanese-likeness of the stimuli mattered, the nonnative listeners would have had difficulties already in the lowest task demand condition in Experiment 2, but this was not the case. In future work, neither Japanese-like nor German-like stimuli but e.g. Chinese-like tonal stimuli should be used in order to avoid this problem.

Instead of supporting the first interpretation possibility, I argue that the performance difference found already in the short ISI condition was caused by the fact that the short ISI did not necessarily prevent listeners from accessing their stored phonological information, but it already involved phonological processing to aid the discrimination. As discussed in Chapter 3, Wayland and Guion (2004) claim that phonological representations were co-activated already in the short ISI condition, although to a lesser extent than in the long ISI condition. Therefore, it should not be strictly dichotomously separated that the short ISI condition triggered only phonetic processing, while the long ISI condition only phonological processing. The higher performance shown by the Japanese listeners could be a piece of evidence that phonological representations were activated already in the short ISI condition, because phonological representations are known to aid effective speech processing (Wayland and Guion, 2004). A counter argument against this interpretation would be that an effect of the ISI condition should have been observed in any case if phonological representations were activated already in the short ISI condition, but to

a lesser extent than in the long ISI condition, because the long ISI condition activated phonological representations to a greater extent than the short ISI condition, and this would have led to more difficulties for the nonnative listeners in the long ISI condition compared to the short ISI condition. Indeed, an indication for the increased difficulties in the long ISI condition for only the nonnative listeners was shown in their RTs that became longer in the long ISI condition.

Note that the German listeners should have had phonological representations of the flat vs. falling pitch contrast that exist in the German intonation system. To explain the difference found between the Japanese and German listeners, I additionally argue that the L1 listeners of a language that uses pitch lexically (= Japanese listeners) were more sensitive to pitch than the L1 listeners of a language that does not use it lexically (= German listeners), but at different linguistic levels (such as at the post-lexical or paralinguistic level), supporting the "levels of representations" account discussed in the introduction in this chapter. The L1 listeners of a language that uses pitch lexically may have established sensitivity to pitch contours by associating lexical meaning and this may lead to generally high ability in pitch processing. Although pitch is used contrastively in German, it can be hypothesised that pitch contrasts are not associated with words in mental lexicon, so that there could be no "one-to-one" associations between the lexicon and pitch. Intonation patterns are discussed to be compositions of individual tones or they are stored as whole tunes (Bartels, 1997; Dainora, 2002; Pierrehumbert and Hirschberg, 1990), but they do not relate to specific words. As future study, it would be interesting to examine whether L1 listeners of tone-languages who are familiar with lexical tone contrasts would achieve d' scores as high as the Japanese L1 listeners and RTs as short as those shown by the Japanese L1 listeners.

Interestingly, the non-learners' RTs were overall shorter than those of the learners' in the short ISI condition, and in the long ISI condition when pairs were presented with the long-vowel structure. The finding that the learners were not at an advantage in RTs than the non-learners is identical to the result of Experiment 2. As discussed in Chapter 3, the indecisiveness observed in the learners' data suggests that two systems competed with each other (the L1 and L2 one) while processing acoustic information. Due to the competition between the two systems, they would have required more time. This discussion about the two competing language systems appears to be more plausible when they compete at the language-specific phonological level and not at the language-general phonetic level. The result therefore indicates that the competition between the L1 and the L2 system already started in the short ISI condition by activating phonological rep-

resentations already in the short ISI condition. Once acoustic speech signals were processed, phonological representations were directly co-activated (Darcy et al., 2012). Another explanation for the overall shorter RTs by the non-learners than by the learners might be that the non-learners were perceptually more sensitive to phonetic differences of the stimuli than the learners due to the lack of the appropriate phonological representations. They could only rely on phonetic information to discriminate the pitch contrast as long as it was still strongly present in the short ISI condition. This explanation presupposes that phonological representations of German intonational pitch contrasts were not used to discriminate the pitch contrast in the experiment. Finally, the learners' longer RTs in comparison to those by the non-learners might be also explained that the learners who had exposure to Japanese were more careful with responding in the experiment, perhaps because they felt being judged with respect to their Japanese performance, while the non-learners felt more free in experiment. Overall in the present experiment, the RT analyses revealed difficulties on part of the nonnative listeners that the d' score analysis did not show.

Moreover, the demand on attention control was increased by presenting the stimulus pairs with various segmental length structures. The dimension of the segmental length was task-irrelevant, although a pitch movement depends on a segmental length structure, because a pitch movement is assigned to segments. The pitch contrast presented in the geminate structure, which was supposed to be nonnative, was expected to be more difficult for the nonnative listeners' than the pitch contrast presented in the singleton or long-vowel structures.

One of the effects of the task-irrelevant segmental length structure on the discrimination of the pitch contrast was found in the learners' RTs. The learners' RTs were longer in the long-vowel condition than in the singleton and geminate conditions compared to the Japanese listeners and the non-learners. In other words, the learners were slower in discriminating the pitch contrasts, when the pairs were presented in the long-vowel structure than in the singleton and geminate structures while the Japanese and the non-learners were not. The long-vowel stimuli with a pitch fall were supposed to be phonologically more natural to German listeners than other stimuli as a pitch fall is naturally accompanied with a longer vowel duration in German (Ladd, 1996). This result suggests that the learners' indecisiveness discussed above was especially true for the stimuli presented with the more native-like phonological structure. When hearing a more native-like sounding stimulus, the learners may have activated their L1 phonological representations to a greater extent than when hearing a less native-like sounding stimulus. The

stronger competition between L1 and L2 phonological representations led to the longer RTs.

Note that not only the long-vowel structure was predicted to be more German-like than the geminate structure, but also the singleton structure was predicted to be German-like since there are singleton words in German lexicon. Why did the nonnative listeners not show any advantage with the singleton pairs? A possible explanation would be that the pitch fall presented in the singleton structure was phonetically not German-like. In German, the duration of a stressed syllable becomes longer and intensity becomes higher when F_0 falls (Niebuhr, 2007). In this experiment, however, the duration of the preceding vowel before a fall did not differ between the singletons in the flat and falling pitch. And this phonetic realisation might have led to an unnaturalness of the sounds of the singleton pairs for the German listeners.

The other effect of the task-irrelevant segmental length structure on the discrimination of the pitch contrast was also found in the RT analyses. There was an interaction between *ISI* and *segmental length*. The RTs in the singleton and geminate conditions became longer in the long ISI condition compared to the long-vowel condition and this was found regardless of participant group. One might interpret this finding as an advantage that the nonnative listeners had in the long ISI condition when the stimuli were presented in the more native-like phonological structure. However, this interpretation cannot explain why this interaction was found regardless of participant group, including the Japanese listeners.

Recall that the effect of the task-irrelevant segmental length structures on the discrimination of the pitch contrast was investigated because Experiment 2 (in Chapter 3) overall showed an effect of pitch on the discrimination of the segmental length contrasts. In Chapter 6, I will proceed with the comparison of the results from Experiments 2 and 3.

Let us go back to the first and foremost research question whether nonnative listeners would have difficulties in discriminating acoustic correlates of pitch contrasts (corresponding to the "input" stage) or in discriminating the contrasts requiring more phonological processing (corresponding to the path "from input to mental representations"). The results in this experiment showed that the nonnative listeners underperformed the native listeners already in the "input" stage, even in the condition with the lowest task demands, both in terms of d' scores and RTs. The nonnative listeners' lower d' scores were kept also in the long ISI condition, showing no effect of ISI. Phonological representations were activated once acoustic information was perceived, and the pitch contrast that exists in the German intonational system did not help the nonnative listeners to

map the Japanese pitch contrast fully, either due to the cross-linguistic phonetic differences of the same phonological category or due to the different linguistic levels at which the prosodic cue exhibits meaningful contrasts in one's L1 and L2. The finding in this experiment suggests that multiple dimensions should be taken into account to define perceived cross-linguistic (dis)similarities that the PAM-L2 requires. Moreover, an effect of the task-irrelevant segmental length structure on the discrimination of the pitch contrast was found. The learners' RTs were longer in the long-vowel condition than in the singleton and geminate conditions compared to the Japanese and the non-learners' RT differences between the segmental length conditions. Moreover, the learners' and the non-learners' RTs became longer in the singleton and geminate conditions in the long ISI condition. These two effects of the task-irrelevant segmental length structure suggest that the stimulus pairs that required lower demand on attention control (that is, the stimulus pairs presented in the more native-like segmental length structures) activated nonnative listeners' L1 phonological representations to a greater extent than the stimulus pairs that required higher demand on attention control. Therefore, the nonnative listeners' performance was less affected when the stimulus pairs were presented in the familiar segmental length structure even when memory load increased. The nonnative listeners' performance decrease suggests that higher demand on attention control impeded successful speech processing under higher memory load, suggesting a more dominant influence of *the central executive* than *the phonological loop*.

CHAPTER 5

IMMEDIATE AND DELAYED IMITATION OF SEGMENTAL LENGTH CONTRASTS AND PITCH FALLS

5.1 Introduction

Experiment 1 showed that L2 speakers produced deviant forms from those produced by L1 speakers. In Experiment 2 and 3, it was analysed whether these deviations related to the L1 speakers' ability to discriminate and/or to store nonnative prosodic contrasts. Under certain experimental conditions, the nonnative listeners could even discriminate a nonnative prosodic contrast, though their ability was found to be vulnerable being easily affected by increased cognitive load. Based on the findings in Experiments 2 and 3, it seems to be reasonable to predict that nonnative listeners will fail to produce nonnative prosodic properties correctly due to the observed difficulties in the initial stages of their speech processing (in the "input" stage). Though, despite the difficulties found in the perception experiments in this thesis, nonnative speakers may still be able to produce the nonnative prosodic properties as accurately as L1 speakers. Theoretical and empirical supports for this assumption are provided later in this section.

In this experiment, nonnative speakers' production accuracy regarding the nonnative consonant length and pitch falls was examined in immediate and delayed imitation tasks. Analogue to the experimental paradigms in the perception experiments, the duration between the offset of the stimuli and the begin of the imitation was varied in order to increase memory load. The immediate imitation task was assumed to test speech production that can possibly take place without accessing phonological representations corresponding to the path from "input" directly to "output" while the delayed imitation task was expected to test speech production that requires phonological representations corresponding to the path from "input", "mental representations" to "output".

There are only very few production studies that investigated the acquisition of L2 consonant length contrasts (e.g. Kabak et al., 2011 testing German L2 - Italian L1 speakers, Han, 1992, 1994 ; Mah and Archibald, 2003 testing English L2 - Japanese L1 speakers). They reported that it is difficult for L2 learners to produce nonnative consonant length contrasts that are not used lexically in one's L1. Kabak et al. (2011) showed that native-like timing of geminate consonants was difficult to produce even for advanced German L2 learners of Italian with a considerable amount of exposure to an L2 (the duration of Italian learning ranged between 5 and 10 years in high school and university and/or the duration of stay in Italy ranged between 6 months and 3 years and all learners self-rated their knowledge of Italian as "very good" to "fluent"). The authors examined the geminate-singleton duration ratios of nonsense word minimal pairs produced by 10 naïve German speakers, 10 advanced German L2 learners of Italian and 8 Italian L1 speakers. The results showed that both German groups were able to produce geminates differentiated from singletons in production despite the absence of this contrast in their L1. However, they found significant differences in the geminate-singleton duration ratios of nonsense word minimal pairs across groups (naïve < advanced learners < Italian). Interestingly, the developmental trajectory in advanced learners' geminate-singleton ratios was not due to an improvement in the timing of geminates, but rather due to the shortening of singletons. This finding suggests that L2 learners are better at readjusting an existing category in their L1 (i.e., singletons) rather than showing an improvement in a novel category. Han (1992) investigated the production of the stop closure durations and voice onset time of geminate and single voiceless stops in Japanese (/tt/, /kk/, /pp/) testing 10 Japanese L1 speakers and 4 advanced American L2 learners. The results indicated that the L1 speakers distinguished between the geminate and single stops by controlling the closure durations in the mean ratio of 2.8:1.0 (ranged from 2.5 to 3.2:1.0), while the L2 learners pronounced the same tokens in diverse and random manners (mean = 2.0:1.0, ranged from 0.9 to 4.0:1.0). This study also showed that the timing control of L2 geminate and single stop consonants was a challenge even for advanced L2 learners. Mah and Archibald (2003) examined the acquisition of different types of consonant and vowel length contrasts (/tt/, /pp/, /kk/, /ɕɕ/, /nn/, /mm/, /a:/, /i:/, /ɯ:/, /ɛ:/, /ɔ:/ each in comparison to short consonants or short vowels. However, they tested only one Canadian L2 learner of Japanese whose Japanese proficiency was extremely limited (= a beginner). She read Japanese minimal pairs containing the consonant and vowel length contrasts. The authors claimed that the subject has acquired a length contrast in her Japanese, as she consistently produced geminate consonants and long vowels significantly longer than

their single/short counterparts (geminate/singleton duration ratios ranged between 1.8 and 4.0 and long-vowel/singleton ones between 2.3 and 3.2). However, the generalisation of their result from only one subject is questionable.

As for production studies investigating L2 pitch accent contrasts, only a limited number of studies exist up to date (Hirano-Cook, 2011; Matsuzaki, 1995; Nakagawa, 2001; Sukegawa, 1999). Except for Sukegawa (1999), they all examined the effect of training to improve the L2 learners' production of Japanese pitch accent and intonation. So far, none of the studies conducted an imitation task.

Sukegawa (1999) reported a case study of pitch realisation of two or three mora words by two advanced Brazilian L2 learners of Japanese. He showed that the learners tended to give a high pitch to the heavy syllables. Hirano-Cook (2011) conducted a read-aloud task using a pretest-posttest paradigm (e.g. Collentine, 1998; Dimitrov and Rumrill, 2003) with 33 American English learners of Japanese (27 learners for an experimental group and 6 for a control group) taking the second or the third-year level of Japanese language courses at a college. They were asked to read aloud 40 noun stimuli (= pretest) in a list (list 1), and then in another list (list 2) that contained the identical stimuli, but with information about the accent location of each word using a symbol. During the training session, participants were asked to read aloud the stimuli in the list 2. When the participants in both groups mispronounced a word, they were instructed to repeat it until they did so correctly. When instructed to repeat, no correct model was provided. The posttest was conducted in the same way as the pretest. During the training, only the participants in the experimental group were made aware of the Japanese accent. No one had formal knowledge of the symbol's use in the list 2 prior to training. Therefore, they did not know how to produce a proper accentual pattern in their pretests. The Japanese L1 listeners judged the accuracy of the productions using a goodness rating. The results of the rating revealed that both the experimental and control group improved their productions in the posttest to the same extent when the stimulus list contained accent information. However, when the list did not provide such an aid, the experimental group improved their productions to a greater extent than the control group. This ability is especially important, since accent information is not provided in Japanese texts.

Matsuzaki (1995) invented a *prosody graph*, a visual simplified pitch contour and examined the effect of the prosody graph testing Korean L2 learners of Japanese. The participants read aloud dialogues after three different practice conditions: 1) no accent information, 2) with a prosody graph, and 3) with a commonly used simple Japanese accent symbol. The utterances produced with the prosody graph had better results than the ut-

terances produced in conditions 1 and 3. Nakagawa (2001) also attempted to instruct prosody to six intermediate and six advanced L2 learners with various L1 backgrounds (including tone languages). The L2 learners were told to pay particular attention to accent and intonation and to practice pronunciation during training. Learners were instructed that the intonation of Japanese phrases appears like a "he"-shape, a string of pitch of high and low will look like the Japanese hiragana character "he" (へ). Nakagawa stated that the L2 learners' intonation forms were improved after training with the increase of proper word accentual patterns. This indicated that the correct realisation of accentual patterns at the word level contributes to a realisation of the "he"-shape of intonation at the sentence level. The study suggests that the correct realisation of accentual patterns was more difficult than improving intonation.

There are further imitation experiments that investigated L1 productions of F_0 contours (e.g. Braun et al., 2006; Pierrehumbert and Steele, 1989). One of the most relevant studies among them is Braun et al. (2006) that tested whether intonation in speech is restricted to a small, limited set of patterns although the pitch of the human voice is continuously variable. They asked 10 subjects to mimic a block of 100 randomly generated intonation contours and then to imitate themselves in several successive sessions. They found that the produced F_0 contours gradually converged towards a limited set of distinct, previously recognised basic English intonation patterns after the first iteration. Their finding showed that phonetic details gradually decayed and the participants' productions approached towards their phonological representations, each time when participants imitated their own productions. Following their findings, in the current experiment, it is expected that participants who do not have phonological representation of a stimulus will have increasing difficulties in imitating the stimuli in the delayed imitation condition compared to the immediate imitation condition, because in the delayed imitation condition, phonetic details of the stimuli are expected to decay while waiting to speak.

The novel aspect in this study is that memory load was varied in order to differentiate imitation processing that makes use of acoustic echo and that requires to access the phonological representations after phonetic information decayed. None of the previous studies considered this distinction in L2 production. The speech production using acoustic echo was tested in an immediate imitation task in which participants imitated pseudo-words immediately after the offset of the stimulus, in other words, the duration between the offset of the stimuli and the start of the imitation was 0. In this condition, acoustic echo was expected to be still available. The theory of working memory (e.g.

Baddeley et al., 1998; Baddeley and Hitch, 1974) also supports the view that the L1 and L2 productions should not differ in the immediate imitation task. In this condition, the acoustic echo content of the stimulus words should not have yet decayed, and participants may use it to execute a phonetic plan (Levelt, 1989, 1999). Moreover, imitation processing in an immediate condition may take place without accessing the phonological representations of a speaker. According to direct realism (Fowler, 1986), an immediate imitation is driven by reflexive phonetic gestures that are mediated automatically in speech perception without requiring the access the phonological representations of the speaker. If this account is valid, the lack of nonnative phonological representations should not disturb the imitation of a sound with a nonnative phonological structure. This imitation processing is also illustrated in the model of speech perception and production (Ramus, 2001; Szenkovits and Ramus, 2005) (see Figure 1.4 in Chapter 1). The model presents a direct connection from *acoustic representation* to *articulatory representation* without mediating *sub-lexical phonological representation*.

Speech production that requires to access the phonological representations after phonetic information decayed was investigated in a delayed imitation task in which participants heard the same words, but waited 2500 ms seconds before speaking. In this condition, memory load was expected to increase compared to that in the immediate imitation task, because the stimulus, which is recognised immediately, has to be kept in the participants' working-memory accessing their phonological representations while they are waiting to speak. Therefore, the absence of nonnative phonological representations in long-term memory might impede successful imitation. This predicted L2 performance decrease in the delayed imitation condition is supported by several theories and models: Baddeley and Hitch (1974), Hintzman (1986) and Hintzman (1988) who supported exemplar theories predicted that imitation accuracy would decrease over longer time delays. While waiting to speak, continuous interactions occur between working memory and long-term memory in order to maintain and refresh the acoustic information through inner speech or rehearsal. Every time the echo in working-memory is communicated to long-term memory again, the next echo will move closer to the central tendency of the stored category, because the feedback loop will attract a representation toward the mean of the stored category. In this way, in L2 processing, the feedback from long-term memory progressively assimilates the input sound to a learners' L1 representation. Idiosyncratic details of the original stimulus will be attenuated in the eventual echo used for output (Goldinger, 1998, 256). After several seconds (Baddeley, 1986; Goldinger, 1998), the echo in working-memory will become the mental category prototype of the learners'

L1 and no longer what was initial perceived. Thus, imitation accuracy should decline in the delayed imitation for learners (see ibid.).

One of the frequently cited studies that used this immediate vs. delayed imitation/ shadowing paradigm was Goldinger (1998). He tested the extent to which idiosyncratic details of the original stimuli decay in the course of time. The dependent variable was shadowing RTs and the independent variables were the number of repetitions, the frequency of the words used as stimuli or talker variability (different vs. same voices of the stimuli). He found that the effects of frequency and talker variability became smaller in the delayed shadowing task, suggesting that such stimulus specific information was attenuated in the course of time. The duration of the delay in the delayed-shadowing condition was 4000 ms. In the current experiment, it was decided to set the time to 2500 ms for the delayed imitation condition and not 4000 ms as in Goldinger (1998). In order to keep the participants' motivation and concentration high, 2500 ms was better than 4000 ms, if there were not any specific reasons in favour of the longer duration. There was no other theoretical argument that speaks only for 4000 ms rejecting 2500 ms. Moreover, this duration corresponded to the long ISI in the perception experiments, so that this method enabled us to compare the results of the perception and production experiments.

At this point, it should be emphasised that the imitation task such as the one presented in this study was a suitable task to reveal difficulties that L2 speakers encounter in their production in daily life situations, because it can vary task demands and can trace back to the basis of speech production and examine whether L2 speakers show difficulties in a task with lower memory load or whether their difficulties arise when memory load increased.

Apart from the imitation processing that requires or does not require to access phonological representations, there is another possible source that may owe deviant L2 productions; difficulties that relate to articulatory or motor processes (Esling and Wong, 1983; Kerr, 2000; Mennen et al., 2010b; Strange, 2007). Languages may vary in their phonetic (or articulatory, voice-quality) setting (Esling and Wong, 1983; Honikman, 1964; Laver, 1994; Mennen et al., 2010b), i.e. in the way in which to make the vocal apparatus (such as lip, tongue, jaw) configured for language-specific habits. Laver (1994) claims that the idea of this setting is applicable at every level of phonetic realisation, including the prosodic level. According to this view, L2 learners may have difficulties in producing nonnative pitch falls, not because it is physiologically impossible to produce them, but because they are not good enough to coordinate vocal apparatus in the way that it is required to produce the nonnative contrast. Mennen and her colleges showed that

German L2 learners of English L1 speakers had different strategies to realise large pitch range in English. The German L2 speakers tended to have a higher pitch level in their English than in their German, but not a wider pitch span in their English as English L1 speakers did (Mennen et al., 2008, 2012, 2010b). This suggests that the majority of L2 learners in her study may only be adjusting one dimension of pitch range (i.e. level), rather than the dimension in which a cross-language difference in setting is apparent (i.e. span). She postulated that the L2 speakers followed different strategies in their production of pitch range settings across the two languages. She further discussed that the narrower pitch span that German L2 speakers of English produced may contribute to foreign accented speech often perceived by English L1 speakers. Note that the difficulties relating to phonetic setting are relevant only for pitch falls, but not for segmental length contrasts, because the realisation of segmental length contrasts does not require different coordination of articulatory apparatus for the short and long segment. In the current study therefore, the difficulties relating to articulatory coordination were excluded for the analysis of the realisation of segmental length contrasts. Moreover, the stimuli used in the current experiment were articulatorily not difficult for both Japanese and German participants, so that all participants should not have had difficulties in articulating the segments of the stimuli.

The accuracy of the imitation was defined as how similar the productions were to the stimulus words. More precisely, the following aspects were measured. For the quality of the production of native and nonnative segmental length contrast, the duration ratios of the critical consonants of the stimuli and of the participants' productions were measured. Duration ratios between singleton and geminate consonants have been extensively analysed in previous studies (e.g. Fujisaki et al., 1975; Harada, 2006; Hirata and Whiton, 2005) confirming that they are the major acoustic correlate and perceptual cue for the distinction between single/geminate consonants in Japanese. As for the quality of the production of a pitch fall, the pitch slope of the fall (in the falling pitch condition) was calculated by dividing the pitch range of the fall (in semitones) by the duration of the fall (in seconds). The analysis of the pitch slope examined whether the nonnative speakers had difficulties in readjusting the L1 phonetic setting to the L2. Since Japanese speakers are known to produce a rapid fall after the F_0 peak preceded by a flat pitch (Kondo, 2009; Sakamoto, 2010; Vance, 1987), it was expected to find a steeper slope in the words with a pitch fall for the productions uttered by the L1 speakers than by the nonnative speakers.

Based on the theoretical backgrounds, the following hypotheses for segmental length contrasts and pitch falls are stated separately. Note that the pitch fall is not completely

nonnative for German participants since a falling vs. a flat pitch are contrastive at the post-lexical or paralinguistic level in the German intonation system, while the consonant length contrast is absent in their L1 lexicon. Moreover, the articulatory difficulties discussed above are true only for the imitation of pitch falls.

As for the segmental length contrasts, the following three hypotheses are possible: First, the consonant length contrast produced by the nonnative speakers may not differ from that of the stimuli and that produced by the Japanese L1 speakers both in the immediate and delayed imitation. This hypothesis lends support to direct realism (Fowler, 1986). Following her theory, the L1 and L2 processing should not differ from each other neither in the immediate nor in the delayed imitation task, because the task will not be interfered by the fact that appropriate phonological representations do not exist in the nonnative listeners' long-term memory. Second, the consonant length contrast produced by the nonnative speakers may not differ from that of the stimuli and of the L1 speakers in the immediate imitation condition. In the delayed imitation task, however, the contrasts produced by the nonnative speakers may differ from those of the stimuli and those produced by the L1 speakers due to the lack of appropriate phonological representations in the nonnative speakers' lexicon. The second hypothesis is plausible if the reflexive gestures that do not mediate mental representations (Fowler, 1986; Liberman et al., 1967) work only in the immediate imitation condition, in which acoustic echo is assumed to be still available. Third, the consonant length contrast produced by the nonnative speakers may differ from that of the stimuli and that produced by the L1 speakers already in the immediate imitation condition. This may hold true in the delayed imitation condition, because phonological representations may be activated already in the immediate imitation condition and auditory information directly co-activates phonological representations (Darcy et al., 2012; Wayland and Guion, 2004).

As for the pitch falls, the pitch slope produced by the nonnative speakers may not differ from that of the stimuli and that produced by the Japanese L1 speakers both in the immediate and delayed imitation. The nonnative speakers will not have difficulties in a different phonetic setting in the L2 and they can keep the nonnative phonetic setting after phonetic information decayed. Second, the pitch slope produced by the nonnative speakers may not differ from that of the stimuli and that produced by the Japanese L1 speakers in the immediate imitation. However, in the delayed imitation, they will differ from each other. After phonetic information decayed, the nonnative speakers may fail to readjust the phonetic setting and the normalised idiosyncratic details of the stimuli. Third, the pitch slope produced by the nonnative speakers may differ from that of the

stimuli and that produced by the Japanese L1 speakers already in the immediate imitation. This may hold true in the delayed imitation condition. The third hypothesis indicates that the nonnative speakers may have difficulties in readjusting their L1 phonetic setting to the L2 (Esling and Wong, 1983; Kerr, 2000; Mennen et al., 2010b; Strange, 2007). The three possible hypotheses depend on whether the readjustment of the L1 phonetic setting to the L2 will be normalised in the course of time, because idiosyncratic details of the original imitation stimulus will be attenuated in the echo which are used for the output (e.g. Baddeley, 1986; Goldinger, 1998).

5.2 Experiment

5.2.1 Methods

Participants

The same participants as in Experiments 2 and 3 took part in Experiment 4.

Materials

The same triplets were used as stimuli as in Experiments 2 and 3. Since the sex of the speaker who recorded experimental stimuli has been reported not to influence difficulties in shadowing or imitation particularly for one sex (Namy et al., 2002), the stimuli produced by one female speaker were used for both male and female participants. Then, considering that this would require male participants to transpose responses, which may have proven difficult for some, the female voice was dropped by 4 semitones. Phonetically, the stimuli with the falling pitch exhibited the typical Japanese pitch accent with a drastic pitch fall preceded by a flat pitch contour. Furthermore, the stimuli were manipulated both in terms of pitch contours and segmental durations of the original recordings. Note that pitch contours were manipulated in Experiment 2, and segmental durations in Experiment 3, but not both of them at the same time. The 6 tokens of each triplet realised in the same segmental length condition and pitch (for example the 6 /punu/ realised in HH) were aligned on the average time location of the phon boundaries, and on the average pitch across realisations. This means that the each set of 6 tokens had the same segmental durations and pitch ranges after the manipulation, see the average segmental durations and pitch ranges Table C.9 in Appendix B.

After the manipulation, a female native speaker of Japanese and a male native speaker of German selected the most naturally sounding tokens from the six tokens of each category. In this way, in total, 36 experimental items were selected (6 triplets x 3 segmental length conditions x 2 pitch conditions). There was no disagreement on the decisions.

Procedure

Participants took a pause (for as long as they wanted, ranging between 0 and several minutes) after Experiment 3 and participated in the imitation experiment in the same room.

Experimental list was constructed by presenting all stimulus words (totalling in 36 trials). Different from Experiments 2 and 3, the stimuli were not presented in a pair this time, but one by one. The same list was used for the immediate and the delayed imitation task. In total, a participant imitated 72 words (36 stimuli x 2 imitation conditions) in a given order. The constrains for the randomisation were to keep at least 3 trials between the stimuli of the same reference pseudoword and at least 2 trials between the stimuli of the same segmental length condition. In this way, similar stimuli were presented separately with a certain distance. The experiment was programmed and presented using *Presentation* (Neurobehavioral Systems). Auditory stimuli were presented via headphones (Sony MDR-CD570).

All participants took part in the immediate imitation task at first, and then in the delayed imitation task. Before each task, participants were given a short description of the experiment and of the procedure on a piece of paper. They were instructed to imitate stimuli as correctly as possible after a cross was shown on a screen. The aim of the study was not communicated to the participants. All written instructions were given in English in the same way for all three participants groups. After reading the description, they sat in front of the computer displaying “Welcome! The experiment will start soon. Press mouse button to start.”. Participants then clicked on the mouse to start a training session. Both the immediate and delayed imitation tasks began with the same 10 training trials using the triplets that were not used as experimental ones (*guna* and *puna*). The training started with the visual presentation of “Training session. Please start to imitate after a cross was shown. Press a button on the box to start.”. After the training, there was a pause (1 minute) before the experimental part started. The experimental part started with the visual presentation of “Experimental session”. The rest of the visual instruction was the same as for the training session. There was no pause during the task. Each trial began with a sinusoid beep of 44100 Hz (500 ms) followed by 500 ms of silence. After

this start signal, the auditory stimulus was presented. In the immediate imitation condition, a cross was shown at the offset of the stimulus. In the delayed imitation condition, it was shown 2500 ms after the offset of the stimulus. Participants were then given a maximum of 2500 ms to imitate before timeout. The intertrial-interval after the timeout was 1000 ms. No feedback was provided during the experiment. Participants' responses were recorded using a portable digital speech recorder (M-Audio Micro Track II Digital-Recorder) via a microphone with a 41kHz sampling rate and 16 bit stereo format.

F_0 extraction and segmental annotation

In total, 6912 data points were recorded (96 participants x 72 trials). F_0 contours were computed using the F_0 tracking algorithm in the Praat toolkit (Boersma and Weenink, 2011). A default range of 70-350 Hz for males and 100-500 Hz for females was used.

The segmental boundary annotation was carried out on the recorded raw data using Praat, and applying standard segmentation criteria (Turk et al., 2005). Five segmental boundaries were considered: | C| V| C| V| , | C| V| C:| V| or | C| V:| C| V| .

5.2.2 Results

Results from segmental durations

From the raw durations (see Table C.7 in Appendix B), normalised relative ratios of short-long consonants or vowels were calculated in the following way: First, the relative durations (with respect to total durations) of a long consonant were divided by those of a short consonant (= ratios of long-short consonant). Then, the ratios of the stimuli were subtracted from those of the participants' productions. In the same way, long-short vowel relative ratios were calculated. The value of 0 means that a ratio was the same as the one of the stimuli. There were 4608 ratios for the analysis.

The LMER-analysis on the consonant $\text{ratios}_{\text{production-stimulus}}$ showed a significant main effect of *language group* (the ratios produced by the learners were smaller than those produced by the Japanese, ß= -0.42, SE = 0.05, t = -8.7, $p < 0.001$, the ratios produced by the non-learners were smaller than those produced by the Japanese, ß= -0.58, SE = 0.06, t = -10.6, $p < 0.001$, the ratios produced by the non-learners were smaller than those produced by the learners, ß= -0.17, SE = 0.05, t = -3.5, $p < 0.001$). Moreover, there was an interaction between *language group* and *pitch* (the difference between the ratios produced by the learners and by the Japanese became smaller in the falling pitch condition compared to the flat pitch condition, ß= 0.21, SE = 0.04, t = 5.0, $p < 0.001$, and the

same was true for the difference between the ratios produced by the non-learners and by the Japanese, ß= 0.22, SE = 0.05, t = 4.4, p < 0.001, but no difference was found between the learners and the non-learners, p = 0.9).

The LMER-analysis on the vowel $\text{ratios}_{\text{production-stimulus}}$ showed a significant main effect of *language group* (the ratios produced by the non-learners were larger than those produced by the Japanese, ß= 0.17, SE = 0.08, t = 2.2, p < 0.03, the ratios produced by the learners tended to be larger than those produced by the Japanese, ß= 0.12, SE = 0.06, t = 1.8, p = 0.08, the two German groups did not differ from each other, p = 0.4). Moreover, there was an interaction between *language group* and *pitch* (the difference between the ratios produced by the learners and by the Japanese became much larger in the falling pitch condition than in the flat pitch condition, ß= 0.17, SE = 0.05, t = 3.5, p < 0.001, and between the ratios produced by the non-learners and by the Japanese, ß= 0.13, SE = 0.06, t = 2.3, p < 0.03, no interaction between the learners and non-learners, p = 0.5).

Figure 5.1 and 5.2 show mean normalised consonant and vowel relative $\text{ratios}_{\text{production-stimulus}}$ and 95%CI bars for each condition. The plots show that the ratios produced by the Japanese speakers did not differ from those of the stimuli in all conditions. As for the consonant duration ratios, those produced by the learners were smaller than those of the stimuli, and those produced by the non-learners were much smaller. This was true for both pitch conditions. The ratios produced by the learners and the non-learners differed from each other. There seems to be no interaction between the imitation condition (henceforth *time* = immediate vs. delayed imitation condition) and *language group* in both pitch conditions.

As for the vowel duration ratios, the ratios produced by the Japanese did not differ from those of the stimuli both in the flat and falling pitch conditions. The ratios produced by the learners and the non-learners were larger than those of the stimuli in both pitch conditions. In the falling pitch condition, the differences between the stimuli and the productions by the learners and the non-learners became greater compared to the Japanese group.

Results from pitch slopes

The slope of the pitch fall was analysed for the contours in the falling pitch condition. Maximum and minimal pitch in Hz were automatically extracted from each of the 6912 data. The maximum pitch occurred before the minimum pitch. The slope was calculated by dividing the pitch range (maximum - minimum pitch) in semitones by the duration of the fall ($\text{duration}_{\text{pitch maximum - minimum}}$ in second), see Table C.8 in Ap-

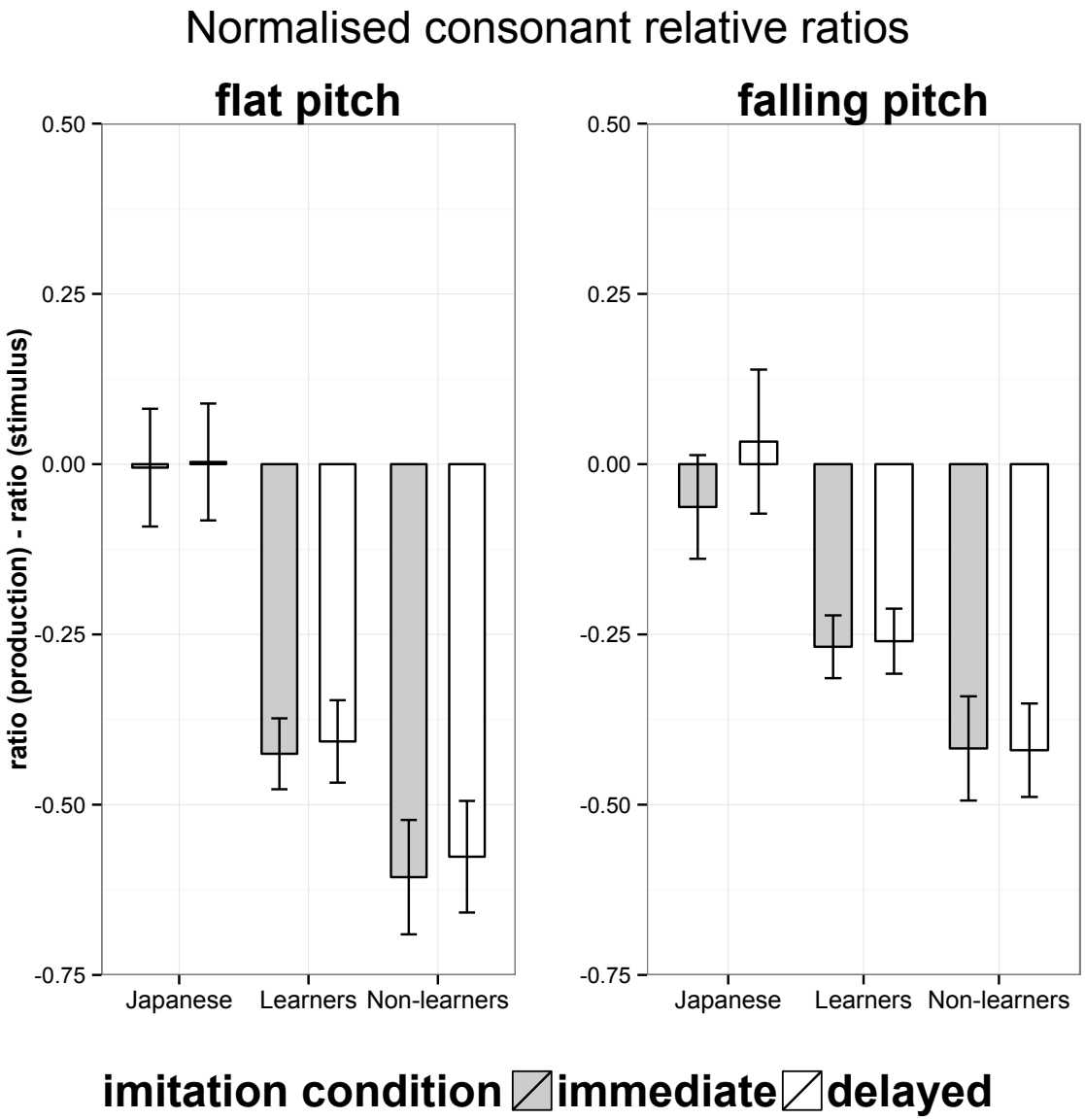

Figure 5.1 *Mean normalised consonant ratios$_{production\text{-}stimulus}$ and 95%CI bars in the flat pitch condition (left) and in the falling pitch condition (right) for each language group and imitation condition*

pendix B. Then, the slope of the stimuli were subtracted from that of the productions (= slope$_{production\ -\ stimulus}$).

The LMER-analysis showed a significant main effect of *language group* (the slope$_{production\ -\ stimulus}$ produced by the Japanese was larger than that produced by the

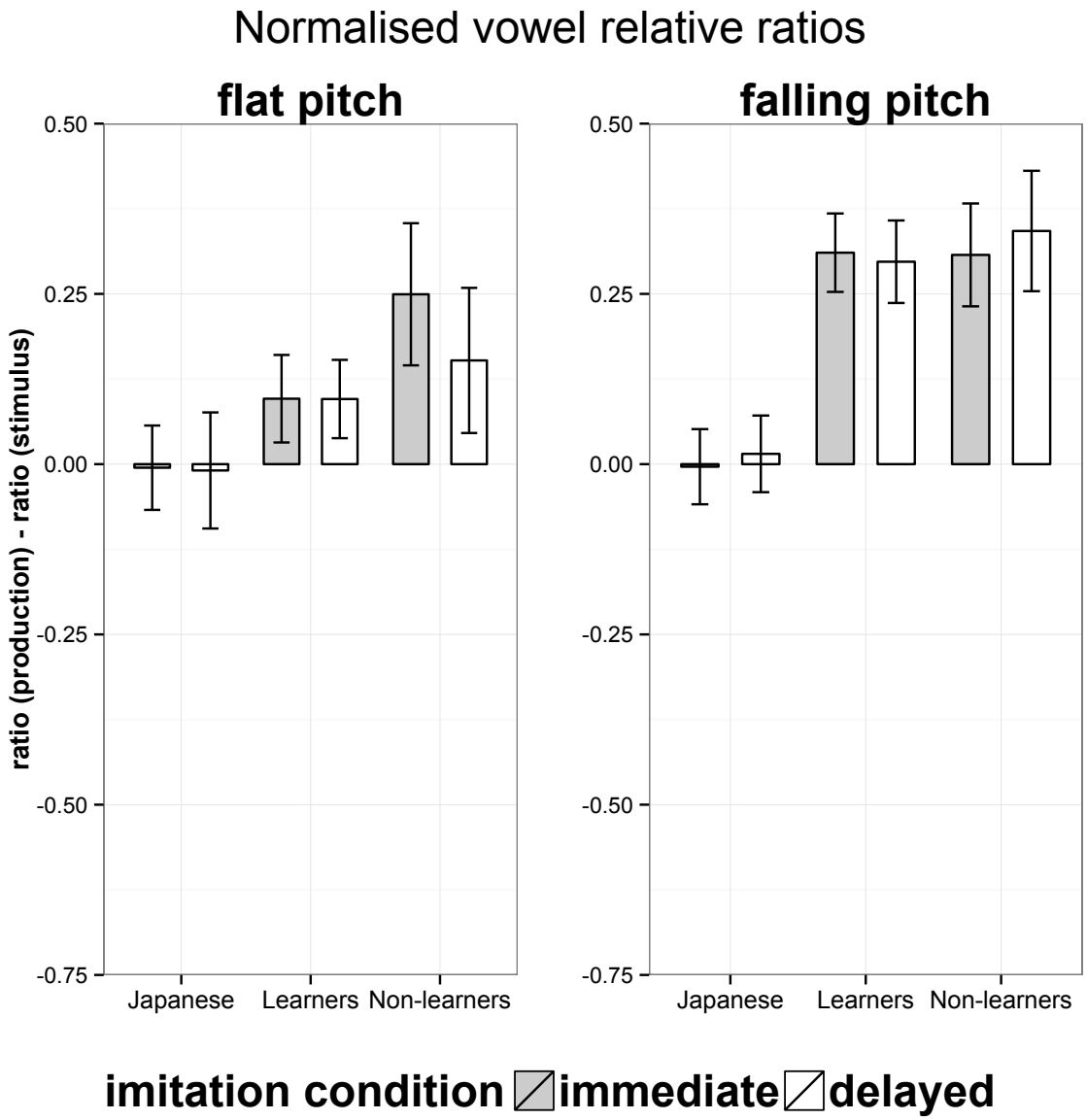

Figure 5.2 *Mean normalised vowel ratios$_{production\text{-}stimulus}$ and 95%CI bars in the flat pitch condition (left) and in the falling pitch condition (right) for each language group and imitation condition*

learners, ß= -20, SE = 3, t = -6.0, $p < 0.001$, and by the non-learners, ß= -20, SE = 4, t = -4.5, $p < 0.001$, the two German groups did not differ from each other, $p = 0.3$).

Figure 5.3 shows that the slope produced by the Japanese speakers was steeper than that produced by the learners and non-learners regardless of the segmental length conditions.

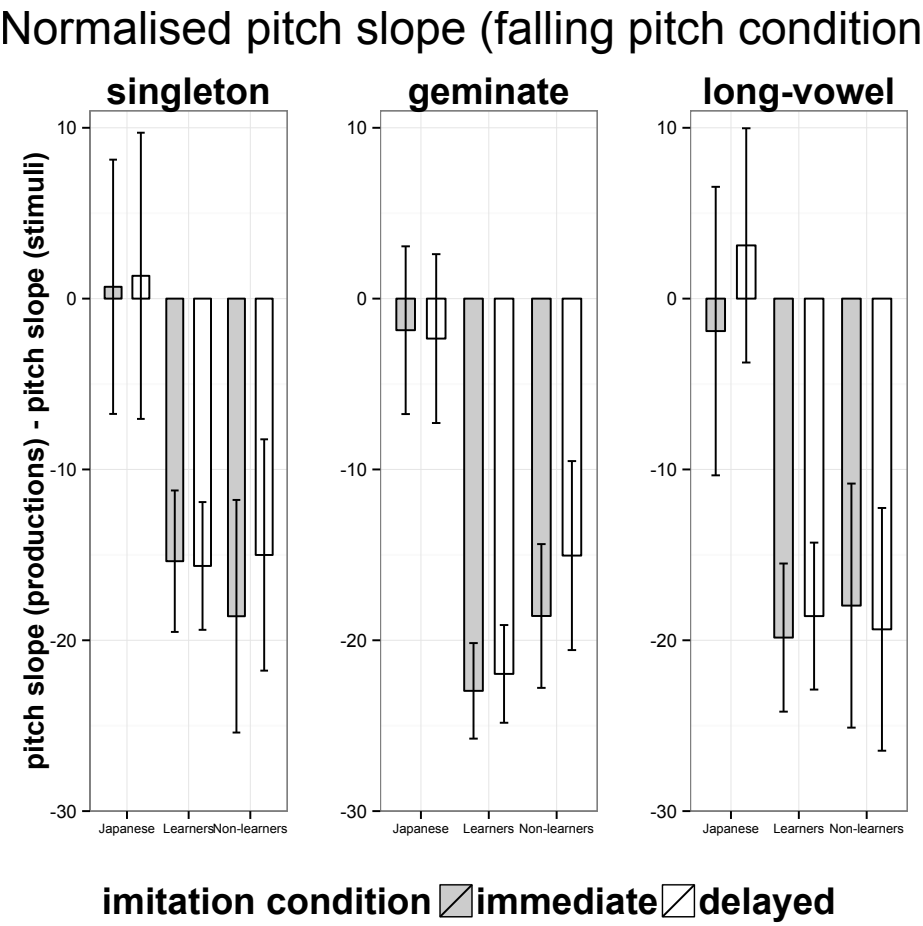

Figure 5.3 *Mean pitch slope$_{production\ -\ stimulus}$ and 95%CI bars for the immediate and delayed imitation condition and for each segmental length condition.*

5.3 Discussion

The consonant length contrast (singletons vs. geminates) was considered to be nonnative for the German participants. The pitch fall was functionally different in the Japanese and German linguistic systems (lexical vs. post-lexical/paralinguistic function). In order to analyse how accurately the participants imitated the stimuli, the duration ratios of short and long consonants were analysed in the productions of singletons and geminates compared to those of the stimuli. As for the productions of pitch, the slopes of the pitch fall were analysed in comparison to those of the stimuli.

The analysis of duration ratios showed that the Japanese speakers produced overall the same duration ratios as those of the stimuli regardless of the types of contrasts (vowel

or consonant duration ratios) and regardless of the pitch and the imitation conditions. As for the consonant duration ratios, the learners produced smaller consonant duration ratios than those of the stimuli and the non-learners even much smaller ones. The learners' performance was in this sense better than the one by the non-learners. This result can be regarded as a positive effect of exposure to the L2. However, there was no effect of *time*, suggesting that it was difficult for the German participants already in the immediate imitation task to produce the nonnative segmental length contrasts.

As for the vowel duration ratios, the Japanese produced the same duration ratios as those of the stimuli and therefore acted as an ideal reference group. Just contrary to the consonant duration ratios, the learners and non-learners produced larger ratios than those of the stimuli in both pitch conditions and imitation conditions. In the flat pitch and immediate imitation condition, the non-learners produced larger pitch ranges than the learners. In the falling pitch condition, they did not differ from each other. It is possible that the nonnative speakers exaggerated the contrasts that were familiar in their L1. And this was all the more true when naïve listeners imitated the stimuli in the lowest task demand condition (=in the immediate and in flat pitch conditions).

For the analysis of the imitation of the pitch contours, the pitch slope of the fall was measured. The results revealed that the Japanese speakers produced the pitch slope as steep as the stimuli, but both the learners and the non-learners produced flatter slopes than the stimuli. The phonetic from found in the data is in line with the claims made in previous studies that Japanese L1 speakers produce a rapid fall after the F_0 peak preceded by a flat pitch (Kondo, 2009; Sakamoto, 2010; Vance, 1987). The difference between the Japanese and German speakers was found already in the immediate imitation condition and this held true in the delayed imitation condition.

Let us recall the hypotheses stated in the introduction. As for the segmental length contrasts, the first hypothesis supported by direct realism (among others Fowler, 1986) predicted no difference between the L1 and L2 speakers' productions of the nonnative segmental length contrasts regardless of the imitation conditions. The second hypothesis supported by the theories of working memory (among others Baddeley and Hitch, 1974) and by the episodic model (among others Hintzman, 1986) claimed that the performance by the L1 and L2 speakers would not differ in the immediate imitation condition. However, the performance by the L2 speakers would decrease in the delayed imitation condition compared to that in the immediate one. These two hypotheses were not confirmed. Instead, the findings in the analyses of the segmental length contrasts support the third hypothesis that predicted a performance difference between the L1 and

L2 speakers already in the immediate imitation task and this held true in the delayed imitation task.

As for the imitation of a pitch fall, the differences in the slope produced by the L1 speakers and by the nonnative speakers were found already in the immediate imitation condition, supporting the view that they had difficulties in the L2 phonetic setting. No differences were found between the immediate and delayed imitation conditions. The first and second hypotheses were not supported by the results of this experiment. The third one was supported by the finding that the nonnative speakers' performance already differed from that of the L1 speakers already in the immediate imitation condition.

The segmental duration ratios and the slope were analysed as measurements for how accurately participants imitated stimuli with respect to the segmental length contrast and the pitch fall. However, it is not totally clear whether the measurement of these variables was sufficient to evaluate the accuracy of the imitation. One of the reasons for my doubt is that the duration ratios of singleton to geminate consonants in the three groups were in most cases larger than the ratio reported in Beckman's experiments (2.25: 1.00) when voice onset time is included, Beckman, 1982). Ratios obtained in the current study (and standard deviations in parentheses) were 3.04: 1.00 (0.98) by the Japanese speakers, 2.89: 1.00 (0.92) by the learners and 2.67: 1.00 (0.82) by the non-learners in the flat pitch condition, and 2.68: 1.00 (0.84) by the Japanese speakers, 2.62: 1.00 (0.85) by the learners and 2.40: 1.00 (0.70) by the non-learners in the falling pitch condition. However, the large ratios observed in the data can be caused by the relatively large duration ratios of the stimuli (see Subsection 5.2.1). Despite these large duration ratios obtained by the nonnative speakers' data, during data annotation, the native Japanese annotator noticed that the geminates produced by the nonnative speakers often did not sound like geminates despite these large duration ratios. This discrepancy would provide a strong motivation to include other measures in future studies. It is known that not only the duration differences of the short and long consonants, but also other factors affect the temporal structure of a word such as F_0 (Kinoshita et al., 2002; Kubozono et al., 2011). Kinoshita et al. (2002) reported that the listeners identified a vowel as a short vowel when the pitch contour was flat and as a long vowel when the pitch contour was falling. Kubozono et al. (2011) revealed that pitch was an additional cue to the perception of geminate consonants as opposed to single ones in Japanese. Moreover, previous studies reported that multiple acoustic features covary with the consonant length distinction in Japanese. In disyllabic CV(C)CV words with single and geminate consonants, preceding vowels (those before durationally contrasted consonants) are consistently longer and following vowels

(those after durationally contrasted consonants) are shorter in the context of geminates than singletons (Campbell, 1999; Fukui, 1978; Han, 1994; Hirata, 2007; Homma, 1981; Idemaru and Guion, 2008; Kawahara, 2006; Ofuka, 2003; Ofuka et al., 2005), although these duration differences are not large enough to affect the total word duration and their mora counts. Taken together, the evaluation of a "good geminate" (vs. singleton) is not a simple matter and requires perceptional evaluation. In the end, considering the fact that speech data were continuous, while the ultimate perception of length contrast is categorical (e.g. Ham, 2001), I will aim at conducting a speeded decision task in which native Japanese listeners will judge the productions of the imitation experiments intuitively. By doing so, it is possible to analyse how easily a production (e.g. a pseudoword) can be categorised as a singleton or a geminate.

To summarise, the Japanese speakers overall imitated stimuli accurately regardless of imitation conditions and of segmental length structures and pitch conditions. As for the segmental length contrasts, the nonnative segmental length contrast was difficult for them to produce already in the immediate imitation condition, while the segmental length contrast that was familiar to them was produced in an exaggerated way. As for pitch, the flatter pitch slope produced by the nonnative speakers in comparison to that of the L1 speakers suggests the difficulties in readjusting the phonetic setting.

While the consonant length contrast was nonnative for the German speakers, the pitch fall was not completely nonnative, as there is such a pitch fall in the German intonational system. The comparison between the two nonnative speakers' groups suggests that the learners showed an advantage over the non-learners when an L2 prosodic category is completely absent in the speaker's L1 (in this experiment, the consonant length contrast for the German participants). When the L2 prosodic category does exist (in this experiment, the pitch fall for the German participants), but is not functionally the same as in their L1, the learners were not advantageous over the non-learners in imitating pitch falls.

Finally, the performance differences between the Japanese and German speakers were overall found already in the immediate imitation condition both for the nonnative segmental length contrasts and for the pitch falls, indicating that phonological representations were activated already in the immediate imitation task.

CHAPTER 6

GENERAL DISCUSSION AND OUTLOOK

This dissertation has investigated possible sources of foreign accented speech by testing different stages of speech processing. In order to examine these stages separately and the stability in L1 and L2 processing in each stage, cognitive load in terms of memory load and the demand on attention control was manipulated in the experiments.

The motivation for the investigation was the deviant L2 productions shown by the German L2 learners of Japanese compared to the productions by the L1 Japanese speakers in high frequency words in Experiment 1. Possible sources for the difficulties shown in the L2 productions were examined. The L2 learners' ability to perceive, to store and to produce nonnative prosodic contrasts was examined step by step in comparison to L1 speakers and naïve speakers (= non-learners). Throughout Experiments 2 to 4, the same participants were tested and the same stimulus words were used (but with different manipulation manners). In the first subsection of this chapter (in Subsection 6.1.1), I will briefly summarise the main results of each experiment. Based on the results, I will then discuss the main topics and insights achieved from my findings. First, I will discuss the relationships between the three stages under investigation; "input", "mental representations" and "output" (in Subsection 6.1.2). After this, I will proceed with a discussion on the mental representations of F_0 and segmental length contrasts and their different status in speech processing (in Subsection 6.1.3 and 6.1.4). Moreover, I will discuss the role of cognitive load placed on working memory in the present research on L2 prosody (in Subsection 6.1.5). Finally, I will provide an outlook for future study (in Section 6.2 and 6.3).

6.1 General discussions

6.1.1 Summary of the results

The first study in Chapter 2 presented a semi-spontaneous production experiment that investigated the coordination of lexical and paralinguistic use of F_0 in the production of very frequent Japanese words (and a German word as a comparison). An experimental situation was designed in which German L2 speakers and Japanese L1 speakers had to repeat the target words while adding paralinguistic information. The results showed that the German L2 speakers varied Japanese pitch accents according to their L1 German intonational categories to convey paralinguistic meanings, while the L1 speakers did not. Moreover, the L2 speakers' organisation of each segmental duration including nasal geminates was deviant from that of the L1 speakers in all attempts. The findings are especially noteworthy as they were found in highly frequent words that the L2 speakers should have encountered very often. Therefore, they suggest that a rich amount of input of the target language did not contribute to the formation of an appropriate L2 prosody. The finding that the L2 speakers produced deviant F_0 contours regardless of the existence or absence of lexical pitch fall and even without adding paralinguistic information (= in the first attempt) led to the assumption that the L2 speakers either did not perceive the pitch information as L1 speakers do, so that they had difficulties in the "input" stage or they did not store the pitch information into the lexicon, suggesting their difficulties in the "mental representations" or they had difficulties in the articulation, in other words, in the "output" stage. In the former case, it could be possible that they had difficulties in the "input" stage, or difficulties that relate to their "mental representations". In the same way, the difficulties in organising segmental durations can relate to one or more of these stages. The findings motivated me to localise the sources of the difficulties. They were tested in the experiments presented in Chapter 3 to 5.

Experiment 2 in Chapter 3 examined the discrimination ability of nonnative segmental length contrasts. The German L2 learners of Japanese showed difficulties in producing geminate consonants in Experiment 1 possibly due to the lack of lexical consonant length contrasts in German. I examined whether German L2 learners of Japanese (and German non-learners and Japanese L1 listeners as comparison groups) were able to discriminate short and long consonant contrasts in comparison to short and long vowel contrasts. Memory load was manipulated by means of two durations of ISIs (300 ms vs. 2500 ms). The short ISI condition tested the discrimination ability of acoustic correlates and the

long ISI condition tested the discrimination ability that involves phonological representations to a greater extent. Moreover, psychoacoustic complexity was added (trials with a task-irrelevant distracting pitch fall that occurred simultaneously with the consonant vs. monotonous flat pitch) to increase the demand on attention control. I predicted a performance decrease in the long ISI condition only in both groups of German listeners whose L1 does not have lexical phonological representations of consonant length contrasts and a negative effect of the distracting pitch only in the German participants due to their difficulties in ignoring the task-irrelevant pitch movement.

The results of the d' scores and the RTs showed a native-like good performance by the learners and the non-learners only in the flat pitch and short ISI condition, in which the task demands were the lowest. In this condition, the non-learners' d' scores and the RTs were as high as the Japanese ones and the learners' d' scores were as high as the Japanese ones. Even the non-learners without exposure to the L2 could discriminate the consonant length contrasts as well as the Japanese L1 listeners and the German learners when the task demands were the lowest, simply by comparing two stimuli at the phonetic level.

However, such reliance on the phonetic comparison did not last long. Once the ISI became longer and the memory load became higher, so that the phonetic information began to be processed phonologically, the nonnative listeners' discrimination ability decreased and differed from that of the L1 listeners. In the flat pitch condition, the learners' and the non-learners' performance decreased in the long ISI condition. The learners' d' scores decreased and differed from those of the L1 listeners and the non-learners' d' scores decreased much more, so that those of the learners' and non-learners' differed from each other. Also the non-learners' RTs that did not differ from those of the L1 listeners in the short ISI condition turned out to be longer than those of the L1 listeners in the long ISI condition.

In the falling pitch condition, the performance by the Japanese L1 listeners was higher than that by the learners and the non-learners already in the short ISI condition. This held true also in the long ISI condition. The two nonnative listeners' groups did not differ from each other. Then, the results of the RTs additionally suggested the learners' indecisiveness. The comparison between the results in the flat and falling pitch conditions indicates a consistent effect of the task-irrelevant pitch on the discrimination of nonnative segmental length contrasts: Consonant length contrasts presented in the falling pitch contour were generally more difficult to discriminate for the learners and for the non-learners than those presented in the flat pitch contour. In the flat pitch condition,

the learners' and non-learners' performance did not differ from the natives' one. On the contrary, the Japanese L1 listeners did not show a difference between the two pitch conditions.

Taken together, these results suggest that the exposure to the L2 helped the learners to establish the phonological representations of the nonnative consonant length contrast than the non-learners (because the learners were affected by the increased memory load to a lesser extent). However, both the learners and the non-learners were strongly affected by the task-irrelevant pitch showing that their speech perception became more vulnerable with the higher demand on attention control. It was difficult for both the learners and the non-learners to ignore the task-irrelevant pitch and to focus their attention only on the task-relevant information. The finding indicates the difficulty to stabilise L2 processing even after starting to establish L2 phonological representations.

Chapter 4 was motivated by the finding in Experiment 1 that the German learners of Japanese ignored the Japanese lexically fixed pitch fall in their productions. The experiment in this chapter examined the discrimination of pitch contrasts. Analogue to Experiment 2, speeded discrimination tasks were conducted by means of the two different durations of ISIs. Additionally, I tested whether native and nonnative segmental length structures would affect the discrimination of pitch contrasts, an effect contrary to the one found in Experiment 2. The results of d' score analysis showed that the Japanese L1 listeners were generally more sensitive to the pitch contrasts than the learners followed by the non-learners. The RT analysis also showed shorter RTs by the Japanese L1 listeners than by the non-learners followed by the learners. The effect of the task-irrelevant segmental length structures on the discrimination of pitch contrasts was found in the nonnative listeners' RTs. First, the learners' RTs were especially long in the long-vowel condition in comparison to those in the singleton and geminate conditions. Second, the learners' and the non-learners' RTs became longer in the long ISI condition when the pairs were presented in the less native-like segmental structures. Both findings suggest that the pairs presented in the more native-like segmental structure activated their L1 phonological representations to a greater extent than the pairs presented in the less native-like segmental structures. In sum, L1 listeners of a language with lexical pitch accents established higher sensitivity to pitch contrasts associating lexical meaning which presumably led to a high discrimination ability of acoustic correlates of the pitch contrasts. The differences between the L1 and the nonnative listeners were found already in the short ISI condition, suggesting the activation of phonological representations once acoustic information comes into play. In the long ISI condition, while requiring more

phonological representations of the contrasts, only the nonnative listeners' performance decreased when the pairs were presented with more unfamiliar segmental length structures even though the segmental length structures were task-irrelevant. Both Japanese and German employ pitch contrasts, though at the different linguistic levels and this difference might be relevant for their different performance. It can be assumed that the L1 listeners of a language with a lexical pitch contrast could maintain the pitch contrast better than the L1 speakers of a language with a non-lexical pitch contrast.

Chapter 5 presented an immediate and a delayed imitation experiment that tested to what extent German learners of Japanese would show difficulties in producing nonnative segmental length contrasts and pitch falls. Following direct realism (Fowler, 1986), the immediate imitation task was set up to test the ability to imitate stimuli not necessarily mediating phonological representations. The delayed imitation task, on the contrary, was considered to require to access those representations. The accuracy of the imitation was investigated by measuring the durational ratios of short and long consonants and the slope of the pitch fall. The analysis of the segmental durations showed that the Japanese L1 speakers' consonant duration ratios did not differ from those of the stimuli, while those by the learners' and the non-learners' were smaller than the L1 speakers' ones already in the immediate imitation task. The learners performed better than the non-learners suggesting a positive L2 learning effect. On the contrary, the duration ratios of short and long vowels showed that the learners and the non-learners produced larger ratios than the Japanese L1 speakers and the stimuli. The nonnative speakers exaggerated the contrast that was familiar to their L1.

The analysis of the pitch slope revealed overall a steeper pitch slope by the Japanese L1 speakers than by the nonnative speakers. Japanese pitch fall is known to be very steep. These differences can relate to the different phonetic settings (Esling and Wong, 1983; Honikman, 1964; Laver, 1994; Mennen et al., 2010b) in Japanese and German.

It is important to remember that the same participants took part in Experiments 2 to 4 on the same day with a short pause between the experiments by being exposed to similar stimuli repeatedly. For example, in Experiments 2 and 3, the same items were presented once in each of ISI condition and in Experiment 4, they repeated the same stimuli twice, once in the immediate and once in the delayed imitation condition. Overall, where no effects were found between the short and long ISI or immediate and delayed imitation condition, one could still argue that there might have been a positive learning effect in the long ISI/delayed imitation condition which has compensated the negative effect of

memory decay. Considering this possible learning effect that could not be teased apart in the experimental paradigm used in this study, the results should be carefully interpreted.

6.1.2 The relationships between the stages of speech processing

One of the aims of the study was to localise the sources of foreign accents in prosody, investigating difficulties that L2 learners could encounter in the "input", "mental representations" and "output" stages. By using the same stimulus words and by testing the same participants, I aimed at investigating the relationships between these stages. The experiments were set up so that they differentiated these stages using two durations of ISI (300 ms vs. 2500 ms) in the perception experiments and the same two durations between the offset of the stimuli and the start of the imitation (immediate vs. delay of 2500 ms) in the production experiments. In this section, the focus is on the changes in the participants' performance due to these time variables.

First of all, the Japanese L1 listeners discriminated both segmental length and pitch contrasts equally well in all ISI conditions and their RTs were overall same . The accuracy of their imitations was constantly high both in the immediate and delayed imitation task. Thus, they performed successfully in "input", "mental representations" and "output" stages. As for the German learners' and non-learners' performance, their sensitivity to nonnative segmental length contrasts was as high as that of the Japanese L1 listeners and their RTs were as short as those shown by the Japanese listeners in the "input" stage and when the task demands were lowest. Once processing required the access of more phonological representations, the non-learners' performance in terms of their sensitivity to the contrast and RTs decreased. Concerning their discrimination ability of pitch contrasts, the learners' and the non-learners' RTs became longer in the long ISI condition, when pitch contrasts were presented in the nonnative segmental length structures. In all other conditions in the two perception experiments, no effect of ISI was found, but there were general differences among the three participant groups.

Despite the fairly good discrimination ability of nonnative segmental length contrasts by the German learners and non-learners in the perception experiment, their imitation in terms of the durational ratios of the vowel and consonant length contrasts differed from those of the Japanese L1 speakers already in the immediate imitation condition. Together with the findings of the perception experiments, it appears to be more convincing and plausible to claim that phonological representations were activated already in the short ISI conditions or in the immediate imitation task, in which phonetic processing was assumed to be involved primarily. For example, the learners' RTs in some

conditions were longer than those of the non-learners, and this difference was found regardless of the ISI condition. This finding can be regarded as evidence that the learners established L2 phonological representations and accessed their stored phonological information to aid their discrimination and they thus needed more time to select either L1 or L2 phonological representations, while the non-learners did not come across that selection process. This happened already in the short ISI condition.

In Chapter 1, I argued that the phonetic and phonological level of speech processing should not be regarded as two separate processes, and suggested to see the two processes as falling on a continuum. In this view, the short ISI condition does not completely prevent listeners from accessing phonological representations, but involves *relatively* more phonetic processing and less phonological processing. The long ISI condition, on the contrary, involves more phonological processing and less phonetic processing *in comparison to* the short ISI condition.

Recall the model of speech perception and production presented by Ramus (2001) and Szenkovits and Ramus (2005) proposing *sublexical phonological representations*, which consist of phonemes. Sublexical phonological representations are placed in short-term memory separately from *phonological lexicon*, which is a permanent long-term storage. The suggested two levels of phonological representations in the models support the claim that sublexical phonological representations may be activated already in the short ISI condition.

Strikingly enough, the effects of memory load were mostly found in the nonnative participants' discrimination ability of pitch contrasts in relation with geminates and partially with singletons, but not in relation with the long-vowel stimuli. These two former segmental length structures were considered to be more nonnative for the German participants than the long-vowel structure. This finding indicates that the nonnative segmental length structure impeded L2 perception regardless of whether the segmental length was target or nontarget. It may be explained by the fact that a pitch movement is dependent on a segmental length structure, but not vice versa. This is because a pitch movement is aligned with morae or syllables but not vice versa. Together with a bulk of the research on infants' word segmentation which shows that infants exploit a rhythmic pattern of an L1 to segment continuous speech stream into discrete units (Goyet et al., 2010; Jusczyk et al., 1999; Nazzi et al., 2006), speech rhythm appears to provide an important structural basis for the processing of prosody. This non-reciprocal dependency between segmental length and pitch will be further discussed in the next subsection.

To summarise, the findings of the series of experiments clearly showed difficulties relating to the nonnative listeners' "mental representations". Additionally, the imitation of a pitch fall indicated that the nonnative speakers had difficulties in the "output" stage in terms of articulation. These two stages contributing to foreign accents are identical to those found in Sakamoto (2010). The novel finding in my thesis is that the "input" stage was found to be easily impeded by the increased cognitive load both in terms of the increased memory load and the higher demand on attention control. The high discrimination ability of acoustic correlates of nonnative prosodic contrasts found in my thesis turned out to be vulnerable, easily affected by other distracting conditions.

6.1.3 F_0 and segmental length contrasts

One of the novel aspects of this thesis was to investigate two types of prosodic contrasts within one study; nonnative segmental length contrasts and pitch contrasts. Both types of contrasts are absent in the German lexicon. However, pitch contrasts exist in the German intonation system. The production experiments (Experiment 1 and Experiment 4) tested the two prosodic properties together within an utterance/nonce-words and the perception experiments examined the opposite process, selective attention to differences in one of the properties. To test this in the discrimination tasks, one of the two prosodic cues was used as a task-irrelevant one for the discrimination of the other cue. In this section, I will propose two different underlying mechanisms for the processing of segmental length and pitch in "input" and "mental representations" and discuss the non-reciprocal effect of the task-irrelevant pitch on the discrimination of segmental length contrasts and the effect of the task-irrelevant segmental length structure on the discrimination of pitch contrasts.

First, given that the phonetic details of contrasts were still available in the short ISI condition, it was assumed that nonnative listeners would exploit the phonetic information to successfully discriminate nonnative prosodic contrasts in that condition. Such an advantage due to the phonetic information in the short ISI condition compared to the long ISI condition was found only in discriminating nonnative segmental length contrasts, but not pitch contrasts: Experiment 2 that tested the discrimination ability of nonnative segmental length contrasts showed that even the learners and the non-learners reached d' scores as high as those of the Japanese in the flat pitch and the short ISI condition. However, their d' scores decreased in the long ISI condition. Experiment 3 that tested the discrimination ability of pitch contrasts showed that the learners and non-learners reached lower d' scores than those of the Japanese already in the short ISI con-

dition. These divergent findings between the segmental length and pitch contrasts pose the question, why the phonetic advantage was not found for the German participants to discriminate pitch contrasts when the phonetic details of the contrasts were still available.

One could also assume that the acoustic correlates of the segmental length contrasts were just more perceptually "extreme" than the pitch contrasts for the German listeners. However, the acoustic measurement of the stimuli showed that the pitch falls had in average over 16 semitones, which was a large pitch range. Previous studies showed that pitch ranges of declarative sentences in German were around 9 semitones (8.1 semitones with a sd = 2.3 semitones for female speakers and 9.7 semitones with a sd = 2.9 semitones for male speakers in Zimmerer et al., 2015 and 7.7 semitones with a sd = 2.6 semitones for female speakers and 10.2 semitones with a sd = 0.57 semitones for male speakers in Brinckmann and Benzmüller, 1999). Though it is difficult to compare the perceived salience of the segmental length contrasts with that of the pitch contrasts, one could assume that both contrasts were perceptually salient enough to obtain the phonetic effect. Further, there are other possible explanations:

It may be the case that a listener is phonetically more sensitive to the acoustic correlates of a contrast of a prosodic cue that has the least linguistic constraint in one's L1 (Hallé et al., 2004). German L1 listeners can be assumed to be phonetically more sensitive to nonnative segmental length contrasts than to pitch contrasts, because a pitch contrast is linguistically (post-lexically and paralinguistically) used in German, while a consonant length contrast is not. Since consonant length contrasts are not present in the mental lexicon of the German listeners, they were phonetically more sensitive to acoustic correlates of consonant length contrasts. Pitch contrasts convey linguistically meaningful contrasts in German, and therefore the German listeners may have been less sensitive to acoustic correlates of pitch contrasts in comparison to those of consonant length contrasts, because the absence of the phonological contrast may enhance the listener's sensitivity to the acoustic correlates of a stimulus. Hallé et al. (2004) that tested a cross-linguistic Chinese tone discrimination (of /pa/, /pi/ and /kwo/ produced in the four Chinese tones) reported that French L1 listeners had a psychoacoustic advantage in discriminating pitch contrasts over English L1 listeners and even a more considerable one over L1 listeners of tone languages, because French listeners are constrained the least by F_0 in their L1 among these three groups. Mandarin Chinese employs lexical F_0 and English uses F_0 to contrast lexical stress and post-lexical information (Beckman, 1986). French uses F_0 to mark segmental and prosodic boundaries (Delattre, 1951) and French intonation is char-

acterised by a sequence of rising pitch movements, which limits a dynamic use of F_0 to convey post-lexical and paralinguistic information as English does.

Another possible interpretation for the asymmetric findings for the discrimination of segmental length contrasts and pitch contrasts is supported by the claim that tone language listeners established fine-grained associations between pitch contours and lexical meaning and this presumably leads to a generally high ability in language-unspecific pitch processing (Deutsch et al., 2006; Pfordresher and Brown, 2009). This could be true also for Japanese L1 listeners, whose L1 employs lexical pitch contrasts. Based on this claim, a listener may be phonetically more sensitive when a prosodic cue conveys lexical contrasts than when it does not. Applying this claim to German, German listeners may be more sensitive to segmental length contrasts than to pitch contrasts for the following reasons: Although pitch is used contrastively in German, pitch contrasts are not associated with lexicon and intonation patterns do not relate to any specific stressed syllables of a word, so that pitch information may be not stored together with a word. As for segmental length contrasts, German listeners may have mental representations of segmental length contrast assigned to specific words in their L1 (e.g. *Stadt* vs. *Staat*). This explains why the German listeners were phonetically more sensitive to the segmental length contrasts than to the pitch contrast in my experiments.

Furthermore, I examined whether a task-irrelevant prosodic cue affected the discrimination of the other prosodic cue and whether the effects were reciprocal. The results of Experiment 2 showed overall a strong effect of the task-irrelevant pitch on the discrimination of the nonnative segmental length contrast. The nonnative listeners' d' scores and the RTs were constantly affected by the task-irrelevant pitch, indicating that the Japanese lexical pitch movement (e.g. a falling lexical pitch accent or an initial low, cf. Haraguchi, 1977) makes the perception of the nonnative consonant length contrast more difficult when it occurs simultaneously. As for the contrary effects, an effect of the task-irrelevant segmental length structures on the discrimination of the pitch contrast was found only in the learners' and non-learners' RTs. The learners' RTs were longer for the stimulus pairs presented in the more German-like segmental length structure than for the stimulus pairs presented in the less German-like segmental length structure. In the long ISI condition, both the learners and the non-learners slowed down to a greater extent in the less German-like segmental length condition than in the more German-like segmental length condition. The task-irrelevant segmental length structure did not affect the sensitivity to the pitch contrasts itself (= d' scores), but the task difficulties (= RTs). In Chapter 4, I argued that the more German-like segmental length structure of the stimuli activated

the nonnative listeners' L1 phonological representations to a greater extent than the less German-like segmental length structures and this probably led to differences in their RTs.

I postulate that the effects of the task-irrelevant prosodic dimensions on the discrimination of the task relevant prosodic contrasts related to what extent words in the mental lexicon were activated. The consistent effects of the task-irrelevant prosodic dimension found in both the d' scores and the RTs for the segmental length contrasts are evidence that the discrimination of segmental length contrasts activated words in nonnative listeners' mental lexicon to the greatest extent, followed by the discrimination of pitch contrasts presented in the long-vowel structure, and finally by the discrimination of pitch contrasts presented in the singleton and geminate structure. As discussed before, the pitch fall presented in the long-vowel structure was phonologically more natural to the German participants than the one presented in the singleton or geminate structure. This led to the greater extent of word activation in the mental lexicon.

Further, there are several arguments to why I hypothesise that the attention to the target segmental length contrasts activated words in the mental lexicon to a greater extent than the attention to the target pitch contrasts did. In order to discriminate segmental length contrasts, participants paid attention to that target cue. While paying attention to segmental length contrasts, segmental features (Hall, 2007) to identify vowels and consonants were inevitably processed together. By activating the vowels and consonants of the stimuli, words in the mental lexicon were activated. Note that pseudo-words were used as stimuli, so that the activation of words in the mental lexicon could have taken place to a lesser degree than when using real words. What is more, it was an initial activation without a final selection of words. On the contrary, pitch is not used for the lexical distinction in German, but to code paralinguistic information such as the attitude and emotion of a speaker. Therefore, it can be hypothesised that the German listeners did not activate words while paying attention to the pitch contrasts so much as while paying attention to the segmental length contrasts. As mentioned above, pitch patterns are assumed not to be stored together with words (= there may be no one-to-one relationships between words and pitch patterns) in the German listeners' mental representations. Another argument would be that to discriminate pitch contrasts, segmental features can be ignored much more easily than when discriminating segmental length contrasts. The stimuli could have been replaced with sine wave sounds or Gaussian noise or with any other tonal sounds like music in order to examine the pitch contrasts. However, to examine segmental length contrasts, vowels and consonants are necessary to build vowel and

consonant length contrasts. Otherwise it would have become a discrimination task for rhythmic contrasts and not for vowel and consonant length contrasts.

To conclude, different underlying processing mechanisms for the different prosodic cues (e.g. F_0 and segmental length) were discussed in this section. Based on such underlying processing differences, "prosody" should not be regarded as a collective (umbrella) term and treated as such, as if there would exist a general processing of "prosody", but should rather be investigated separately when talking about e.g. segmental length and pitch. The discussion further highlighted the uniqueness of prosody; the same cue is used multiply at different linguistic levels (at lexical, post-lexical and paralinguistic levels) in different ways in each language. Indeed, lateralisation and neuronal representations of the same prosodic cue differ depending on one's L1 and L2s (e.g. Barry, 1981; Näätänen et al., 2007; Tamaoka et al., 2014; Zhao et al., 2011).

6.1.4 Lexical, post-lexical and paralinguistic prosody (F_0)

The experiments exploited the fact that Japanese and German employed the same prosodic cue at different linguistic levels. In Experiment 1, the German L2 learners of Japanese and the Japanese L1 speakers were exposed to an experimental situation in which lexical and paralinguistic use of F_0 had to be coordinated simultaneously. The productions of the Japanese L1 speakers showed that the lexical use of F_0 was maintained and therefore the realisation of paralinguistic use of F_0 was restricted. Lexical use of F_0 outweighed the paralinguistic one. This priority of lexical prosody in the processing order is also shown in Levelt's prosody generator (see Chapter 2). The lexical use of prosody is assigned to metrical and segmental spellout, before global prosody is generated in *Prosody Generator*, suggesting that the lexical use of prosody is taken into the processing before the paralinguistic use of prosody is considered, and that the paralinguistic use of prosody will be generated in a way allowing no interference with its lexical use. As for speech perception, Experiment 3 tested whether a post-lexical use of F_0 in an L1 would help a listener to discriminate a pitch contrast in comparison to the listeners whose L1 exhibits a lexical use of F_0. The results suggested that the post-lexical use of F_0 was not advantageous to discriminate the pitch contrast. Listeners of a language that employs F_0 lexically were more sensitive to the pitch contrast than listeners of a language with the post-lexical use of F_0.

6.1.5 Cognitive load

Cognitive load was considered to be one of the keys to understand the difference between L2 and L1 processing. In Experiments 2 and 3, two aspects of cognitive load were manipulated. One was memory load and the other was attention control. Memory load relates to *the phonological loop* in working memory (e.g. Baddeley and Hitch, 1974), which is one of the slave systems of *the central executive*, to which attention control relates. In Experiment 2, a clear effect of memory load was found in the flat pitch condition, namely when the processing only dealt with length contrast without the distracting pitch movement. Once pitch movement came into play, the effect of ISI vanished. In the falling pitch condition, the effect of memory load was not found, because the sensitivity by the nonnative listeners was generally much lower than in the flat pitch condition, which can be regarded as a floor effect. In Experiment 3, an effect of memory load was not observed in the d' score analysis, but in the RT analysis. Moreover, the effect of memory load was found only together with an effect of attention control: The RTs became longer for the stimulus pairs that required higher demand on attention control once the required memory load increased. The listeners whose L1 exhibits lexical pitch contrasts were more advantaged than the listeners whose L1 does not. This nonnative listeners' performance decrease indicates that the higher demand on attention control impeded successful speech processing under the higher memory load, suggesting a more dominant influence of *the central executive* than *the phonological loop* while processing speech stimuli.

Taken together, in the current experiments, the effect of the demand on attention control was stronger than that of memory load. This can be discussed on the basis of the hierarchical relationships of the main components in working memory (e.g. Baddeley, 2010; Baddeley and Hitch, 1974) (see Figure 1.6 in Section 1.6). The stronger influence of *the central executive* than *the phonological loop* indicates that successful processing in the subordinated system in working memory requires successful processing in *the central executive*, but not vice versa. This non-reciprocal dependency between the two systems in working memory can be predicted from the structure of the model (e.g. Baddeley, 2010; Baddeley and Hitch, 1974). Beyond the experimental situation, my finding suggests that learners, who were establishing L2 phonological representations, still did not fully succeed in automatising the L2 speech processing, because more stable attention control in *the central executive* presupposes that the speech is processed more automatically (McLaughlin et al., 1983).

Speeded AX discrimination tasks and imitation tasks were conducted, which required relatively low task demands and that therefore enabled us to increase and better control the task demands of the experiments. The tasks with low demands also enabled us to find that the processing of L2 processing is vulnerable. Such vulnerability would have been otherwise difficult to reveal had task demands been generally higher. The vulnerability and unstable L2 processing affected by various task demands is an issue that is still rarely considered in current L2 perception and production models. If at all, some studies investigated differences in the perception of auditory, phonetic and phonological information in nonnative speech (Best et al., 2001; Best and Tyler, 2007).

The task demands controlled in the present study may be translated into various distracting factors in speech perception in more natural settings. Therefore, the performance differences between L1 and L2 listeners and the performance decrease once task demands became higher suggest the instability of the L2 speech perception under various distracting factors in daily communication.

6.1.6 Definition of (di)simmilarities of a cross-linguistic prosodic contrast

As discussed in Chapter 1, it seems to be necessary to distinguish the dimensions along which perceived (di)simmilarities of a cross-linguistic prosodic contrasts are defined. PAM-L2 (So, 2010; So and Best, 2011, 2014) predicts that the existence of the same phonological category in one's L1 and L2 will help listeners to successfully map an L2 prosodic category onto an L1 category, regardless of at which linguistic level the prosodic cue builds a meaningful contrast in the L1 and L2. However, in Experiment 4, it was found that the pitch contrast that exists in the German intonational system did not help the nonnative listeners to map the Japanese pitch contrast fully, either due to the cross-linguistic phonetic differences of the same phonological category or due to the different linguistic levels at which the prosodic cue exhibits meaningful contrasts in one's L1 and L2. This finding suggests that multiple dimensions should be taken into account to define cross-linguistic (dis)similarities in perception.

In the L2 Intonation Learning theory (Mennen, 2015), Mennen proposed four dimensions that characterise intonation along with similarities and differences between L1 and L2 (ibid., 173):

1. The inventory and distribution of categorical phonological elements (= systemic dimension)

2. The phonetic implementation of these categorical elements (= realisational dimension)

3. The functionality of the categorical elements or tunes (= semantic dimension)

4. The frequency of use of the categorical elements (= frequency dimension)

Let's take the Japanese and German pitch accents as examples for each of the dimensions proposed by Mennen (2015). The first dimension refers to typological similarities or differences in the inventory of structural phonological elements (such as pitch accents, accentual phrases, prosodic words and boundary phenomena) (Mennen, 2015, 174).

For example, the only Japanese pitch accent form is a falling one, while there are both falling and rising pitch accent forms in German. As for boundary tones, there are both falling and rising boundary tones in both languages. Thus, the phonological inventory of boundary tones is similar between the two languages than the one of pitch accents. The second dimension refers to cross-language similarity/dissimilarity in how the systemic prosodic elements are phonetically implemented or realised. As mentioned before, phonetic realisations of a Japanese falling pitch accent differ from those of a German falling pitch accent. As for the third dimension, Mennen (2015) explained that it concerns the use of structural elements or tunes for conveying meaning. While a pitch accent realised with a drastic pitch fall is attitudinally neutral in Japanese, the one in German conveys attitudinal meanings such as frustration or anger (e.g. Gibbon, 1998).

Finally, the fourth dimension refers to cross-language similarities and differences in the frequency of use of the prosodic inventory. More frequently used varieties are expected to be easier to acquire than those that are used less frequently (Gass and Mackey, 2002). To the four dimensions presented by Mennen (2015), I will propose to add the fifth dimension;

5. The linguistic level of categorical phonological elements (= linguistic level dimension)

As discussed before, the same prosodic cue can convey meanings cross-linguistically at different linguistic levels. While Japanese pitch accents primarily convey lexical meanings, German pitch accents do not, but post-lexical or paralinguistic meanings. The linguistic levels at which a prosodic cue plays a role can influence the difficulty of acquisition of that cue.

Finally, Experiment 2 revealed that the processing of L2 processing is vulnerable under increased task demands (due to the increased memory load and demand on attention control). Such variabilities of L2 perception under different task demands and cognitive load that are required in L2 perception have been rarely taken into account in the existing models up to date (with a notable exception in Best et al., 2001 that proposed to substantiate differences in the perception of auditory, phonetic and phonological information in nonnative speech). The findings of this thesis show that perceived similarities for nonnative listeners cannot be simply defined by means of a cross-linguistic comparison of phonological categories, but they depend on task demands under which two sounds are presented.

6.2 Outlook

This dissertation has produced several results that invite further investigation. First, the learners generally performed better than the non-learners showing a positive learning effect of the L2. However, RT analyses in the perception experiments showed that the learners took longer for the responses than the non-learners. I postulated that the learners and the non-learners applied different strategies to complete the tasks. Differences in the underlying mechanisms for the non-learners and the learners need further attention in future experiments. Second, in order to pinpoint the sources of the difficulties found in the semi-spontaneous production experiment, I traced speech perception and production back to their bases by conducting AX discrimination tasks and imitation tasks. Little is still known however as to how L1 and L2 listeners and speakers may appropriately integrate the prosody of words at higher linguistic levels such as at the utterance level. Prosody at these higher linguistic levels is what is normally required in daily life speech processing. I will therefore examine the integration of F_0 and segmental length in L2 perception and production at the word level and at the short-sentence level. Third, after investigating prosody at the utterance level to understand better the processing of prosody required in daily life communication, I will aim at providing didactic implications from the findings of my thesis and evaluate teaching methods on L2 prosody using a pretest-posttest paradigm (e.g. Collentine, 1998; Dimitrov and Rumrill, 2003). The transfer from scientific experiments to actual classroom settings is one of the important contributions of a study on L2, since didactic methods in this area are far from being developed sufficiently.

Finally, other analysis methods for L2 production data can be further investigated. The production data in Experiment 1 were analysed by means of a manual annotation. However, L2 data were found to be more difficult to annotate than L1 data due to the variability of the L2 data. The production data in Experiment 4 were analysed by measuring segmental durations and the local pitch slope. However, it was questionable whether the imitation accuracy could be reduced to such local phonetic measurements. One of the possibilities to deal with these difficulties and problems would be to include a perceptional evaluation through L1 listeners' judgement. Speech data are continuous, while the perception of e.g. length contrast is categorical (e.g. Ham, 2001). Therefore, as future work, the production data of this study can be evaluated by Japanese L1 listeners using a speeded decision task. Besides the judgement responses, RTs in such a speeded decision task can provide us information as to how certain the listeners were to give the response.

Another meaningful investigation would be to analyse the data using data-driven (semi-)automatic methods such as Functional Principal Component Analysis (FPCA, Ramsay and Silverman, 2005) or Self Organising Maps (SOM, Kohonen, 2001). Such semi-automatic analysis methods do not require annotators to categorise an utterance subjectively and do not reduce the dynamic variation to static statistical indexes or to select discrete points to be measured subjectively. For example, the FPCA makes it possible to run statistics on a whole curve. In this way, there is no risk to destroy essential information in global curves through subjective decisions.

6.3 An exploratory example: data-driven analyses of F_0

As a part of future work, I exploratorily applied the aforementioned two data-driven analysis methods (FPCA and SOM) to the data obtained in Experiment 1.

Up to date, there has been a number of attempts to apply data-driven (semi)automatic analysis methods to F_0. For example, polynomial equations are a way to describe curves including F_0 contours in a mathematical expression using variables and constants (Grabe et al., 2007). A fixed "dictionary" of orthogonal shapes (horizontal line, sloped line, parabola and cubic or "wave") is used to define the shapes of F_0. This method solves the problem well for short curves, while longer ones may need a more flexible set of basic shapes. Other approaches used in speech technology to fit F_0 curves include the Fujisaki model (Fujisaki, 1992), MOMEL (Hirst and Espesser, 1993) or Tilt (Taylor, 2000).

While not proposing such data-driven semi-automatic data analysis methods as a full substitute of the annotation analysis, one can take the opportunity to investigate the

pros and cons of different approaches to the same set of data so that a methodological triangle enables us to understand the nature of the data better. By all means, it would be desirable and important to have an additional way of analysing L2 data for the reliability of the analysis. In this thesis, I explanatorily applied the FPCA and the SOM to the data obtained in Experiment 1 (N = 90). Before investigating larger data points of Experiment 4 (N = over 7000) extensively, it is reasonable to evaluate the novel methods with smaller data set as the first step. In the following, the results are briefly presented. Technical descriptions and exposition are limited to the concepts and steps that are required in order to follow the results. Further details on the procedure described in the following are found in Appendix A as well as in the given literature.

6.3.1 FPCA

FPCA, an extension of ordinary PCA, produces a parametrisation of the set of input curves or trajectories including F_0 contours in terms of a small set of *Principal Components* (PC1, PC2, etc.), each one capturing a different and independent shape variation found in the data. The output of the FPCA is both numerical and graphical, where the former allows further statistical investigations (e.g. hypothesis tests), while the latter allows a graphical interpretation of results. Since the FPCA is purely data-driven and is neither based on a prosodic model nor on the perception of a human annotator, it provides a complementary perspective to the analysis based on the annotation using a ToBI system. Additionally, the FPCA makes it possible to bring together the analysis of the shape of F_0 contours and that of segment durations in one joint analysis (the procedure proposed by Gubian et al., 2011).

Procedure

In this thesis I follow the method proposed by Gubian et al. (2011) for the automatic analysis of F_0 contours. This is an FPCA-based procedure that has been adapted to the specific needs and constraints characterising the study of speech prosody. Two main concepts are introduced in this section forming the essence of this method, namely how the F_0 signal is represented and how it is analysed.

The input to the FPCA consists of F_0 contours and of temporal positions of segments (syllables or morae). In order to minimise gender effects, F_0 values were expressed in normalised semitones according to the formula;

$$F_0(t_i)[\text{norm. st}] = 12\log_2 F_0(t_i)[\text{Hz}] - \frac{1}{n}\sum_{i=1}^{n} 12\log_2 F_0(t_i)[\text{Hz}], \tag{6.1}$$

where t_i's and n are the time positions of F_0 samples and the total number of samples for a specific contour, respectively. Some adjustments had to be done in order to accommodate to the absence of F_0 signal during voiceless sounds, since the FPCA represents contours as continuous curves without missing values. For this reason, the first fricative /s/ in *sumimasen* was eliminated from the analysed time interval, and internal boundaries were placed as follows: | ~~s~~| mi | ma | se | n| (a strike-through line indicates the cut phone). The missing F_0 values for the second /s/ were linearly interpolated F_0 values. For the word *Entschuldigung*, segmented as | ~~Ent.sch~~ ul | di | gung| , the first syllable /ent/ was eliminated because some participants did not produce it. For the same reason as above, the initial fricative was not taken into account.

By performing a first operation called *smoothing*, each of the extracted F_0 contours is interpolated by a continuous smooth curve represented by a mathematical function (Ramsay and Silverman, 2005). A second transformation, called *landmark registration*, aligns the curves with the syllabic and moraic boundaries. The purpose of this operation is to obtain an analysis of F_0 contours that is synchronised to those boundaries, so that it is possible to locate e.g. a pitch rise or fall within certain syllables or morae. This step is equivalent to a time normalisation carried out on each segment separately, but the operation is global and smooth. Contrary to most traditional approaches to time normalisation, where the analysis of normalised contours and of segment duration occur separately, FPCA allows to carry out a joint analysis. This is because the difference between original and normalised durations are internally represented by smooth time warping curves, which paired to their corresponding normalised F_0 curves produce a complete representation of the original F_0 contours. This representation is the input to the FPCA, the actual statistical analysis tool.

Each input curve is described in terms of *PC scores*, which quantify how much each PC is applicable to that particular curve. For example, PC1 may capture the variation in height of a peak in the curves, while PC2 captures its shift in time. Then each curve will be parametrised in terms of a PC1 score determining the height of that peak and a PC2 score determining its position. The crucial difference between implementing this

parametrisation manually and applying the FPCA is that the latter finds the principal shape traits automatically from the data itself, while in the case of ad-hoc parametrisations, e.g. in terms of peak heights, slope etc., the investigator has to identify the shape traits of interests and to implement a convenient quantitative representation. Once the FPCA is completed, the analysis terminates by using PC scores as variables in an ordinary statistic (or statistic model).

Results

The FPCA procedure (see details in Appendix B) was applied on the 90 F_0 contours from *sumimasen*. I considered only the first PC that explained 60.4% of the variance in their respective FPCA model. The amount of variance explained by a functional PC reflects the "importance" of a variation measured across the entire time interval under study. Figure 6.1 shows mean PC1 scores with 95% CIs.

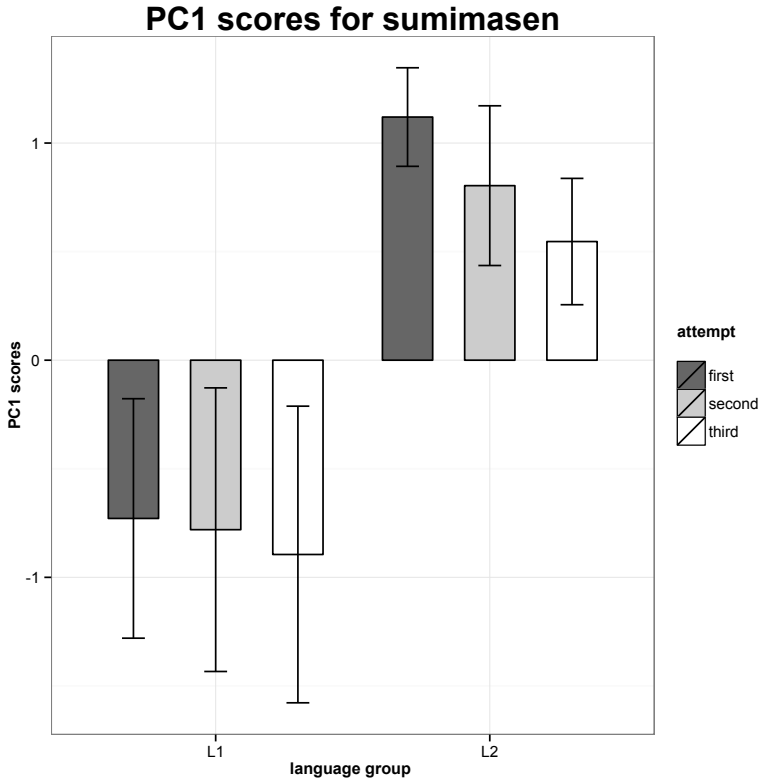

Figure 6.1 *Mean PC1 scores and 95% CI error bars for each language group and for each attempt.*

The plots show a large difference between the L1 and L2 speakers' PC1 scores and an interaction between language groups and the number of the attempt. The L2 speakers' PC1 scores decreased in the repetitions, while those of the L1 speakers did not change across the number of the attempt. In order to relate scores to the corresponding contour shapes, Figure 6.2 shows the panel that translates the PC1 score values into effects on the shape of time-normalised F_0 contours by applying Equation (1) to PC1 scores. The three lines in the plot correspond to a PC1 score value of -1, 0 and 1 in Figure 6.1. Figure 6.2 shows that PC1 modulates a variation that mainly affects the rightmost part of the curves, corresponding to the morae /se/ and /n/. The comparison between Figure 6.1 and Figure 6.2 shows that the L1 speakers constantly produced F_0 contours with a drastic pitch fall, while the L2 speakers produced flat F_0 contours in the first attempt and more falling F_0 contours in the third attempt.The L1 speakers generated curves that exhibit a steep falling accent in all attempts, while this phonetic characteristic of the Japanese pitch accent was not found in the L2 data.

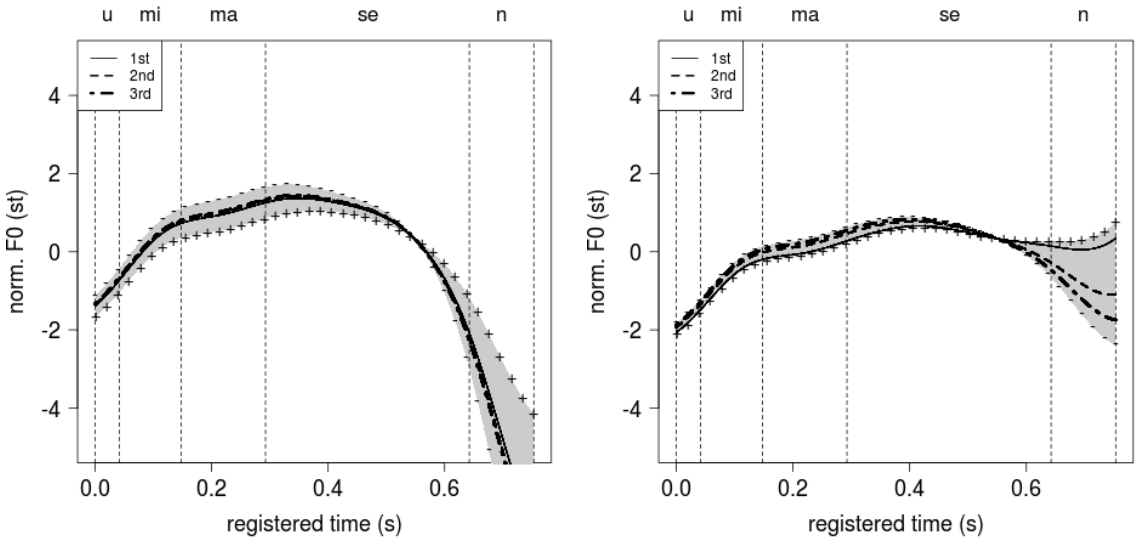

Figure 6.2 *Time-normalised F_0 curves for the L1 speakers (left) and for the L2 speakers (right) that correspond to the PC1 scores.*

The outcome visually confirmed that the variations of the L1 speakers' F_0 curves were situated within the category of a falling boundary tone, while the variations of the L2 speakers might lie across phonological categories (falling, flat and rising boundary

tones). This observation suggests the limitation of a data-driven analysis that does not account for perceptual categorical boundaries of the contours.

6.3.2 SOM

SOMs, also known as Kohonen Maps, are a well established Machine Learning technique that can be used for clustering or as a classifier based on feature vectors (for further details, see Kohonen, 2001). I made use of this machine learning for the visual analysis of F_0 contours finding similar clusters within the data based on vectors (= F_0 contours). While the FPCA abstracted the real underlying data, the SOM visualises original contours.

Procedure

The system consists of three components. The *Data Input* where all input files are read and converted into the internal data model. In this case, the smoothed and landmark-registered F_0 contours that formed also the input for the FPCA, the corresponding time vectors, segmental boundaries from the manual annotation done in Experiment 1, and meta data such as participants' information including the language group (L1 or L2), the number of the attempt (1, 2, 3) were put into the system as input. The second part covers *Machine Learning*, in which the SOM algorithm is carried out. Here the level of detail can be changed interactively by selecting the number of desired cells or by removing data in every iteration. After some attempts with different numbers of cells, I set up 3 by 4 as the number of the cells, because this number of cells seemed to present the 90 data capturing the main characters of the F_0 contours. The visualisation based on the SOM result is realised within the last component *Training and Interactive Visualisations*, between which the analysis loop may go back and forth. This iterative process refines the analysis in the loop, in which the SOM-result can be manipulated and serves as an input for the next iteration. For example, it is possible to delete cells directly on the grid. Another interaction is to move cells and to pin them afterwards, making this cell fixed for the next SOM-training.

Based on the visualisation, one is able to visually perceive and understand the data comparing other sources of results (such as of a manual annotation). This combination of human knowledge and reasoning with automated computational processing is the key idea underlying visual analytics (Thomas and Cook, 2006).

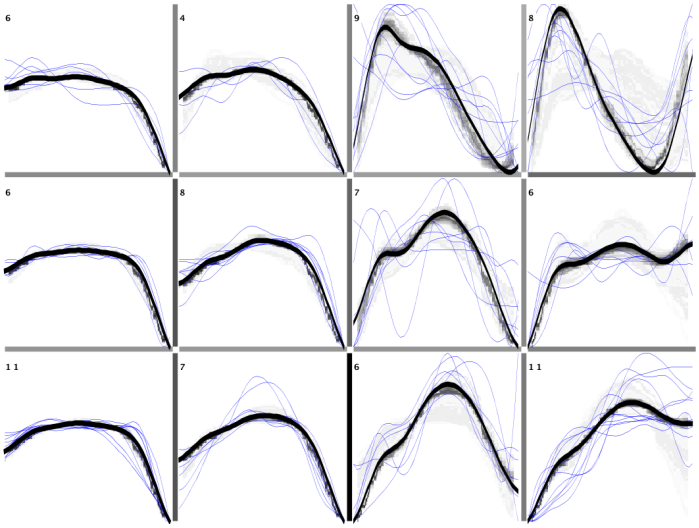

Figure 6.3 *12 clusters found in the first run of SOM including all data. Thin lines represent the F_0 contours within the cluster. The number in each cell is the number of the contour of that cluster (hereafter). Black borders between the cells show the cluster (dis)similarity (thicker, more similar).*

Results

Figure 6.3 shows the distribution of all data including L1 and L2 ones. The number in each cell presents the number of the contours in that cluster. As the next step, the cells were coloured according to the speaker groups, see Figure 6.4. Clusters were coloured according to each speakers group. The colour saturation represents the number of speakers in a cell, relatively to each speakers group. The left heatmap shows that most Japanese L1 speakers' contours are situated in the bottom left corner, in which falling contours with a steep pitch fall are presented. The right heatmap for the L2 speakers shows another type of distribution. Most German L2 speakers are situated in the right half of the heatmap in which both falling and rising contours are presented. Figure 6.5 presents the most frequent three clusters for each language group from the heatmaps in Figure 6.4. Figure 6.5 makes it clear that the Japanese produced homogeneous clusters, while the Germans did not.

Moreover, in order to investigate how the German L2 speakers varied their contours in the repetitions, the Japanese L1 speakers' data were removed and the second run of SOM only for the German speakers' data was undertaken. Figure 6.6 shows the clusters according to the number of the attempt. As a tendency, more rising contours were found

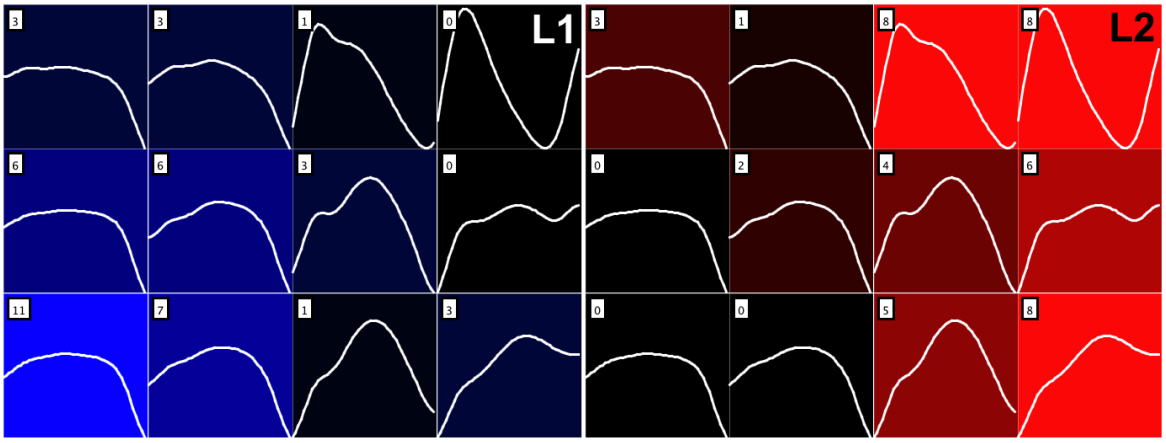

Figure 6.4 *Heatmaps for L1 (left) and L2 (right) speakers.*

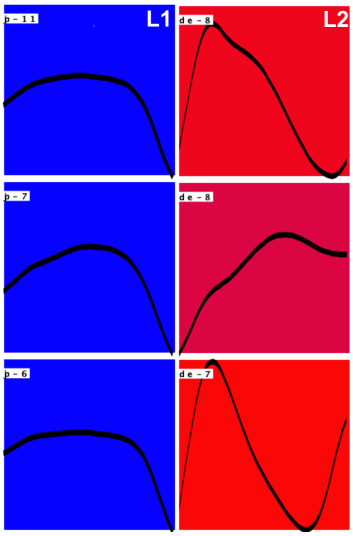

Figure 6.5 *The most frequent three clusters for each speakers group (L1 speakers in left and L2 speakers in right). The output is based on the first run of SOM.*

in the first attempt, in contrast to more falling contours in the third attempt. However, this tendency is difficult to generalise due to the small number of the data.

6.3.3 Evaluation of the methods

Three analysis methods (manual annotation presented in Chapter 2, the FPCA and the SOM) were employed for the same data. All three methods showed the common essen-

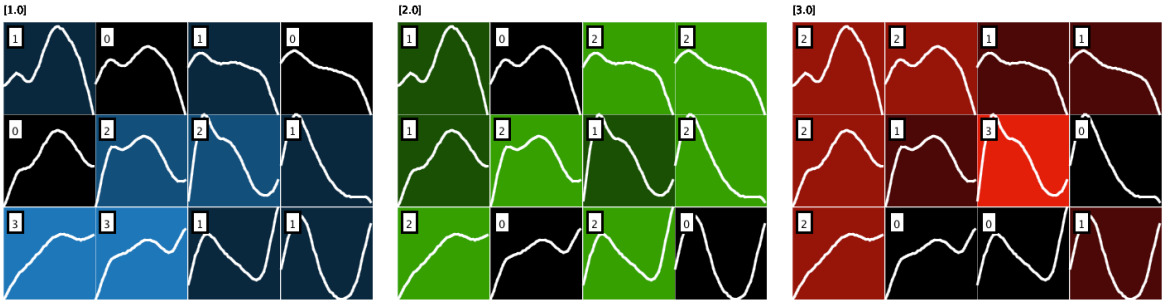

Figure 6.6 *Heatmaps for each number of attempt. Only German L2 speakers' data are included. Clusters were coloured according to the the number of the contours of that cell relatively to each attempt.*

tial findings that the Japanese L1 speakers were faithful to the lexical restriction of F_0 and did not vary F_0 contours to convey the paralinguistic information, while the German L2 speakers varied F_0 contours. Moreover, they all showed that German L2 speakers produced rising F_0 contours more frequently in the first attempt than in the third attempt. In the third attempt, more falling contours were found than in the first attempt. The FPCA additionally showed that the L2 speakers did not produce the typical Japanese pitch accent with a flat pitch followed by a steep pitch fall. This was the most influential difference between the groups. Moreover, the FPCA provided numerical distances between each contour and the statistics on which it could be run. The additional value provided by the SOM was that it categorised the contours without any prior knowledge. Moreover, it was useful to get an overview of a large data set visually. For example, this makes it possible to easily detect deviant contours or data containing errors. Furthermore, the visualised SOM-results were easier to understand than other statistical analyses or projections (such as those of the FPCA). Without opening the door to subjective decisions in the annotation of F_0 contours, it was possible to see the clear differences between L1 and L2 productions. Finally, it is worth mentioning that the SOM provided an interactive and a flexible way for the analysis of prosodic data as opposed to other systems. For example, it was easy to filter the data (e.g. in order to examine only a subset of the data) or to adjust the grid size (see examples in Sasha et al., 2015). The disadvantages of such data-driven method analyses are that they did not take categorical thresholds of auditory perception into account, so that subjective visual decisions at sight were required at the end. As for the SOM, it did not provide information whether two contours in the neighbouring cells auditory belong to a same category. To assist the categorisation based on auditory

perception, the SOM can visualise the numerical distances between the cells/groups as the thickness of the borders between the cells. Orientating the different thickness of the borders, one can distinguish similar from less similar cells. Also, the distance of the cells within the grid reflects their distance. In addition, some approaches (e.g. Schreck, 2010) apply a second clustering method on the SOM result delivering the information about which cells are similar.

Both manual annotation and data-driven methods offer advantages and disadvantages while they complement each other. On the one hand, the manual annotation corresponds more to our categorical interpretation of F_0 contours. The fully automatic analyses using the FPCA or the SOM ignore this aspect. In the case of this study, it would have been difficult to conclude that the German L2 speakers changed F_0 categorically without the additional information from the manual annotation. On the other hand, the big variations found for the L2 data were problematic for the manual annotation, but not for the data-driven methods. There were also some pros and cons between the two data-driven methods. While the FPCA showed abstracted F_0 contours constructed by PCs, the SOM visualised typical contours for both language groups without destroying or abstracting original data. On the other hand, the SOM did not allow us to run statistics on the data, while the FPCA outputted numbers on which common statistical analyses can be done. Ultimately, it is important to note that subjective decisions are involved both when using manual annotations and data-driven methods, but in different domains and in different manners. Even the FPCA and the SOM, which are supposed to be objective in some aspects, required subjective decisions on the interpretations and categorisations of the outputs.

Such data-driven methods proved technically more advantageous for larger data sets, because the manual annotation becomes more tedious with larger data points. The FPCA and the SOM are generalisable to all kinds of data for which feature vectors can be derived, and the systems are capable of handling a much larger data volume than covered by this experiment. It can be emphasised that the data-driven methods do not require to decide beforehand which aspects of the functional data are important to annotate, so that they are also suitable for an explorative analysis, without prior expectation as to what kind of differences between the groups will come out. Human perception can reason about their problems, leverage domain knowledge and perceive patterns efficiently, while computers (to undertake a data-driven data analysis) can process huge amounts of data incredibly fast.

In this case, the methodological triangle enabled us to get a more complete picture of the data. The FPCA and the SOM are purely data-driven and are neither based on a prosodic model nor on the perception of a human annotator. They provide a complementary perspective to the analysis based on the manual annotation. Moreover, the FPCA and the SOM can also be used as a tool to examine which annotation categories may be useful for a data set before starting to annotate the data. While not proposing the FPCA and the SOM as a full substitute of the annotation analysis, I will take further opportunities to use these methods.

6.4 Conclusions

The aim of this thesis was to investigate the sources of the difficulties found in the productions of L2 prosody. Processing of nonnative pitch contrasts and segmental length contrasts in the "input", "mental representations" and "output" stages were tested by manipulating the cognitive load placed on working memory in terms of memory load and attention control. The findings from the experiments led me to the following conclusions. First, the nonnative speakers were successful in the "input" stage when task demands were the lowest without distracting factors. Once distracting factors came into play, their performance decreased. When more phonological representations were required and accessed in the "mental representations" stage, the nonnative speakers' performance differed from the native speakers' one. In the "output" stage, the nonnative speakers were not successful, suggesting that they might have had articulatory difficulties. Second, manipulating cognitive load, it was shown that L2 processing was vulnerable to increased cognitive load, while L1 processing was more stable. The results showed that it was difficult to automatise L2 processing. Third, the learners generally performed better than the non-learners, showing a positive learning effect of the L2. However, RT analyses revealed that the non-learners were faster than the learners in responding, suggesting that they applied different strategies for the task. Fourth, the study has shown that there is no general "prosodic processing" that can be applied to all kinds of prosodic cues in the same way. The results of segmental length contrasts and pitch contrasts in this study differed in numerous aspects, suggesting that their underlying mechanisms were different. These differences were caused by the fact that each prosodic cue is used or not used at different linguistic levels in different languages. The combination of "what prosodic cue" and "at which linguistic level" in an L1 and an L2 makes the processing of that cue unique. At least in this thesis, it was shown that the same prosodic property used

in an L1 and in an L2 was not sufficient for the successful processing of that property, but it must be used at the same linguistic level in both languages. The distinction between the lexical vs. non-lexical use of prosody was relevant. Fifth, perceived (dis)similarities of a cross-linguistic prosodic contrast should be defined taking multiple dimensions into account, such as not only phonological, but also phonetic differences, the frequency of use of the category or the linguistic levels at which the prosodic cue builds a meaningful contrast in an L1 and L2. Moreover, the variability of L2 perception that is influenced by cognitive load and task demands should be discussed more when modelling the processing of (L2) prosody. Finally, the findings suggest that the phonetic and phonological processing of prosody are not mutually exclusive and they are on a continuum. Depending on the gradual change of the cognitive load placed on working memory, more or less phonetic and phonological processing are required.

The contributions offered to the field by this study are as follows. The first contribution is that the cross-sectional investigation by testing the same participants using the same stimuli captured the relationships between speech perception, mental representations and production, which are connected and contribute to foreign accented speech in prosody. The analysis of multiple stages of speech processing is a valuable contribution to the field as most of the individual studies has investigated only a single factor at a time. The second contribution is that I showed the vulnerability of L2 processing by manipulating experimental task variables, and underscored the importance of the role of cognitive load to account for the L2 processing. The investigation on memory load and attention control that increased cognitive load was original. The third contribution is that by tracing speech perception and production back to their basis and conducting basic simple experimental tasks, I revealed the fine-graded differences in the processing of different prosodic properties in the L1 processing. The fourth contribution is that two prosodic cues were investigated within one study. The first experiment tested the integration of the two cues and the other experiments tested how to control attention in order to ignore one prosodic cue to better process the other one, in which I did not find reciprocal effects. The latter investigation was my original. Since prosody is layered at different linguistic levels multiply and almost uniquely in each language, not only the investigation on the differences of the same prosodic cue across languages, but also on the differences among several prosodic cues within a language is fundamental for further theoretical contributions.

BIBLIOGRAPHY

Akaba, S. (2008). *An Acoustic Study of the Japanese Short and Long Vowel Distinction*. PhD thesis, University of Kansas.

Aliaga-Garcia, C., Mora, J. C., and Cerviõ-Povedano, E. (2011). L2 speech learning in adulthood and phonological short-term memory. In Dziubalska-Kolaczyk, K., editor, *Poznań Studies in Contemporary Linguistics PSiCL*, volume 47. De Gruyter.

Altmann, H., Berger, I., and Braun, B. (2012). Asymmetries in the perception of non-native consonantal and vocalic length contrasts. *Second Language Research*, 28(4):387–413.

Ambrazaitis, G. I. and Niebuhr, O. (2008). Dip and hat pattern: a phonological contrast of German? In *Proceedings of the 4th International Conference of Speech Prosody*, Campinas, Brazil.

Anderson, J. (1987). The markedness differential hypothesis and syllable structure difficulty. In Ioup, G. and Weinberger, S., editors, *Interlanguage phonology: the acquisition of a second language sound system*, pages 279–291. Newbury House/Harper & Row, New York.

Anderson, P. J. (2002). Assessment and development of executive functioning (EF) in childhood. *Child Neuropsychology*, 8(2):71–82.

Anderson-Hsieh, J. and Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning*, 38(4):561–613.

Antoniou, M., Wong, P. C. M., Ingvalson, E., and Wang, S. (2013). Cognitive factors contribute to speech perception: Implications for sound change actuation. In *Poster presented at Workshop on Sound Change Actuation*, Chicago, USA. University of Chicago.

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, 66(1-2):46–63.

Asano, Y. (2015). Coordination of lexical and paralinguistic f_0 in L2 production. In *Proceedings of the 18th international congress of phonetic Sciences*, Glasgow, UK: the University of Glasgow. ISBN 978-0-85261-941-4. Paper number 577.

Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In Spence, K., editor, *The psychology of learning and motivation*, volume 2, pages 89–195. Academic Press, Oxford, UK.

Atkinson, R. C. and Shiffrin, R. M. (1971). The control processes of short-term memory. Technical Report 173, Institute for Mathematical Studies in the Social Sciences, Stanford University.

Ayusawa, T., Nishinuma, Y., Lee, M. H., Arai, M., Odaka, K., and Hoki, N. (1995). Analysis of perceptual data on the Tokyo accent: Results from 10 language groups. In *Research bulletin of new program for 'Japanese studies', 2nd meeting of study report about 'Japanese' in the international society*, pages 25–32.

Ayusawa, T. and Odaka, K. (1998). Analysis of perceptual data on the Tokyo accent: Results from 21 language groups. In *Collection of research papers about 'Japanese' in the international society, new program for 'Japanese studies'*, pages 57–71.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, 4(11):417–423.

Baddeley, A. (2010). Working memory. *Current Biology*, 20(4):136–140.

Baddeley, A., Gathercole, S., and Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, 105(1):158–173.

Baddeley, A. and Wilson, B. A. (2002). Prose recall and amnesia: implications for the structure of working memory. *Neuropsychologia*, 40(10):1737–1743.

Baddeley, A. D. (1986). *Working memory*. Oxford University Press, Oxford.

Baddeley, A. D. (2003). Working memory and language: an overview. *Journal of Communication Disorders*, 36(3):189–208.

Baddeley, A. D. and Hitch, G. J. (1974). Working memory. In Bower, G. H., editor, *The psychology of learning and motivation*, volume 8, pages 47–90. Academic Press, London.

Baddeley, A. D. and Logie, R. H. (1999). Working memory: The multiple component model. In Miyake, A. and Shah, P., editors, *Models of working memory: Mechanisms of active maintenance and executive control*, pages 28–61. Cambridge University Press, New York.

Balota, D. A. and Chumbley, J. (1985). The locus of word-frequency effects in the pronunciation task: Lexical access and/or production? *Journal of Memory and Language*, 24:89–106.

Bänziger, T. and Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*, 46(3–4):252–267.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Barry, C. (1981). Hemispheric asymmetry in lexical access and phonological encoding. *Neuropsychologia*, 19(3):473–478.

Bartels, C. (1997). *Towards a compositional interpretation of English statement and question intonation*. PhD thesis, University of Massachusetts, Massachusetts, USA.

Baumann, S. and Grice, M. (2006). The intonation of accessibility. [invited contribution to journal of pragmatics]. *Journal of Pragmatics*, 38:1636–1657.

Baumann, S., Grice, M., and Benzmüller, R. (2001). GToBI – a phonological system for the transcription of German intonation. In Puppel, S. and Demenko, G., editors, *Prosody 2000: Speech recognition and synthesis*, 21–28. Adam Mickiewicz University.

Beckman, M. E. (1982). Segment duration and the 'mora' in Japanese. *Phonetica*, 39:113–135.

Beckman, M. E. (1986). *Stress and non-stress accent*. Foris, Dordrecht.

Beckman, M. E. and Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, 3:54.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics. Identifying Influential Data and scourses of Colinearity*. Wiley Series in Probabiliby and Mathematical Statistics. Wiley, New York.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In Goodman, J. and Nusbaum, H. C., editors, *The development of speech perception: The transition from speech sounds to spoken words*, chapter 6. MIT Press, Cambridge MA, 167–224 edition.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In Strange, W., editor, *Speech perception and linguistic experience*. York Press, Timonium MD.

Best, C. T. (1996). A direct realist perspective on cross-language speech perception. In Strange, W., editor, *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research*, chapter 6, pages 167–200. York Press, Timonium MD.

Best, C. T., McRoberts, G. W., and Goodell, E. (2001). Discrimination of nonnative consonant contrasts varying in perceptual assimilation to the listeners native phonological system. *Journal of Acoustical Society of America*, 109:775–794.

Best, C. T. and Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In Munro, M. J. and Bohn, O.-S., editors, *Second Language Speech learning: The role of language experience in speech perception and production*, pages 13–34. John Benjamins, Amsterdam.

Best, J. R., Miller, P. H., and Jones, L. L. (2009). Executive functions after age 5: Changes and correlates. *Developmental review : DR*, 29(3):180–200.

Bialystock, E. (1992). Attentional control in children's metalinguistic performance and measures of field independence. *Develop. Psychol.*, 28(4):654–664.

Bialystock, E. (2005). Consequences of bilingualism for cognitive development. In Kroll, J. F. and de Groot, A. M. B., editors, *Handbook of Bilingualism: Psycholinguistic Approaches*. Oxford University Press, Oxford, UK.

Bialystok, E., Craik, F. I., and Luk, G. (2012). Bilingualism: Consequences for mind and brain. *Trends in Cognitive Sciences*, 16(4):240–250.

Bin, K. (1993). Nihongo sokuon no choushuu handan ni kansuru kenkyuu [study on perception of japanese geminate]. *Sekai no nihongo Kyooiku 3*, 3:237–249.

Bloch, B. (1950). Studies in colloquial Japanese IV phonemics. *Language*, 26(1):86–125.

Blumenfeld, H. K. and Marian, V. (2007). Constraints on parallel activation in bilingual spoken language processing: Examining proficiency and lexical status using eye-tracking. *Language and Cognitive Processes*, 22(5):633–660.

Bock, J. K. (1982). Towards a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, 89:1–47.

Boersma, P. and Weenink, D. (2011). Praat: doing phonetics by computer [computer program] version 5.2.20.

Bohn, O.-S. (1995). Cross-language speech perception in adults: first language transfer doen't tell it all. In Strange, W., editor, *Speech perception and linguistic experience*, pages 279–304. York Press, Timonium MD.

Bohn, O.-S. (2002). On phonetic similarity. In Burmeister, P., Piske, T., and Rohde, A., editors, *An integrated view of language development: Papers in honor of Henning Wode*, pages 191–216. Wissenschaftlicher Verlag Trier.

Bongaerts, T. (1999). Ultimate attainment in L2 pronunciation: The case of very advanced late L2 learners. In Birdsong, D., editor, *Second language acquisition and the Critical Period Hypothesis*, chapter 6, pages 133–159. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Bongaerts, T., Planken, B., and Schils, E. (1995). Can late starters attain a native accent in foreign language? In Singleton, D. and Lengyel, Z., editors, *The Age Factor in Seond Language Acquisition*, pages 30–50. Multilingual Matters, Clevedon.

Borràs-Comes, J., M., V. M., and Pieto, P. (2010). The role of pitch range in establishing intonational contrasts in Catalan. In *Proceedings of the 5th International Conference on Speech Prosody*, pages 1–4, Chicago.

Boula de Mareüil, P. and Vieru-Dimulescu, V. (2006). The contribution of prosody to the perception of foreign accent. *Phonetica*, 63:247–267.

Braun, B. (2006). Phonetics and phonology of thematic contrast in German. *Language and cognitive processes*, 26:224–235.

Braun, B., Dainora, A., and Ernestus, M. (2011). An unfamiliar intonation contour slows down on-line speech comprehension. *Language and cognitive processes*, 26:350–375.

Braun, B., Galts, T., and Kabak, B. (2014). Lexical encoding of L2 tones: The role of L1 stress, pitch accent and intonation. *Second Language Research*, 30(3):323–350.

Braun, B. and Johnson, E. (2011). Question or tone 2? How language experience and linguistic function guide pitch processing. *Journal of Phonetics*, 39(4).

Braun, B., Kochanski, G., Grabe, E., and Rosner, B. S. (2006). Evidence for attractors in English intonation. *Journal of Acoustical Society of America*, 119(6):4006–4015.

Brinckmann, C. and Benzmüller, R. (1999). The relationship between utterance type and f_0 contour in German. In *Proceedings 6th European Conference on Speech Communication and Technology (Eurospeech 99)*, pages 21–24, Budapest, Hungary.

Bürki-Cohen, J., Miller, J. L., and Eimas, P. D. (2011). Perceiving non-native speech. *Language and Speech*, 44(2):149–169.

Burnham, D. and Francis, E. (1997). The role of linguistic experience in the perception of Thai tones. In Abramson, A., editor, *Southeast Asian linguistic studies in honour of Vichin Panupong*, pages 29–47. Chulalongkorn University Press.

Campbell, N. (1999). A study of Japanese speech timing from the syllable perspective. *J. Phon. Soc. Japan*, 3(2):121–130.

Chen, A. (2003). Reaction time as an indicator of discrete intonational contrasts in English. In *Proceeding of the 8th EuroSpeech, Geneva, Switzerland*, pages 97–100.

Chen, A. (2005). *Universal and language-specific perception of paralinguistic intonational meaning*. LOT.

Chen, A. and Mennen, I. (2008). Encoding interrogativity intonationally in a second language. In *Proceedings of the 4th International Conferences on Speech Prosody*, pages 513–516.

Chun, D. (2002). *Discourse intonation in L2: from theory and research to practice*. John Benjamins B.V.

Cibelli, E. (2012). Shared early pathways of word and pseudoword processing: Evidence from high-density electrocorticography. *UC Berkeley Phonology Lab Annual Report*, pages 111–127.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Physiological Measurements*, 20:37–46.

Cohen, J. (1969). *Statistical power analysis for the behavioural sciences*. Academic Press, New York, 2nd edition.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1):155–159.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49:997–1003.

Coleman, J. A., Ruediger, G., and Raatz, U., editors (1994). *University Language Testing and the C-Test*. AKS-Verlag Bochum.

Collentine, J. (1998). Processing instruction and the subjective. *Hispania*, 81(576–587).

Colzato, L. S., Bajo, M. T., van den Wildenberg, W., and Paolieri, D. (2008). How does bilingualism improve executive control? a comparison of active and reactive inhibition mechanisms. *Journal of Experimental Psychology: Learning, Memory and Cognitition*, 34(2):302–312.

Costa, A., Hernandes, M., and Sebastian-Galles, N. (2008). Bilingualism aids conflict resolution: Evidence from the ANT task. *Cognition*, 106:59–86.

Costa, A. and Santesteban, M. (2004). Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners. *Journal of Memory and Language*, 50(4):491–511.

Couper-Kuhlen, E. (1986). *An Introduction to English Prosody*. Edward Arnold and Niemeyer, London and Tübingen.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research*, 169:323–338.

Cowan, N. and Morse, P. (1986). The use of auditory and phonetic memory in vowel discrimination. *Journal of Acoustical Society of America*, 79:500–507.

Crowder, R. G. (1982). Decay of auditory memory in vowel discrimination. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 8:153–162.

Crowder, R. G. and Morton, J. (1969). Precategorical acoustic storage (PAS). *Perception & Psychophysics*, 5:365–373.

Cumming, G. (2011). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*. Routledge, New York.

Cumming, G. (2013). The new statistics: Why and how. *Psychological Science*.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, 28(3):369–382.

Cutler, A., Cooke, M., Lecumberri, M. L. G., and D., P. (2007). L2 consonant identification in noise: Cross-language comparisons. In *Proceedings of the 8th Interspeech*, pages 1585–1588, Antwerp, Belgium.

Cutler, A. and Otake, T. (1999). Pitch accent in spoken-word recognition in Japanese. *Acoustical Society of America*, 105(3):1877–1888.

Cutler, A., Weber, A., and Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics*, 34(2):269–284.

Dainora, A. (2002). Does intonational meaning come from tones or tunes? Evidence against a compositional approach. In Bell, B. and Marlien, I., editors, *proceedings of the 1st international conference on Speech Prosody*, Aix-en-Provence, France.

Dallett, K. M. (1964). Intelligibility and short-term memory in the repetition of digit strings. *Journal of Speech and Hearing Research*, 7:362–368.

Darcy, I., Dekydtspotter, L., Sprouse, R. A., Glover, J., Kaden, C., McGuire, M., and Scott, J. H. (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):5–40.

De Boor, C. (2001). *A practical guide to splines*. Springer, New York.

De Luca, C. R. and Leventer, R. J. (2008). Developmental trajectories of executive functions across the lifespan. In Anderson, P., Anderson, V., and Jacobs, R., editors, *Executive functions and the frontal lobes: a lifespan perspective*, pages 3–21. Taylor & Francis, Washington D. C.

Delattre, P. (1951). *Principes de phonétique française*. Middlebury College, Middlebury.

Deutsch, D., Henthorn, T., Marvin, E., and Xu, H. (2006). Absolute pitch among American and Chinese conservatory students: Prevalence differences, and evidence for a speech-related critical period. *Journal of Acoustical Society of America*, 119:719–722.

Dijkstra, T. and van Heuven, W. J. B. (1998). The BIA-model and bilingual word recognition. In Grainger, J. and Jacobx, A., editors, *Localist connectionist approaches to human cognition*. LEA, Hove.

Dimitrov, D. M. and Rumrill, P. D. (2003). Pretest-posttest designs and measurement of change. *Rehabilitation*, 20(2):159–165.

Dobel, C., Lagemann, L., and Zwitserlood, P. (2009). Non-native phonemes in adult word learning: evidence from the N400m. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1536):3697–3709.

Dombrowski, E. and Niebuhr, O. (2005). Acoustic patterns and communicative functions of phrase-final f0 rises in German: Activating and restricting contours. *Phonetica*, (62):176–195.

Dombrowski, E. and Niebuhr, O. (2010). Shaping phrase-final rising intonation in German. In *Proceedings of the 5th International Conference on Speech Prosody*, pages 1–4, Chicago, USA.

Dupoux, E., Peperkamp, S., and Sebastián-Gallés, N. (2001). A robust method to study stress "deafness". A robust method to study stress "deafness". *Journal of Acoustical Society of America*, 110(3):1606–1618.

Dupoux, E., Sebastián-Gallés, N., Navarrete, E., and Peperkamp, S. (2008). Persistent stress "deafness": the case of French learners of Spanish. *Cognition*, 106:682–706.

Dupoux, E., P. C. S.-G. N. and Mehler, J. (1997). A destressing deafness in French. *Journal of Memory and Language*, 36:165–179.

Eckman, F. (1977). Markedness and the contrastive analysis hypothesis. *Language Learning*, 27:315–330.

Eckman, F. (2008). Typological markedness and second language phonology. In Edwards, H., Zampini, J. G., and Zampini, M. L., editors, *Phonology and Second Language Acquisition*, pages 95–115. Benjamins.

Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford University Press, Oxford.

Enomoto, K. (1992). Interlanguage phonology: The perceptual development of durational contrasts by English-speaking learners of Japanese. *Edinburgh Working Papers in Applied Linguistics*, 3:25–36.

Esling, J. H. and Wong, R. F. (1983). Voice quality settings and the teaching of pronunciation. *TESOL Quarterly*, 17(1):89–95.

Féry, C. (1993). *German Intonational Patterns*. Niemeyer, Tübingen.

Flege, J. E. (1995). Second language speech learning: theory, findings, and problems. In Strange, W., editor, *Speech perception and linguistic experience*, pages 233–277. York Press, Timonium MD.

Flege, J. E. (1999). Age of learning and second-language speech. In Birdsong, D., editor, *Second language acquisition and the Critical Period Hypothesis*, pages 101–132. Lawrence Erlbaum Associates, Hilsdale, NJ.

Flege, J. E., Bohn, O.-S., and Jan, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, 25:437–470.

Flege, J. E., MacKay, I., and Piske, T. (2002). Assessing bilingual dominance. *Applied Psycholinguistics*, 23:567–598.

Flege, J. E., Munro, M. J., and Mackay, I. (1995a). Effects of age of second-language learning on the production of English consonants. *Speech Communication*, 16:1–26.

Flege, J. E., Munro, M. J., and Mackay, I. R. A. (1995b). Factors affecting strength of the perceived foreign accent in second language. *Journal of Acoustical Society of America*, 97:3125–3134.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14:3–28.

Fowler, C. A. (1990a). Listener-talker attunements in speech. *Haskins Laboratories Status Report on Speech Research*, SR-101/102:110–129.

Fowler, C. A. (1990b). Sound-producing sources as objects of perception: Rate normalization and nonspeech perception. *Journal of Acoustical Society of America*, 88:1236–1249.

Fowler, C. A. and Rosenblum, L. D. (1989). The perception of phonetic gestures. *Haskins Laboratories Status Report on Speech Research*, 99/100:102–117.

Francis, A. L., Ciocca, V., Ma, L., and Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *Journal of Phonetics*, 36:268–294.

Fujisaki, H. (1992). Modeling the process of fundamental frequency contour generation. In Tohkura, Y., Vatikiotis-Bateson, E., and Sagisaka, Y., editors, *Speech Perception, Production and Linguistic Structure*, pages 313–328. Ohmsha.

Fujisaki, H. and Hirose, K. (1993). Analysis and perception of intonation expressing paralinguistic information in spoken japanese. In *ISCA Workshop on Prosody*, pages 254–257.

Fujisaki, H. and Kawashima, T. (1969). Some experiments on speech perception and a model for the perceptual mechanism. *Annual Report of the Engineering Research Institute (Tokyo)*, 28:67–73.

Fujisaki, H. and Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism. *Annual Report of the Engineering Research Institute (Tokyo)*, 29:207–214.

Fujisaki, H., Nakamura, K., and Imoto, T. (1975). Auditory perception of duration of speech and non-speech stimuli. In Fant, G. and Tatham, M. A. A., editors, *Auditory Analysis and Perception of Speech*, pages 197–219. Academic Press, London.

Fukui, S. (1978). Nihongo no heisaon no enchoo tanshuku ni yoru sokuon hisokuon to shite no chooshu (Perception for the Japanese stop connsonants with reduced and extended durations. *Bull. Phon. Soc. Japan*, 159:9–12.

Gábor, K. and Mihály, R. (2008). Handling L2 input in phonological STM: The effect of non-L1 phonetic segments and non-L1 phonotactics on nonword repetition. *Language Learning*, 58(3):597–624.

Garrett, M. F. (1982). Production of speech: Observations from normal and pathological language use. In Ellis, A. W., editor, *Normality and pathology in linguistic performance: Slips of the tongue, ear, pen and hand*, pages 19–76. Academic Press, New York.

Gass, S. and Mackey, A. (2002). Frequency effects and second language acquisition. *Studies in Second Language Acquisition*, 24:249–260.

Gathercole, S. E. (1999). Cognitive approaches to the development of short-term memory. *Trends in Cognitive Sciences*, 3:410–418.

Gathercole, S. E., Hitch, G. J., Service, E., and Martin, A. J. (1997). Phonological short-term memory and new word learning in children. *Dev. Psychol.*, 33(6):966–979.

Gerrits, E. (2001). *The categorisation of speech sounds by adults and children: a study of the categorical perception hypothesis and the development weighting of acoustic speech cues*. PhD thesis, Universititeit Utrecht.

Gibbon, D. (1998). Intonation in German. In Hirst, D. and Cristo, A., editors, *Intonation systems: a survey of twenty languages*, chapter 4, pages 78–95. Cambridge University Press.

Gick, B., Wilson, I., Koch, K., and Cook, C. (2004). Language-specific articulatory settings: evidence from inter-utterance rest position. *Phonetica*, 61:220–233.

Gluszek, A. and Dovidio, J. F. (2010). The way they speak: A social psychological perspective on the stigma of nonnative accents in communication. *Personality and Social Psychology Review*, 14(2):214–237.

Goblirsch, K. G. (1990). *Consonant strength and quantify in Upper German dialects*. PhD thesis, University of Minnesota.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognitition*, 22:1166–83.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105:251–279.

Goyet, L., de Schonen, S., and Nazzi, T. (2010). Syllables in word segmentation by French learning infants: an ERP study. *Brain Research*, (1332):75–89.

Grabe, E., Kochanski, G., and Coleman, J. (2007). Connecting intonation labels to mathematical descriptions of fundamental frequency. *Language and Speech*, 50(3):281–310.

Gårding, E. (1981). Contrastive prosody: A model and its application. *Studia Lingusitica*, 35:146–165.

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, 1:67–81.

Greenberg, J. H. (1966). *Universals of Language*. MIT Press, Cambridge, MA and London, 2nd edition.

Grice, M., Baumann, S., and Benzmüller, R. (2005). German intonation in autosegmental-metrical phonology. In Sun-Ah, J., editor, *Prosodic Typology. The Phonology of Intonation and Phrasing*, pages 55–83. Oxford University Press, Oxford.

Grice, M., Reyelt, M., Benzmüller, R., Mayer, J., and Batliner, A. (1996). Consistency in transcription and labelling of German intonation with GToBI. In *Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia, USA.

Gubian, M., Asano, Y., Asaridou, S., and Cangemi, F. (2013). Rapid and smooth pitch contour manipulation. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association*, pages 31–35, Lyon, France.

Gubian, M., Boves, L., and Cangemi, F. (2011). Joint analysis of f0 and speech rate with functional data analysis. In *Proceedings of the 36th International Conference on Acoustics, Speech and Signal Processing*, pages 4972–4975, Prague, Czech Republic Prague, Czech Republic.

Guiora, A. (1990). A psychological theory of second language pronunciation. *Toegepaste Taalwetenschap in Artikelen*, 37:15–23.

Gussenhoven, C. (2004). *The phonology of tone and intonation*. Research surveys in linguistics. Cambridge University Press, Cambridge ; New York. 2003065202 Carlos Gussenhoven. ill. ; 24 cm. Includes bibliographical references (p. 321-344) and index.

Gutknecht, C. (1979). Intonation and language learning: The necessity for an integrative approach. *Studies in Second Language Acquisition*, 1(2):25–36.

Hall, T. A. (2007). Segmental features. In De Lacy, P., editor, *The Cambridge Handbook of Phonology*, pages 311–334. Cambridge University Press.

Hallé, P. A., Chang, Y. C., and Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics*, 32:395–421.

Ham, W. (2001). *Phonetic and Phonological Aspects of Geminate Timing*. Routledge.

Han, M. S. (1962). The feature of duration in Japanese. *Onsei no kenkyuu*, 10:65–80.

Han, M. S. (1992). The timing control of geminate and single stop consonants in Japanese: A challenge for nonnative speakers. *Phonetica*, 49:102–127.

Han, M. S. (1994). Acoustic manifestations of mora timing in Japanese. *Acoustical Society of America*, 96(1):73–82.

Harada, T. (2006). The acquisition of single and geminate consonants by English-speaking children in a Japanese immersion program. *Studies in Second Language Acquisition*, 28(4):601–632.

Haraguchi, S. (1977). *The tone pattern of Japanese: An autosegmental theory of tonology*. Kaitakusha, Tokyo.

Harnad, S. (1987). *Categorical Perception: The Groundwork of Cognition*. Cambridge University Press, New York.

Hatasa, Y. A. and Tohsaku, Y. (1997). Spot as a placement test. Technical report, University of Hawai'i Press.

Hayashi, R. (1996). Wie nehmen deutsche Japanischlerner den japanischen Akzent wahr? *Der Keim*, 20:15–23.

Hayes-Harb, R. and Masuda, K. (2008). Development of the ability to lexically encode novel L2 phonemic contrasts. *Second Language Research*, 24(1):5–33.

He, X. (2012). *Mandarin-accented Dutch Prosody*. PhD thesis, Radboud Universiteit Nijmegen, Nijmegen.

Hilchey, M. D. and Klein, R. M. (2011). Are there bilingual advantages on nonlinguistic interference tasks? implications for the plasticity of executive control processes. *Psychonomic Bulletin & Review*, 18:625–658.

Hill, K. T. and Miller, L. M. (2010). Auditory attentional control and selection during cocktail party listening. *Cerebral Cortex*, 20(3):583–590.

Hintzman, D. L. (1986). "schema abstraction' in a multiple-trace memory model. *Psychological Review*, 93:411–428.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, 95:528–551.

Hirano-Cook, E. (2011). *Japanese pitch accent acquisition by learners of Japanese: Effects of training on Japanese accent instruction, perception, and production.* PhD thesis, University of Kansas.

Hirata, E., Ayusawa, T., Nakagawa, C., and Odaka, K. (1997). Perception test of tokyo accent in japanese. In *Study Report*, pages 61–72. The Ministry of Education, Japan.

Hirata, Y. (1990). Tango level, bun level ni okeru sokuon no kikitori - eigo wo bogo to suru NNS no baai [perception of geminates at word and sentence level - the case study for English learners of Japanese. *Onseigakkai kaihou*, 195:4–10.

Hirata, Y. (2007). Durational variability and invariance in Japanese stop quantity distinction: Roles of adjacent vowels. *J. Phon. Soc. Japan*, 11:9–22.

Hirata, Y. and Whiton, J. (2005). Effects of speaking rate on the single/geminate stop distinction in Japanese. *Journal of Acoustical Society of America*, 118(3):1647–1660.

Hirose, K. and Minematsu, N. (2004). Use of prosodic features in speech recognition. In *Proceedings of the 8th International Conference on Spoken Language Processing*, pages 1445–1448, Jeju, Island.

Hirschfeld, U. (1994). *Untersuchungen zur phonetischen Verständlichkeit Deutschlernender*, volume Forum Phoneticum Bd. 57. Hector, Frankfurt a. M.

Hirschfeld, U. and Trouvain, J. (2007). *Teaching prosody in German as foreign language*, pages 171–187. Mouton de Gruyter, Berlin.

Hirst, D. and Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. pages 75–85.

Holcomb, P. J. and Neville, H. J. (1990). Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. *Language and Cognitive Processes*, 5(4):281–312.

Holm, S. (2007). The relative contributions of intonation and duration to intelligibility in Norwegian as a second language. In *Proceedings of the 16th International Congress of the Phonetic Sciences*, Saarbr´ucken, Germany.

Homma, Y. (1981). Durational relationships between Japanese stops and vowels. *Journal of Phonetics*, 9(3):273–281.

Honda, M. (2007). The role of prosody in Japanese: The use of pitch information in spoken word recognition by L1 and L2 speakers. In *Proceedings of the 5th Cambridge Postgraduate Conference in Linguistics*, pages 96–103.

Honikman, B. (1964). Articulatory settings. In Abercrombie, D., Fry, D. B., MacCarthy, P. A. D., Scott, N. C., and Trim, J. L. M., editors, *In honour of Daniel Jones*, pages 73–84. Longman.

Hyman, L. M. (2001). Tone systems. In Haspelmath, M., K´onig, E., Oesterreicher, W., and Raible, W., editors, *Language typology and language unviersals: An international Handbook*, volume 2, pages 1367–1380. Walter de Gruyter.

Hyman, L. M. (2006). Word-prosodic typology. *Phonology*, 23:225–257.

Hyman, L. M. (2007). How (not) to do phonological typology: The case of pitch-accent. *Berkeley Phonology Lab Annual Report*, pages 654–685.

Hyman, L. M. (2009). How (not) to do phonological typology: the case of pitchaccent. *Language Science*, 31:213–238.

Hyman, L. M. and Wilson, S. (1992). Review of Harry van der Hulst & Norval Smith (eds), *Autosegmental studies on pitch accent* (dordrecht: Foris). *Language*, 67:356–363.

Idemaru, K. and Guion, S. G. (2008). Acoustic covariants of length contrast in Japanese stops. *Journal of the international Phonetic Association*, 38:167–186.

Idemaru, K. and Guion-Anderson, S. (2010). Relational timing in the production and perception of Japanese singleton and geminate stops. *Phonetica*, 67(1-2):25–46.

Imada, T., Hari, R., Loveless, N., McEvoy, L., and Sams, M. (1993). Determinants of the auditory mismatch response. *Electroencephalogr. Clin. Neurophysiol.*, 87(3):144–53.

Isaacs, T. and Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics*, 32:113–140.

Ishihara, S. (2011). Japanese focus prosody revisited: Freeing focus from prosodic phrasing. *Lingua*, 121(13):1870–1889.

Jiang, N. (2012). *Conducting Reaction Time Research in Second Language Studies*. Routledge, New York & London.

Jilka, M., Anufryk, V., Baumotte, H., Lewandowski, N., Rota, G., and Reiterer, S. (2007). Assessing individual talent in second language production and perception. In Rauber, A. S., Watkins, M. A., and Baptista, B. O., editors, *New Sounds 2007: Proceedings of the 5th International Symposium on the Acquisition of Second Language Speech*, pages 243–258. Federal University of Santa Catarina.

Johansson, S. (1978). *Studies in error gravity*. Gothenburg University, Gothenburg.

Johnson, K. (1997). Speech perception without speaker normalisation: An exemplar model. In Johnson, K. and Mullemix, J. W., editors, *Talker variability in speech processing*, pages 145–166. Academic Press.

Johnson, K. (2004). Cross-linguistic perceptual differences emerge from the lexicon. In Agwuele, G., Warren, W., and park, S.-H., editors, *Proceedings of the 2003 Texas Linguistics Society Conference: Coarticulation in Speech Production and Perception*, pages 26–41, Sommerville, MA. Cascadilla Press.

Jun, S.-A. and Oh, M. (2000). Acquisition of second language intonation. In *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China.

Jusczyk, P. W., Houston, D., and Newsome, M. (1999). Th e beginnings of word segmentation in english-learning infants. *Cognitive Psychology*, 39(3–4):159–207.

Kabak, B., Reckziegel, T., and Braun, B. (2011). Timing of second language singletons and geminates. In *The 17th International Congress of Phonetic Sciences in Hong Kong, 17-21 August 2011*, pages 994–997.

Kaisse, E. M. (1984). *Connected Speech: the interaction of syntax and phonology*. Academic Press, New York.

Kaisse, E. M. and Shaw, P. A. (1985). On the theory of Lexical Phonology. *Phonology Yearbook*, 2:1–30.

Kawahara, S. (2006). Contextual effects on the perception of duration. *Journal of Acoustical Society of America*, 119(5):3243.

Kerr, J. (2000). Articulatory setting and voice production: Issues in accent modification. *Prospect*, 15(2):4–15.

Kin, O. (2005). Ks no hatsuon no ayamari ni kansuru kousatsu - choukai tesuto no kekka bunseki wo chuushin ni [study on pronunciation error by Korean learners]. *Nichigo-nichibungaku*, 26:123–134.

Kinoshita, K., Behne, D., and Arai, T. (2002). Duration and f0 as perceptual cues to Japanese vowel quantity. In *Proc. of the 7th International Conf. on Spoken Language Processing Denver*, pages 757–760.

Kitamura, C. and Burnham, D. (2003). Pitch and communicative intent in mother's speech: Adjustment for age and sex in the first year. *Infancy*, 41(1):85–110.

Kohonen, T. (2001). *Self-organising maps*, volume 30. Springer.

Kondo, M. (2009). Is acquisition of L2 phonemes difficult? Production of English stress by Japanese speakers. In Bowles, M. e. a., editor, *Proceedings of the 10th Generative Approaches to Second Language Acquisition Conference (GASLA 2009)*, pages 105–112, Somerville, MA. Cascadilla Proceedings Project.

Kormos, J. and Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition*, 11:261–271.

Kraehenmann, A. (2001). Swiss German stops: Geminates all over the word. *Phonology*, 18(1):109–145.

Kraehenmann, A. (2003). *Quantity and prosodic asymmetries in Alemannic*. De Gruyter, Berlin & New York.

Kraehenmann, A. and Lahiri, A. (2008). Duration differences in the articulation and acoustics of Swiss German word-initial geminate and singleton stops. *Acoustical Society of America*, 123(6):4446–4455.

Krishnan, A., Swaminathan, J., and Gandour, J. T. (2009). Experience- dependent enhancement of linguistic pitch representation in the brainstem is not specific to a speech context. *J. Cogn. Neurosci.*, 21:1092–1105.

Kritzman, J., Soke, E., Marian, V., and Kraus, N. (2014). Bilingualism increases neural response consistency and attentional control: Evidence for sensory and cognitive coupling. *Brain and Language*, 128:34–40.

Kubozono, H. (1999). *Nihongo-no onsei [The Japanese sounds.]*. Iwanami Shoten, Tokyo.

Kubozono, H. (2011a). Geminate obstruents and accent in Japanese. *NINJAL Project Review*, 6:3–15.

Kubozono, H. (2011b). Japanese pitch accent. In van Oostendorp, M., Ewen, C. J., Hume, E., and Rice, K., editors, *The Blackwell Companion to Phonology*, pages 2879–2907. Blackwell Publishing Ltd.

Kubozono, H., Takeyasu, H., Giriko, M., and Hirayama, M. (2011). Pitch cues to the perception of consonant length in Japanese. In *Proc. of the17th International Congress of Phonetic Sciences Hong Kong*, pages 1150–1153.

Kuhl, P. K. (1991). Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50(2):93–107.

Kuhl, P. K. (1993). Early linguistic experience and phonetic perception: Implications for theories of developmental speech production. *Journal of Phonetics*, 21:125–139.

Kuhl, P. K. and Iverson, P. (1995). Linguistic experience and 'perceptualmagnet effect'. In Strange, W., editor, *Speech perception and linguistic experience*. York Press, Timonium MD.

Lacerda, F. (1995). The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. In *Proceedings of the Xlll international congress of phonetic Sciences, Stockholm*, volume 2, pages 140–147.

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

Ladd, D. R. (2015). *Simultaneous Structure in Phonology*. Oxford University Press, Oxford.

Lado, R. (1957). *Linguistics across Cultures*. University of Michigan Press, Ann Arbor.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Laver, J. (1994). *Principles of phonetics*. Cambridge University Press, Cambridge.

Lavie, N. and Hirst, A. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General*, 133(3):339–354.

Lecumberri, M. L. G. and Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *Journal of Acoustical Society of America*, 119:2445–2454.

Lehiste, I. (1969). *Suprasegmentals*. MIT Press, Cambridge, MA.

Lenneberg, E. H. (1967). *Biological Foundations of Language*. John Wiley & Sons, Inc., New York.

Levelt, Willem, J. M. (1989). *Speaking. From intention to articulation*. MIT Press, Cambridge, MA.

Levelt, Willem, J. M. (1999). Producing spoken language: A blueprint of the speaker. In Brown, C. and Hagoort, P., editors, *The neurocognition of language*, chapter 4, pages 83–122. Oxford University Press.

Li, A.-J., Jia, Y., Fang, Q., and Dang, J.-W. (2013). Emotional intonation modeling: A cross-language study on Chinese and Japanese. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, pages 1–6.

Li, X., Yang, Y., and Hagoort, P. (2008). Pitch accent and lexical tone processing in chinese discourse comprehension: an erp study. *Brain Research*, 1222:192–200.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6):431–461.

Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1):1–36.

Liberman, A. M. and Mattingly, I. G. (1989). A specialisation for speech perception. *Science*, 243:489–494.

Lingel, S., Pappert, S., and Pechman, T. (2006). The prosody of German pp-attachment ambiguities: Evidence from production and perception. In *Poster presentation at the 12th Annual Conference on Architectures and Mechanisms for Language Processing (AMLaP)*, Nijmegen, The Netherlands.

Liscombe, J. (2007). *Prosody and Speaker State: Paralinguistics, Pragmatics, and Proficiency*. PhD thesis, Columbia University.

Loftus, G. R. (1993). A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments, & Computers*, 25(2):250–256.

Lombard, É. (1911). Le signe de l'élévation de la voix. *Annales des maladies de l'oreille et du larynx*, 37:101–119.

Luce, P. A., Feustel, T. C., and Pisoni, D. B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 25(1):17–32.

Macmillan, N. and Creelman, C. (2005). *Detection theory: a user's guide*. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Macmillan, N. and Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, 98:185–199.

Maekawa, K. (2004). Production and perception 'paralinguistic' information. volume 03. Processing of Specch Prosody.

Mah, J. and Archibald, J. (2003). Acquisition of L2 length contrasts. In *Proceedings of the 6th Generative Approaches to Second Language Acquisition Conference*, pages 208–212, Ottawa, USA.

Major, R. C. (2008). Transfer in second language phonology. In Hansen Edwards, J. G. and Zampini, M. L., editors, *Phonology and Second Language Acquisition*, pages 19–39. John Benjamins, Amsterdam.

Massaro, D. (1972). Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review*, 79(2):124–145.

Masuko, S. and Kiritani, S. (1992). Nihongo gakushuusha ni okeru moora onso no shuutoku ni tsuite [acquisition of morae by L2 learners of Japanese. *Nihongo kyouikugakkai shunki taikai yokoushuu*, pages 19–24.

Masuoka, T. and Tabuchi, Y. (1992). *Kiso nihongo bunpou kaiteiban (Basic Japanese Grammer renewed)*. Kuroshio shuppan.

Matsuzaki, H. (1995). The way to show the Japanese prosodic features and their effects in language learning. *Journal of the Department of Japanese, Tohoku University*, 5:85–96.

McAllister, R., Flege, J. E., and Piske, T. (2002). The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian. *Journal of Phonetics*, 30:229–258.

McCawley, J. D. (1968). *The phonological component of Grammar of Japanese*. Mouton, The Hague.

McGuire, M. (2010). A brief primer on experimental designs for speech perception research. Technical report, UC Santa Cruz.

McLaughlin, B., Rossman, T., and McLeod, B. (1983). Second language learning: An information-processing perspective1. *Language Learning*, 33(2):135–158.

Mennen, I. (1998). *Second language acquisition of intonation: The case of Dutch near-native speakers of Greek*. PhD thesis, University of Edinburgh.

Mennen, I. (2004). Bi-directional interference in the intonation of Dutch speakers of Greek. *Journal of Phonetics*, 32(4):543–563.

Mennen, I. (2007). *Phonological and phonetic influences in non-native intonation*, pages 53–76. Mouton De Gruyter, Berlin & New York.

Mennen, I. (2015). Beyond segments: Towards a L2 intonation learning theory. In Delais-Roussarie, E., Avanzi, M., Herment, S., and Mennen, I., editors, *Prosody and Language in Contact*, chapter 9, pages 171–188. Springer, Berlin & Heidelberg.

Mennen, I., Chen, A., and Karlsson, F. (2010a). Characterising the internal structure of learners intonation and its development over time. *New sounds : International symposium on the acquisition of second language speech*.

Mennen, I., Schaeffler, F., and Docherty, G. (2008). A methodological study into the linguistic dimensions of pitch range differences between German and English. In *Proceedings of the 4th International Conferences on Speech Prosody*, pages 527–530, Campinas, Brazil.

Mennen, I., Schaeffler, F., and Docherty, G. (2012). Crosslanguage difference in f0 range: a comparative study of English and German. *JASA*, 131(3):2249–2260.

Mennen, I., Scobbie, J., De Leeuw, E., Schaeffler, F., and Schaeffler, S. (2010b). Measuring language-specific phonetic settings. *Second Language Research*, 26(1):191–215.

Menning, H., Imaizumi, S., Zwitserlood, P., and Pantev, C. (2002). Plasticity of the human auditory cortex induced by discrimination learning of non-native, mora-timed contrasts of the Japanese language. *Learning & Memory*, 9:253–267.

Mesgarani, N. and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397):10.1038/nature11020.

Michalsky, J. (2014). Scaling of final rises in German questions and statements. In *Proceedings of the 7th International Conference on Speech Prosody*, pages 978–982, Dublin, Ireland.

Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.*, 2:81–97.

Minagawa, Y. (1996). Sokuon no shikibetsu ni okeru akusento-gata to shiinshu no yooin - kankoku, tai, chuugoku, ei, supeingo bogowasha no baai -. In *Heisei 8 nendo nihongo kyooiku gakkai shunkitaikai yokoushuu*, pages 97–102.

Minagawa, Y. (1998). NNS ni yoru heisashiin no jikanseigyo - gengo rizumu no kotonaru bogo washa no hikaku - [nonnative speakers' production of closure durations - a comparison between L2 learner groups with different L1 backgrounds. *Nihon onseigakkai zenkoku taikai yokoushuu*, pages 103–108.

Minagawa, Y. and Kiritani, S. (1997). Hibogo no onin tairitsu no shikibetsu ni okeru onkyouteki tegakari ni tsuite - sokuon, hisokuongo no baai [the use of acoustic correlates for the identification of nonnative consonant length contrasts]. *Nihon onkyou gakkai kouen ronbunshuu*, pages 385–386.

Minematsu, N. and Hirose, K. (1995). Role of prosodic features in the role of prosodic features in the human process of perceiving spoken words and sentences in Japanese. *J. Acoust. Soc. Jpn.*, 16:311–320.

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

Morton, J. (1969). The interaction of information in word recognition. *Psychol. Rev.*, 76:165–178.

Munro, M. J. and Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1):73–97.

Munro, M. J. and Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, 38:289–306.

Muroi, K. (1995). Eigo bogowasha no nihongo no tokushuhaku no chikaku to sanshutsu ni okeru shomondai [problems of perception and production of Japanese morae—the case of native English sprakers]. *Sophia Linguistica*, 38:41–60.

Näätänen, R., Paavilainen, P., Rinne, T., and Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology*, 118(12):2544–2590.

Nabelek, A. K. and Donahue, A. M. (1984). Perception of consonants in reverberation by native and non-native listeners. *Journal of Acoustical Society of America*, 75:632–634.

Nagahara, H. (1994). *Phonological phrasing in Japanese*. PhD thesis, University of California, Los Angeles.

Nakagawa, C. (2001). "he" no ji gata intonation ni chuumoku sita prosody sidoo no kokoromi. [prosody instruction with the use of "he" shaped intonation]. *Journal of Japanese Language Teaching*, 110(140–149).

Namy, L. L., Nygaard, L. C., and Sauerteig, D. (2002). Gender differences in vocal accommodation:. *Journal of Language and Social Psychology*, 21(4):422–432.

Nazzi, T., Lakimova, I., Bertoncini, J., Frédonie, S., and Alcantara, C. (2006). Early segmentation of fluent speech by infants acquiring French: emerging evidence for crosslinguistic differences. *Journal of Memory and Language*, 54:283–299.

Nespor, M. and Vogel, I. (1986). *Prosodic phonology*. Foris Publications, Dordrecht.

Niebuhr, O. (2007). The signalling of German rising-falling intonation categories - the interplay of synchronisation, shape, and height. *Phonetica*, (64):174–193.

Nishihata, C. (1993). Heisajizokujikan wo hensuu to shita nihongo sokuon no chikaku no kenkyuu - NS to chuugokugo bogowasha no hikaku [study on perception of Japanese geminates with a variable of closure durations - comparison between native speakers and Chinese learners of Japanese. *NIhongo kyouiku*, 81:128–140.

Nishinuma, Y., Arai, M., and Ayusawa, T. (1996). Perception of tonal accent by Americans learning Japanese. In *Proceedings of the 4th International Conference on Spoken language Processing*, pages 646–649, Philadelphia.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, 43:25–53.

O'Brien, M. G., Jackson, C. N., and Gardner, C. E. (2014). Cross-linguistic differences in prosodic cues to syntactic disambiguation in german and english. *Applied psycholinguistics*, 35(1):27–70.

Ofuka, E. (2003). Sokuon /tt/ no chikaku: Akusento gata to sokuon hisokuongo no onkyooteki tokuchoo ni yoru chigai (Perception of a Japanese geminate stop /tt/: The effect of pitch type and acoustic characteristics of preceding/following vowels). *J. Phon. Soc. Japan*, 7(1):70–76.

Ofuka, E., Mori, Y., and Kiritani, S. (2005). Perception of a Japanese geminate stop: The effect of the duration of the preceding/following vowel. *Journal of Phonetic Society of Japan*, 9(2):59–65.

Otake, T., Hatano, G., Cutler, A., and Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, 32(2):258–278.

Oviatt, S., Levow, G. A., Moreton, E., and MacEachern, M. (1996). Modeling global and focal hyperarticulation during human-computer error resolution. *Journal of Acoustical Society of America*, 104(5):3080–98.

Pastore, R. E. (1987). Categorical perception: Some psychophysical models. In Harnad, S., editor, *Categorical Perception*, pages 29–52. Cambridge University Press, New York.

Pennington, M. and Richards, J. (1986). Pronunciation revisited. *TESOL Quarterly*, 20:207–225.

Petrone, C. and Niebuhr, O. (2014). On the intonation of German intonation question: the role of the prenuclear region. *Language and Speech*, 57(1):108–46.

Pfordresher, P. Q. and Brown, S. (2009). Enhanced production and perception of musical pitch in tone language speakers. *Attention, Perception, & Psychophysics*, 71:1385–1398.

Pierrehumbert, J. B. (2001). Exemplar dynamics: word frequency, lenition and contrast. In *Frequency and the Emergence of Linguistic Structure*, pages 137–158. John Benjamins, Amsterdam.

Pierrehumbert, J. B. and Hirschberg, J. (1990). *The Meaning of Intonational Contours in the Interpretation of Discourse*, pages 271–311. MIT Press, Cambridge.

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

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, 13:253–260.

Pisoni, D. B. (1975). Auditory short-term memory and vowel perception. *Memory & Cognition*, 3(1):7–18.

Pisoni, D. B. and Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, 15(2):285–290.

Porretta, V. J. and Tucker, B. V. (2015). Perception of non-native consonant length contrast: The role of attention in phonetic processing. *Second Language Research*, 31(2):239–265.

Prieto, P. and Roseano, P., editors (2010). *Transcription of Intonation of the Spanish Language*. Lincom Europa, Munich.

Prior, A. and MacWhinney, B. (2010). A bilingual advantage in task switching. *Bilingualism: Language and Cognition*, 13:253–362.

Qin, Z. and Mok, P. (2013). Discrimination of Cantonese tones by speakers of tone and non-tone languages. *Kansas Working Papers in Linguistics*, 34.

Quené, H. (2011). Software tools - adjustdurpitch.praat.

R Development Core Team (2008). *R: A language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabbitt, P. (1966). Recognition memory for words correctly heard in noise. *Psychonomic Science*, 6:383–384.

Ramsay, J. O., Hookers, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer Verlag, New York, NY.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis - 2nd Ed.* Springer.

Ramus, F. (2001). Outstanding questions about phonological processing in dyslexia. *Dyslexia*, 7(4):197–216.

Rasier, L. and Hiligsmann, P. (2007). Prosodic transfer from L1 to L2: Theoretical and methodological issues. *Nouveaux cahiers de linguistique française*, 28:41–66.

Repp, B. (1984). Categorical perception: Issues, methods, and findings. In Lass, N. J., editor, *Speech and Language: Advances in Bsic Research and Practice*, volume 10, pages 243–335. Academic Press.

Rimmele, J. M., Golumbic, E. Z., Schröger, E., and Poeppel, D. (2015). The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. *Cortex; a journal devoted to the study of the nervous system and behavior*, 68:144–154.

Ringbom, H. (1994). Contrastive analysis. In Asher, R. and Simpson, J., editors, *The Encyclopedia of language and linguistics*, volume 2, pages 737–742. Pergamon Press, Oxford.

Rodriguez-Fornells, A., Balaguer, R. D., and Munte, T. F. (2006). Executive control in bilingual language processing. *Language Learning*, 56:133–190.

Rosen, V. M. and Engle, R. W. (1998). Working memory capacity and suppression. *Journal of Memory and Language*, 39:418–436.

Sabol, M. A. and de Rosa, D. V. (1976). Semantic encoding of isolated words. *Journal of Experimental Psychology: Human Learning and Memory*, 2(1).

Sakamoto, E. (2010). *An investigation of factors behind foreign accent in the L2 acquisition of Japanese lexical pitch accent by adult English speakers*. PhD thesis, The University of Edinburgh.

Sasha, D., Asano, Y., Rohrdantz, C., Keim, D., B., B., and Butt, M. (2015). Self organising maps for the visual analysis of pitch contours. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*. Vilnius, Lithuania.

Scherer, K. R., Ladd, D. R., and Silverman, K. E. A. (1984). Vocal cues to speaker affect: Testing two models. *Journal of Acoustical Society of America*, 76:1346–1356.

Schneider, K., Dogil, G., and M´obius, B. (2011). Reaction time and decision difficulty in the perception of intonation. In *Interspeech 2011*, pages 2221–2224.

Schouten, M. and Van Hessen, A. J. (1992). Modelling phoneme perception: I. categorical perception. *Journal of Acoustical Society of America*, 92:1841–1855.

Schreck, T. (2010). *Visual-Interactive Analysis With Self- Organising Maps - Advances and Research Challenges*, pages 83–96. Intech.

Schwab, S. and Llisterri, J. (2011). Are French speakers able to learn to perceive lexical stress contrasts? In *Proceedings of the 17th International Congress of Phonetic Sciences. Hong Kong, China*, pages 1774–1777.

Scovel, T. (1988). *A time to speak. A psycholinguistic inquiry into critical period for human speech.* Newbury House, Rowley, Mass.

Searle, J. R. (1969). *Speech acts: An essa in the philosophy of language.* Cambridge University Press, Cambridge.

Seiler, G. (2009). Sound change or analogy? Monosyllabic lengthening in German and some of its consequences. *Journal of Comparative Germanic Linguistics*, 12:220–272.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, 10:209–241.

Shattuck-Hufnagel, S. and Turk, A. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, 25(2):193–247.

Shepard, R. N., Kilpatric, D. W., and Cunningham, J. P. (1975). The internal representation of numbers. *Cognitive Psychology*, 7:82–138.

Shibata, T. and Shibata, R. (1990). Akusento wa doo'ongo o donoteido benbetsu shiuruno ka?: Nihongo, eigo, chuugokugo no baai [how much can accent distinguish homophones?: Cases of Japanese, English and Chinese]. 17:317–327.

Shockley, K., Sabadini, L., and Fowler, C. (2004). Imitation in shadowing words. 66(3):422–429.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366.

So, C. K. (2010). Categorising Mandarin tones in Japanese pitch-accent categories: The role of phonetic properties. In *Paper presented at Interspeech 2010 Second language studies: Acquisition, learning, education and technology*, Tokyo, Japan. Waseda University.

So, C. K. and Best, C. T. (2008). Do English speakers assimilate Mandarin tones to English prosodic categories? In *Proceedings of Interspeech*, Brisbane.

So, C. K. and Best, C. T. (2011). Categorising Mandarin tones into listeners' native prosodic categories: The role of phonetic properties. *Poznań Studies in Contemporary Linguistics*, 47(1):133–145.

So, C. K. and Best, C. T. (2014). Phonetic influences on English and French listeners' assimilation of Mandarin tones to native prosodic categories. *Studies in Second Language Acquisition*, 36:195–221.

Sonu, M., Kato, H., Tajima, K., Akahane-Yamada, R., and Sagisaka, Y. (2013). Non-native perception and learning of the phonemic length contrast in spoken Japanese: Training Korean listeners using words with geminate and singleton phonemes. *Journal of East Asian Linguistics*, 22(4):373–398.

Stent, A. J., Huffman, M. K., and Brennan, S. E. (2008). Adapting speaking after evidence of misrecognition: Local and global hyperarticulation. *Speech Communication*, 50(3):163–178.

Stern, H. H. (1983). *Fundamental Concepts of Language Teaching*. Oxford University Press, Oxford.

Sternberg, S. (1966). High speed scanning in human memory. *Science*, 153:652–654.

Sternberg, S. (1975). Memory scanning: New findings and current ocntroversies. *Quarterly Journal of Experimental Psychology*, 27:1–32.

Stockwell, R., Bowin, J., and Martin, J. W. (1965). *The grammatical structures of English and Spanish*. University of Chicago Press, Chicago.

Strange, W. (2007). Cross-language phonetic similarity of vowels: Theoretical and methodological issues. In Bohn, O.-S. and Munro, M. J., editors, *Language Experience in Second Language Speech Learning: in honorof James Emil Flege*, pages 35–55. John Benjamins.

Strange, W., Akahane-Yamada, R., Kubo, R., Trent, S., and Nishi, K. (2001). Effects of consonantal context on perceptual assimilation of American English vowels by Japanese listeners. *Journal of Acoustical Society of America*, 109:1692–1704.

Studdert-Kennedy, M., Liberman, A. M., Harris, K., and Cooper, F. S. (1970). The motor theory of speech perception: A reply to lane's critical review. *Psychological Review*, 77(234–249).

Studdert-Kennedy, M., Shankweiler, D. P., and Pisoni, D. B. (1972). Auditory and phonetic processes in speech perception: Evidence from a dichotic study. *Cognitive Psychology*, 3(3):455–466.

Sukegawa, Y. (1999). Pitch realisation of 2/3-mora-words by Brazilian learners of Japanese. *Journal of Phonetic Society of Japan*, 3(3):13–25.

Sussman, E. (2005). Integration and segregation in auditory scene analysis. *Acoustical Society of America*, 117(3):1285–1298.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Szenkovits, G. and Ramus, F. (2005). Exploring dyslexic' phonological deficit I: Lexical vs sub-lexical and input vs output processes. *Dyslexia*, 11:253–268.

Tajima, K., Hiroaki, K., Amanda, R., Reiko, A.-Y., and Kevin G., M. (2008). Training English listeners to perceive phonemic length contrasts in Japanese. *The Journal of the Acoustical Society of America*, 123:397–413.

Takagi, N. (2002). The limits of training Japanese listeners to identify English /r/ and /l/: eight case studies. *Journal of Acoustical Society of America*, 111:2887–96.

Tamaoka, K., Saito, N., Kiyama, S., Timmer, K., and Verdonschot, R. G. (2014). Is pitch accent necessary for comprehension by native Japanese speakers? –an ERP investigation. *Journal of Neurolinguistics*, 27(1):31–40.

Taylor, P. (2000). Analysis and synthesis of intonation using the tilt model. *The Journal of the Acoustical Society of America*, 107(3):1697–1714.

Thomas, J. J. and Cook, K. A. (2006). A visual analytics agenda. *IEEE Comput. Graph. Appl.*, 26(1):10–13.

Toda, T. (1998). Nihongogakushuusza ni yoru sokuon, chooon, hatsuon no chikaku hanchuuka. *Bungei gengo kenkyuu gengo hen (tsukuba daigaku bungei, gengogaku kei)*, 33:65–82.

Toda, T. (2001). The influence of pronunciation instruction on accent perception. *Waseda University Center for Japanese Language Studies*, 14:67–88.

Toda, T. (2003). *Second Language Speech Perception and Production: Acquisition of Phonological Contrasts in Japanese*. University Press of America.

Tomaschek, F., Truckenbrodt, H., and Hertrich, I. (2011). Processing German vowel quantity: Categorial perception or perceptual magnet effect? In *Proceedings of the 17th International Congress of the Phonetic Sciences, Hong Kong*.

Trager, G. L. and Bloch, B. (1941). The syllable phonemes of English. *Language*, 17:223–246.

Turk, O., Schöder, M., Bozkurt, B., and Arslan, L. (2005). Voice quality interpolation for emotional text-to-speech synthesis. In *Proceedings of the 7th Interspeech*, pages 797–800, Lisbon.

Uchida, T. (1993). Characteristics of auditory cognition of long vowels and double consonants for Chinese students in learning Japanese language. *Japanese Journal of Educational Psychology*, 41:414–423.

Ueyama, M. (2000). *Prosodic Transfer: An Acoustic Study of L2 English vs. L2 Japanese*. PhD thesis, University of California.

Ueyama, M. and Jun, S.-A. (1998). *Focus realisation in Japanese English and Korean English intonation*, volume 7. CSLI/Stanford University Press.

Van Wijk, C. (1987). The PSY behind PHI: A psycholinguistic model for performance structures. *Journal of Psycholinguistic Research*, 16:185–99.

Vance, T. J. (1987). *An Introduction to Japanese Phonology*. State University of New York Press.

Venditti, J. (1997). Japanese ToBI labelling guidelines. *Ohio State University Working Papers in Linguistics*, 50:127–62.

Venditti, J. (2000). *Discourse Structure and Attentional Salience Effects on Japanese Intonation*. PhD thesis, The Ohio State University.

Walsh, D. L. (1993). Limiting-domains in lexical access: Processing of lexical prosody. In Dickey, M. and Tunstall, S., editors, *University of Massachusetts Occasional Papers in Linguistics 19: Linguistics in the Laboratory*, pages 133–155. GLSA, Amherst.

Wardhaugh, R. (1970). The contrastive analysis hypothesis. *TESOL Quarterly*, 4:123–130.

Warner, N. and Arai, T. (2001). Japanese mora-timing: A review. *Phonetica*, 58(1-2):1–25.

Watabe, S. and Hiratou, Y. (1985). Nionsetsugo ni okeru museiharetsuon to sokuon no handankyoukai to senkouboin no nagasa no kankei [relationship between perceptional boundaries of plosive geminates in dysyllabic words and durations of preceding vowels]. *Onsei gengo*, 1:1–8.

Wayland, R. P. and Guion, S. G. (2004). Training English and Chinese listeners to perceive Thai tones: A preliminary report. *Language Learning*, 54(4):681–712.

Werker, J. F. and Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics*, 37:35–44.

Werker, J. F. and Tees, R. C. (1984a). Cross-language speech perception: evidence for perceptual reorganisation during the first year of life. *Infant Behavior and Development*, 7:49–63.

Werker, J. F. and Tees, R. C. (1984b). Phonemic and phonetic factors in adult crossâĂŘlanguage speech perception. *The Journal of the Acoustical Society of America*, 75(6):1866–1878.

Wheeldon, L. (2000). Generating prosodic structure. In Wheeldon, L., editor, *Aspects of Language Production*, chapter 9, pages 249–274. Psychology Press.

Wichmann, A. (2000). The attitudinal effects of prosody, and how they relate to emotion. In *ISCA Workshop on Speech and Emotion*. Belfast, Northern Ireland.

Wiese, R. (2000). *The Phonology of German*. Oxford University Press, Oxford.

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., and Johnsrude, I. S. (2012). Effortful listening: The processing of degraded speech depends critically on attention. *The Journal of Neuroscience*, 32(40):14010–14021.

Willi, U. (1996). *Die segmentale Dauer als phonetischer Parameter von 'fortis' und 'lenis' bei Plosiven im Zürichdeutschen*. Steiner (=ZDL Beihefte 92), Stuttgart.

Wilson, A., Kato, H., and Tajima, K. (2005). Native and non-native perception of phonemic length contrasts in Japanese: Effects of speaking rate and presentation context. *The Journal of the Acoustical Society of America*, 1(117):2425.

Wilson, I. (2006). *Articulatory settings of French and English monolingual and bilingual speakers*. PhD thesis, University of British Columbia.

Wilson, I., Horiguchi, N., and Gick, B. (2007). Japanese articulatory setting: The tongue, lips, and jaw. In *Paper presented at UltraFest IV*. New York University.

Winters, S. and O'Brien, M. G. (2013). Perceived accentedness and intelligibility: The relative contributions of f0 and duration. *Speech Communication*, 55:486–507.

Wood, C. C. (1976). Discriminability, response bias, and phoneme categories in discrimination of voice onset time. *Journal of Acoustical Society of America*, 60:1381–1389.

Yang, R.-X. (2011). The phonation factor in the categorical perception of mandarin tones. In *Proceedings of the 17th International Congress of Phonetic Science (ICPhS XVll)*, pages 2004–2007.

Zhao, J., Guo, J., Zhou, F., and Shu, H. (2011). Time course of Chinese monosyllabic spoken word recognition: evidence from ERP analyses. *Neuropsychologia*, 49(7):1761–1770.

Zimmerer, F., Andreeva, B., Jügler, J., and M´obius, B. (2015). Comparison of pitch profiles of German and French speakers speaking French and German. In *Proceedings of the 18th International Congress of the Phonetic Sciences*, Glasgow, UK: the University of Glasgow. ISBN 978-0-85261-941-4. Paper number 183.

APPENDIX A

RAPID AND SMOOTH PITCH CONTOUR MANIPULATION

Introduction

The artificial manipulation of natural speech is common practice in the preparation of stimuli for perception experiments. A number of speech processing software tools, like the PSOLA (Pitch Synchronous Overlap Add Method) re-synthesis tool available in Praat (Boersma and Weenink, 2011), offer the possibility to modify the shape of F_0 and intensity contours extracted from recorded utterances, and to selectively alter segment duration. However, these tools investigate the perceptual effect of those phonetic features in isolation, e.g. by keeping the original F_0 contour and varying segment duration or vice versa. Changing prosodic parameters in isolation can give rise to stimuli that sound unnatural if constraints on the co-variation of the parameters are violated.

There are further operational limitations of the common practice. For example, in the case of gradual modification of F_0 contours between two reference shapes, the creation of intermediate shapes using default Praat tools requires two operations, namely (i) aligning corresponding segmental boundaries in time, and (ii) stylising the reference F_0 contours using straight line segments. The manipulation is carried out by changing the position of the junction points (usually only one) in the stylised contours using a graphical editor or a script (e.g. both options available in Praat). For example, in Yang (2011) artificial mixtures of two Mandarin tones are generated from a three-points stylisation of each tone and by gradually moving the point in the middle. Similarly, in Dombrowski and Niebuhr (2010) the shape of a pitch rise is changed from concave to convex by imposing a three-points stylisation and by varying the middle point height, and in Ambrazaitis and Niebuhr (2008) a modulation between two more complex shapes (dip and hat) is obtained by moving more than one point at the same time.

While segmental alignment is performed in order to preserve the anchoring of F_0 movements to the segmental material, F_0 contour stylisation is not necessarily justified by theoretical or experimental reasons. On one hand, stylisation may help in reducing the complexity of a prosodic model by isolating simple shape features. On the other hand, there is always the risk of losing important detail, e.g. the type of curvature (concave or convex) of a rising gesture (Dombrowski and Niebuhr, 2005). Moreover, stylisation is carried out manually, which entails empirical judgement and time consuming procedures.

While full F_0 contour grafting does not necessarily involve stylisation, which is sometimes applied nonetheless (e.g. in Niebuhr, 2007), segmental alignment remains a requirement. Ideally, the time warping involved in the alignment should alter the utterance structure as little as possible, in order to minimise the risk of introducing unwanted perception effects. Praat provides manual editing facilities for the selective manipulation of segment duration, and scripted procedures are also available, like Quené (2011) or the one used in Boula de Mareüil and Vieru-Dimulescu (2006). To my knowledge, the available scripts do not take into account possible effects on parameters that are linked to segment duration, which may cause noticeable discontinuity effects.

The contour manipulation method presented in this chapter eliminates the limitations described above. First, the proposed method makes it possible to vary parameters in combination, and to import or to "transplant" one or more speech features from other stimuli. Second, it implements smooth deformation of F_0 and intensity contours in order to align them to given segmental boundaries. Third, it provides a transparent way to combine two or more contours in desired proportions, e.g. for the creation of hybrids or for averaging.

The procedures are automatic and controlled by the user through a few parameters, like the degree of "elasticity" of contour deformation. This method improves the experimental procedures involved in stimuli manipulation in two ways. First it minimises the risk of introducing unwanted alterations in the original speech samples. Second, it provides for the automatic generation of stimuli in large quantities, enabling the rapid examination of several experimental conditions at design time, and the auditory assess to remove unnatural sounding tokens select the best ones.

The rest of the chapter is structured as follows. First, the method is described in its main principles, which are adapted from techniques applied in Functional Data Analysis (Ramsay and Silverman, 2005). After this, a brief general description of the software tools is given that carry out all the necessary operations. Then, the use case in this thesis

is reported in which the method has been applied to manipulate and generate stimuli. Finally, I draw conclusions.

Method

Smoothing

The basic principles of the proposed method are illustrated by referring to the manipulation of F_0 contours; similar considerations apply for intensity contours. The first operation to be carried out on the input data is called *smoothing*, which transforms a sampled F_0 contour into a continuous curve represented by a mathematical function of time $f(t)$. This function is constructed by combining a set of so called basis functions such that the combination fits the sampled data. In the case of features like F_0, whose contours in time can assume arbitrary shapes and do not present periodicity, it is customary to adopt B-splines as basis (De Boor, 2001). An example of smoothing is shown in Figure A.1. The user can control the degree of smoothing through a number of parameters (see Section Software).

Once contours are represented by functions of time, expressing combinations of them becomes trivial. For example, to create hybrids between two base contours A and B, the arithmetic operation $(1-\alpha)\cdot f_A(t)+\alpha\cdot f_B(t)$ produces the desired combinations in proportions controlled by the parameter α.

Landmark registration

Suppose A and B are realisations of the same three-syllabic word. The operation $(1-\alpha)\cdot f_A(t)+\alpha\cdot f_B(t)$ defined above combines values of contours A and B at corresponding points in time. However, the inevitable segment duration differences between the two realisations would mix shape traits belonging to different syllables, which would blur the timing relation between F_0 movement and segmental content.

This problem is solved by a convenient transformation applied to the time axis that alters each contour in such a way that corresponding segmental boundaries get aligned in time. This operation, called *landmark registration*, is carried out automatically and it is based on the position of each boundary (landmark) on each of the input contours. The time warping carried out by landmark registration guarantees that the qualitative aspects of the curves are not altered. Moreover, the local speech rate alterations are spread gradually throughout the entire contour, which minimises discontinuity effects. Figure A.2

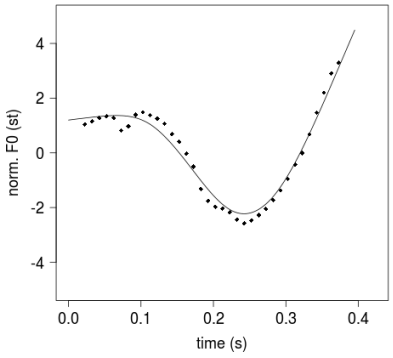

Figure A.1 *Example of a smoothed F_0 contour. Dots represent F_0 samples obtained from the pitch tracker available within Praat. The curve is a B-spline. This contour is extracted from a realisation of a three-syllabic word. The y-axis reports F_0 values in semitones after the global mean value was subtracted (corresponding to the zero level).*

shows the effect of landmark registration on an F_0 contour extracted from a three-syllabic word, where the syllable boundaries are shifted to a desired position.

Manipulation and re-synthesis

In this subsection, it is illustrated how to combine smoothing and landmark registration in order to create of a set of stimuli whose F_0 contours are combinations of two base contours A and B, extracted from two realisations of the same word or phrase in two different conditions (e.g. yes-no question vs. statement); analogous steps are required in other manipulation schemes.

First, F_0 contours are extracted from utterance A and B, and all relevant landmarks, like syllable boundaries, are marked (e.g. on a Praat textgrid). Then F_0 contours are smoothed and turned into functions, $f_A(t)$ and $f_B(t)$, which have different duration and whose landmarks are not synchronised yet. Suppose one wants to carry out re-synthesis on the recording of utterance A, which can be called as the *base* utterance. Before mixing F_0 contours of A and B it is needed to synchronise utterance B on the landmarks of the base. Thus, landmark registration is carried out on the boundaries of utterance B by imposing a time warp that aligns its boundaries on those of the base A. This is internally represented by a warping function $h_{B\to A}(t)$. After this, function $h_{B\to A}(t)$ is applied on $f_B(t)$ to obtain a different function $f_B(t_A)$, which has (qualitatively) the shape of $f_B(t)$ but

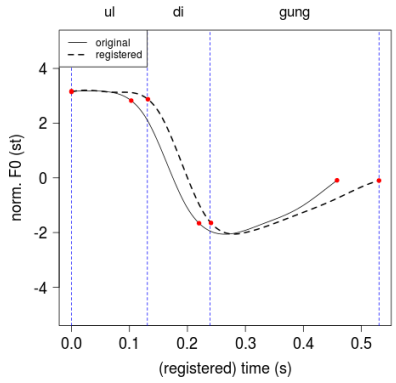

Figure A.2 *Example of landmark registration of a smoothed F_0 contour extracted from a realisation of a three-syllabic word. Dots show the position of syllable boundaries, vertical dashed lines the position where the boundaries are going to be shifted by landmark registration. The solid curve is the original F_0 contour, the dashed curve the contour after registration.*

is aligned with the landmarks of the base A. At this point one can create a number of mixtures of the form $f_\alpha(t) = (1-\alpha)\cdot f_A(t) + \alpha \cdot f_B(t_A)$, for instance, for $\alpha = 0, 0.2, 0.4, \ldots, 1.0$, where the value $\alpha = 0$ will produce a stimulus that should be identical to the original A and will be employed in the experiment in order to control for the re-synthesis effect, as well as being a useful sanity check for the re-synthesis. Finally, all the $f_\alpha(t)$ are converted into PitchTiers and used in Praat PSOLA re-synthesiser to modify the shape of the F_0 contour of the base utterance A.

Software

The software to carry out all the operations described above consists of a main R script (R Development Core Team, 2008) and a number of auxiliary R and Python scripts (www.python.org). The core functionalities are based on the `fda` library (Ramsay et al., 2009), with minor modifications. The software accepts Praat formats as input (e.g. TextGrids) and produces output also in Praat formats (e.g. PitchTiers) or wave files by calling Praat.

The main script is not intended to be executed in a single call, because the procedure is composed by a cascade of operations, some of which require the user to set a number of parameters, like those controlling smoothing, whose outcome has to be checked

by plotting (cf. Figure A.1). A simple expedient has been devised in order to alleviate the problem of having gaps in F_0 contours due to voiceless sounds. This is a hindrance when such a contour is transplanted on speech material where voiced sounds occupy the F_0 gap, as reported in Winters and O'Brien (2013). The input contour is smoothly interpolated by (automatically) padding extra samples at a level computed by averaging neighbour sample values.

Use Case

In this section the stimuli manipulation for my experiments in Chapter 3 to Chapter 4 is presented that was carried out using the method and the software introduced above.

The experiment investigated the influence of L1 on the discrimination of segmental length and pitch contrasts in an L2. A series of AX (same-different) discrimination tasks were designed.

Six non-sense disyllabic words were selected as the base stimuli that respected Japanese phonological structure and differed from each other in manner of articulation and voicing of the medial consonant. Each word was then created in three versions; triplets, which differed either in the duration of the first vowel or in the duration of the second consonant, resulting in a singleton (CVCV), a geminate (CVC:V) and a long-vowel (CV:CV) as counterparts (e.g., /punu/, /pu:nu/, /pun:u/).

All of the tokens were produced either with a lexical pitch accent on the first syllable (High-Low) or with no pitch accent at all (High-High). Each token was recorded six times by the same L1 speaker of Japanese, in order to present different tokens to the participants. Five segmental boundaries were considered: | C| V| C| V| , | C| V| C:| V| or | C| V:| C| V| .

AX tasks were designed to measure discrimination in one cue, either pitch or duration, while keeping the other constant. As the first step, the annotation of segmental boundaries of the stimuli was carried out using Praat, following standard segmentation criteria (Turk et al., 2005). More specifically, for plosives, closure duration (starting from a clear drop in the amplitude of the waveform and a drop in higher frequency energy, especially F2, in the spectrogram) was marked up to the release burst and for fricatives the duration of friction (as determined by the presence of aperiodic noise in the waveform). For nasals, measurement started when the amplitude in the waveform dropped and the waveform showed less high frequency components (drop in high frequency energy in spectrogram). As the second step, pitch contours of the stimuli were smoothed as

illustrated in Subsection Smoothing. Then, in order to keep segment duration constant across stimuli, landmark registration was applied as illustrated in Subsection Landmark Registration, where the 12 realisations of a given segmental pattern (e.g. six /pu:nu/ realisations for each of the two pitch patterns) were aligned on the average time location of those boundaries across realisations. Conversely, to keep pitch constant across stimuli in the duration contrast condition, a cascade of two landmark registrations was applied as follows. An average F_0 contour was created by first aligning the $N = 18$ contours $f_i(t)$, $i = 1, \ldots, N$ (i.e. six realisations times three duration patterns) on the landmarks of one of them, i.e., the first one, obtaining $f_i(t_1)$. This allowed to compute a time-aligned average contour $f_A(t_1) = \frac{1}{N} \sum_i f_i(t_1)$. This average shape was re-aligned on the segmental timing of each of the N realisations, obtaining N pitch-normalised contours $f_A(t_i)$, which were eventually imposed on the recorded stimuli. The naturalness of the stimuli was highly satisfactory in both manipulations. Moreover, F_0 padding was successfully used in order to accommodate for gaps due to voiceless segments (cf. Section Software).

Conclusions and future work

In this chapter, a method for the rapid and effective manipulation of F_0 and segmental duration values was presented aiming at the re-synthesis of stimuli for speech perception experiments. The method provides an automation layer between the level of specification of segmental alignment constraints and contour linear combinations on one hand, and the lower level provided by state-of-the-art editors, like the one available in Praat (PSOLA). The effectiveness of the method was illustrated by the use case, where it was successfully applied in real experimental conditions.

APPENDIX B

PARTICIPANTS' DEMOGRAPHICS (EXPERIMENT 2–4)

Experiment 2–4

	age (years)		sex		tone language	musical training
	range	mean	female	male		
Japanese L1 speakers	20-31	22.1	14	10	0	15
German non-learners	19-30	22.8	16	8	0	16
German L2 learners	20-34	25	18	30	11	23

Table B.1 *Overview of self-reported participants' information*

There was no bilingual participant. German L2 learners learned Japanese for between 6 and 156 months (mean learning duration: 38 months), 37 of them were students of Japanese Studies and they stayed in Japan for 0 up to 42 months (mean duration: 5.6 months). The L2-score of the participants ranged between 12 and 50 (mean score: 30.6, min. = 0, max. = 50).

APPENDIX C

RESULTS (EXPERIMENT 2–4)

Experiment 2

		flat pitch		falling pitch	
		ISI		ISI	
		short	long	short	long
Japanese	same	98.2	98.2	99.5	95.1
	consonant	98.6	99.0	98.6	98.6
	vowel	99.4	98.1	98.6	97.5
Learners	same	94.3	94.1	93.8	95.8
	consonant	83.3	87.1	84.4	88.0
	vowel	88.0	88.2	88.9	93.8
Non-learners	same	96.4	93.7	96.7	95.5
	consonant	87.4	85.3	79.2	76.4
	vowel	92.5	93.1	94.4	88.6

Table C.1 *Mean response accuracy (in %).*

		flat pitch		falling pitch	
		ISI		ISI	
		short	long	short	long
Japanese	consonant	5.36 (0.26)	5.43 (0.02)	4.94 (0.04)	4.93 (0.05)
	vowel	5.40 (0.18)	5.38 (0.22)	4.93 (0.06)	4.92 (0.07)
Learners	consonant	4.96 (0.85)	4.80 (1.00)	4.41 (0.83)	4.52 (0.81)
	vowel	5.14 (0.50)	4.95 (0.93)	4.71 (0.68)	4.78 (0.66)
Non-learners	consonant	4.91 (1.06)	4.20 (1.11)	3.93 (1.19)	3.92 (1.57)
	vowel	5.08 (0.67)	5.18 (0.62)	4.94 (0.04)	4.69 (0.61)

Table C.2 *Average d' scores and standard deviations.*

		flat pitch		falling pitch	
		ISI		ISI	
		short	long	short	long
Japanese	same	279 (233)	310 (266)	285 (366)	331 (377)
	consonant	341 (216)	331 (238)	334 (277)	373 (379)
	vowel	293 (235)	362 (328)	334 (399)	391 (453)
Learners	same	554 (351)	647 (454)	498 (419)	544 (418)
	consonant	918 (611)	881 (578)	913 (601)	953 (659)
	vowel	748 (564)	815 (573)	684 (608)	695 (625)
Non-learners	same	487 (307)	611 (470)	392 (357)	494 (420)
	consonant	765 (500)	823 (601)	799 (563)	912 (713)
	vowel	649 (533)	696 (499)	624 (558)	751 (644)

Table C.3 *Average RTs and standard deviations (in ms).*

Experiment 3

	same		singleton		geminate		long-vowel	
	ISI		ISI		ISI		ISI	
	short	long	short	long	short	long	short	long
Japanese	98.3	98.3	97.1	98.2	99.6	97.5	98.2	99.3
Learners	91.9	91.9	94.2	93.0	92.1	93.7	89.4	88.9
Non-learners	92.0	92.0	95.0	93.9	92.6	93.6	88.3	88.5

Table C.4 *Mean response accuracy (in %).*

	singleton		geminate		long-vowel	
	ISI		ISI		ISI	
	short	long	short	long	short	long
Japanese	4.79 (1.00)	4.69 (1.24)	5.10 (0.79)	4.42 (1.28)	4.74 (1.08)	4.87 (0.91)
Learners	3.87 (1.34)	4.01 (1.29)	4.16 (1.41)	4.18 (1.43)	4.17 (1.27)	4.24 (1.46)
Non-learners	3.89 (1.07)	3.68 (1.19)	3.80 (1.39)	3.60 (1.13)	3.98 (1.20)	3.94 (1.49)

Table C.5 *Average raw d' scores and standard deviations.*

	same		singleton	
	ISI		ISI	
	short	long	short	long
Japanese	312 (253)	350 (304)	398 (354)	419 (396)
Learners	558 (406)	659 (472)	707 (517)	794 (549)
Non-learners	482 (367)	566 (417)	562 (419)	645 (465)

	geminate		long-vowel	
	ISI		ISI	
	short	long	short	long
Japanese	290 (226)	320 (278)	348 (294)	357 (322)
Learners	618 (480)	712 (533)	762 (547)	801 (586)
Non-learners	480 (365)	572 (429)	528 (392)	601 (463)

Table C.6 *Average raw RTs and standard deviations.*

Experiment 4

flat pitch condition					
		immediate imitation		delayed imitation	
		segment duration		segment duration	
		short	long	short	long
Japanese	**consonant**	121 (44)	316 (51)	120 (40)	312 (23)
	vowel	103 (23)	321 (50)	107 (24)	328 (56)
Learners	**consonant**	132 (54)	270 (31)	133 (51)	270 (60)
	vowel	95 (21)	334 (55)	100 (20)	348 (53)
Non-learners	**consonant**	135 (50)	244 (62)	135 (48)	240 (58)
	vowel	90 (22)	339 (61)	98 (30)	346 (60)

falling pitch condition					
		immediate imitation		delayed imitation	
		segment duration		segment duration	
		short	long	short	long
Japanese	**consonant**	131 (49)	275 (59)	131 (50)	289 (56)
	vowel	106 (25)	311 (41)	111 (24)	324 (47)
Learners	**consonant**	140 (58)	269 (63)	141 (55)	266 (60)
	vowel	98 (25)	329 (47)	104 (26)	348 (53)
Non-learners	**consonant**	142 (52)	253 (71)	144 (50)	260 (33)
	vowel	95 (22)	326 (60)	100 (27)	342 (61)

Table C.7 *Mean segmental durations (in ms) and standard deviations for the productions in each experimental condition.*

	singleton		**geminate**		**long-vowel**	
	imitation condition					
	immediate	**delayed**	**immediate**	**long**	**immediate**	**delayed**
Japanese	155 (43)	155 (45)	128 (14)	127 (14)	140 (42)	145 (42)
Learners	139 (35)	139 (33)	98 (13)	99 (13)	121 (33)	122 (34)
Non-learners	136 (36)	139 (33)	101 (13)	103 (17)	123 (34)	122 (32)

Table C.8 *Average raw pitch slopes and standard deviations.*

flat pitch												
Singleton				Long vowel				Geminate				
g	u	b	u	g	uu	b	u	g	u	bb	u	
68	145	80	188	41	339	80	201	26	145	220	170	
pitch range 2.31				pitch range 1.30				pitch range 1.03				
g	u	n	u	g	uu	n	u	g	u	nn	u	
38	131	67	200	29	386	65	173	26	145	214	170	
pitch range 0.63				pitch range 1.22				pitch range 0.58				
g	u	p	u	g	uu	p	u	g	u	pp	u	
113	118	128	180	20	379	107	180	31	156	345	222	
pitch range 1.09				pitch range 2.04				pitch range 1.53				
p	u	n	u	p	uu	n	u	p	u	nn	u	
46	106	65	175	22	435	65	241	30	176	179	200	
pitch range 1.36				pitch range 1.02				pitch range 0.60				
s	u	f	u	s	uu	f	u	s	u	ff	u	
301	113	97	188	201	370	117	176	194	154	286	242	
pitch range 0.1.60				pitch range 0.94				pitch range 0.65				
z	u	s	u	z	uu	s	u	z	u	ss	u	
79	123	129	193	66	363	136	156	108	164	334	194	
pitch range 1.12				pitch range 2.61				pitch range 2.14				
falling pitch												
Singleton				Long vowel				Geminate				
g	u	b	u	g	uu	b	u	g	u	bb	u	
52	148	106	182	41	397	80	201	29	144	221	295	
pitch range 12.5				pitch range 13.0				pitch range 13.2				
g	u	n	u	g	uu	n	u	g	u	nn	u	
26	144	75	180	31	393	99	145	32	137	210	166	
pitch range 10.5				pitch range 11.3				pitch range 12.2				
g	u	p	u	g	uu	p	u	g	u	pp	u	
35	112	152	145	40	360	150	144	24	119	377	159	
pitch range 16.8				pitch range 14.5				pitch range 10.8				
p	u	n	u	p	uu	n	u	p	u	nn	u	
31	131	67	170	31	407	76	176	19	140	154	184	
pitch range 13.8				pitch range 14.9				pitch range 10.8				
s	u	f	u	s	uu	f	u	s	u	ff	u	
235	107	139	157	216	352	136	189	188	145	304	172	
pitch range 12.8				pitch range 12.7				pitch range 16.3				
z	u	s	u	z	uu	s	u	z	u	ss	u	
93	122	138	190	104	348	160	179	75	126	298	174	
pitch range 12.9				pitch range 13.5				pitch range 14.7				

Table C.9 Mean segment durations (in ms) and pitch range (in semitones) of the stimuli.

APPENDIX D

*M*ODEL SPECIFICATIONS IN THE STATISTICAL ANALYSES

```
# Experiment 2 (discrimination of segmental length contrasts)
# d' scores analysis
samedif.lmer1=lmer(dprime_c_v ~group*time*pitch+(1+time+pitch|vpnr), data =
data[abs(scale(resid(samedif.lmer))) < 2.5,])
# flat pitch condition
samedif.lmer1=lmer(dprime_c_v ~group*time+(1|vpnr), data=
data_HH[abs(scale(resid(samedif.lmer)))<2.5,])
# falling pitch condition
samedif.lmer1=lmer(dprime_c_v ~group*time+(1|vpnr), data=
data_HL[abs(scale(resid(samedif.lmer)))<2.5,])
# normalised d' scores (long - short ISI condition)
samedif.lmer1=lmer(long_short ~group*pitch+(1|vpnr),
data=data[abs(scale(resid(samedif.lmer)))<2.5,])
# flat pitch condition
samedif.lmer1=lm(long_short ~group,
data=data_HH[abs(scale(resid(samedif.lmer)))<2.5,])
# falling pitch condition
samedif.lmer1=lm(long_short ~group,
data=data_HL[abs(scale(resid(samedif.lmer)))<2.5,])
# RT analysis
rt.lmer1=lmer(rt_vc ~group*pitch*time+(1+pitch+time|vpnr),
data=data[abs(scale(resid(rt.lmer)))<2.5,])
# normalised RTs (long-short ISI)
rt.lmer1=lmer(rt_c_v_3_1 ~group*pitch+(1|vpnr),
data=data[abs(scale(resid(rt.lmer)))<2.5,])
# flat pitch condition
rt.lmer1=lm(rt_c_v_3_1 ~group, data=data_HH[abs(scale(resid(rt.lmer)))<2.5,])
# falling pitch condition
rt.lmer1=lm(rt_c_v_3_1 ~group, data=data_HL[abs(scale(resid(rt.lmer)))<2.5,])
# Experiment 3 (discrimination of pitch contrasts)
```

```
# d' scores analysis
samedif.lmerA=lmer(dprime~group+segment+(1+segment+ ISI|vp),
data=data[abs(scale(resid(samedif.lmer)))<2.5,])
# RT analysis
rt.lmerA=lmer(mean_rt~experiment*segment+segment*
group+(1+experiment+segment|vpnr), data=data[abs(scale(resid(rt.lmer)))<2.5,])
# experiment 4 (imitation experiment)
# analysis of consonant and vowel ratios
# vowel ratio
ratio.lmerA=lmer(relative_ratio~group*pitch+(1+pitch|vpnr), data=
data_v_contrast[abs(scale(resid(ratio.lmer)))<2.5,])
# consonant ratio
ratio.lmerA=lmer(relative_ratio~group*pitch+(1+pitch|vpnr), data=
data_c_contrast[abs(scale(resid(ratio.lmer)))<2.5,])
# analysis of pitch slopes
slope.lmerA=lmer(pitchslope~group*segment+(1|vpnr), data=
data[abs(scale(resid(slope.lmer)))<2.5,])
# geminate data
slope.lmer1=lmer(pitchslope~group*time+(1|vpnr), data=
data_g[abs(scale(resid(slope.lmer)))<2.5,])
# singleton data
slope.lmerA=lmer(pitchslope~(1|vpnr), data=
data_s[abs(scale(resid(slope.lmer)))<2.5,])
# long-vowel data
slope.lmerA=lmer(pitchslope~(1|vpnr), data=
data_l[abs(scale(resid(slope.lmer)))<2.5,])
```