TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

LEARNING OF SECOND LANGUAGE SOUND CONTRASTS

Effects of training on the perception and production of second language vowel and consonant duration

Antti Saloranta

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. B OSA – TOM. 589 | HUMANIORA | TURKU 2022

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

LEARNING OF SECOND LANGUAGE SOUND CONTRASTS

Effects of training on the perception and production of second language vowel and consonant duration

Antti Saloranta

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. B OSA – TOM. 589 | HUMANIORA | TURKU 2022

University of Turku

Faculty of Technology
Department of Computing
Phonetics
Doctoral Programme in Technology (DPT)

Supervised by

Professor Maija S. Peltola
Phonetics
Department of Computing
University of Turku

Reviewed by

Associate Professor, Natalia Kartushina
Faculty of Social Sciences
University of Oslo

Senior lecturer, Katja Mäntylä
Faculty of Humanities and Social Sciences
University of Jyväskylä

Opponent

Professor Emeritus, Heikki Lyytinen
Faculty of Education and Psychology
University of Jyväskylä

The originality of this publication has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

Cover Image: Essi Vastamäki

ISBN 978-951-29-8973-7 (PRINT)
ISBN 978-951-29-8974-4 (PDF)
ISSN 0082-6987 (Print)
ISSN 2343-3191 (Online)
Painosalama, Turku, Finland 2022

Dedicated equally to everyone who helped me get here, consciously or not.

UNIVERSITY OF TURKU
Faculty of Technology
Department of Computing
Phonetics
ANTTI SALORANTA: Learning of second language sound contrasts:
Effects of training on the perception and production of second language
vowel and consonant duration
Doctoral Dissertation, 188 pp.
Doctoral Programme in Technology (DPT)
May 2022

ABSTRACT

Listen-and-repeat training has been shown to be an effective tool for the learning of non-native phoneme contrasts. The aim of this thesis was to examine the learning of non-native duration contrasts using short-term listen-and-repeat training with adult learners. Learning results were examined with event-related potentials, behavioral discrimination and identification tasks, and production tasks. In addition, the learning of non-native vowel duration in a classroom setting was investigated.

Study I focused on listen-and-repeat training with articulation instructions in the learning of a non-native vowel quality contrast. The results showed that the participants were able to adjust their production of the contrast after only one training session. In Studies II and III, participants underwent two days of training with a non-native vowel duration contrast, and their performance was measured on trained, untrained and non-linguistic duration contrasts. Study II used behavioral discrimination and production tasks, and Study III additionally used the electric event-related potentials MMN and N1. Study II showed tentative training-related improvements in the discrimination and production of the trained contrast. Study III found increased MMN amplitudes, decreased N1 latencies and improved behavioral discrimination performance for the trained stimuli. Changes to general detection of acoustic duration contrasts were also seen in the elicitation of an N1 response for the untrained contrast. No training-related changes in production were observed. Study IV used a two-day training paradigm on stop and sibilant duration contrasts, measuring learning results with MMN, behavioral discrimination and a production task. No training-related improvement was observed, but a clear difference emerged between the consonant types, with sibilants proving less difficult for the participants than stops. Study V examined the learning of non-native vowel duration contrasts on an intensive language course. Similarly to the listen-and-repeat studies, perceptual identification performance improved, but no improvement was observed in production. Overall, these results suggested that perception of non-native vowel duration contrasts can be improved fairly quickly, and they are learned more easily than production. More research is needed regarding the learning of consonant duration.

KEYWORDS: second language acquisition; phonetic training; event-related potentials

TURUN YLIOPISTO
Teknillinen tiedekunta
Tietotekniikan laitos
Fonetiikka
ANTTI SALORANTA: Vieraan kielen äänne-erojen oppiminen: Harjoittelun vaikutus vieraan kielen vokaali- ja konsonanttikestoerojen havaitsemiseen ja tuottoon
Väitöskirja, 188 s.
Tohtoriohjelman nimi: Teknologian tohtoriohjelma (DPT)
Toukokuu 2022

TIIVISTELMÄ

Kuuntele ja toista –harjoittelun on osoitettu olevan tehokasta vieraan kielen äänne-erojen oppimisessa. Tämän opinnäytetyön tavoitteena oli tutkia vieraan kielen kestoerojen oppimista aikuisilla oppijoilla lyhyen kuuntele ja toista –harjoittelun avulla. Oppimistuloksia mitattiin herätevasteilla, behavioraalisilla erottelu- ja identifikaatiokokeilla, sekä tuottokokeilla. Lisäksi tutkittiin vieraan kielen vokaalikestoerojen oppimista luokkahuoneessa.

Osatutkimus I tutki artikulaatio-ohjeilla tuettua kuuntele ja toista –harjoittelua vieraan kielen vokaalilaatueron oppimisessa. Tulokset osoittivat, että osallistujat pystyivät muokkaamaan tuottoaan jo yhden harjoituskerran jälkeen. Osatutkimuksissa II ja III käytettiin kahden päivän kuuntele ja toista –harjoitusta vieraan kielen vokaalikeston oppimiseen. Oppimista mitattiin harjoitellulla, harjoittelemattomalla ja ei-kielellisellä kestoerolla. Osatutkimuksessa II käytettiin behavioraalista erottelua ja tuottokoetta, ja osatutkimuksessa III lisäksi MMN- ja N1-herätevasteita. Osatutkimuksessa II todettiin alustavia parannuksia harjoitellun eron havaitsemisessa ja tuotossa. Osatutkimuksessa III havaittiin harjoitellulla erolla kasvanut MMN-amplitudi, laskenut N1-vasteen latenssi sekä parantunut behavioraalinen erottelukyky. N1-vasteen elisitoituminen harjoittelemattomalle erolle viittasi myös yleisen akustisen erottelukyvyn paranemiseen. Tuoton suhteen muutoksia ei havaittu. Osatutkimuksessa IV käytettiin kaksipäiväistä harjoittelua vieraan kielen klusiili- ja sibilanttikestoerojen harjoitteluun. Oppimista mitattiin MMN:n, erottelukokeen sekä tuottokokeen avulla. Harjoittelu ei parantanut konsonanttien havaitsemista eikä tuottoa, mutta konsonanttien välillä havaittiin selkeä ero, jonka perusteella sibilantit olivat koehenkilöille helpompia kuin klusiilit. Osatutkimus V tarkasteli vieraan kielen vokaalikestoerojen oppimista intensiivisellä kielikurssilla. Kuten aiemmissa osatutkimuksissa, behavioraalinen havaitseminen kehittyi, mutta tuotossa ei havaittu parannusta. Kokonaisuudessaan tutkimukset viittasivat siihen, että vieraan kielen vokaalikestoerojen havaitsemista voidaan parantaa suhteellisen nopeasti harjoittelulla, ja se kehittyy helpommin kuin kestoerojen tuotto. Konsonanttien keston oppimisen suhteen tarvitaan jatkotutkimusta.

ASIASANAT: vieraan kielen oppiminen; foneettinen harjoittelu; herätevasteet

Acknowledgements

The full number of people who have, knowingly or unknowingly, taken part in this thesis process is impossible to tally. Obviously, everyone involved in the academic side, the research and the writing, is to be thanked, but it's the people outside academia that have been the ones to really help me withstand the intensely frustrating ebb and flow of the process. There are some people and institutions whose role has been particularly significant, and I'd like to now thank them as best as I can.

First of all, I would like to thank all of my funders, who in a very concrete way made this whole thesis possible. My sincerest thanks to the Alfred Kordelin Foundation and Turku University Foundation whose grants were essential in helping me focus fully on my research, particularly in the early stages of the thesis. Doctoral Program Utuling, Doctoral Program MATTI and the Department of Future Technologies also supported me in various ways, and of course I would like to thank my current Doctoral Program in Technology, as well as the University of Turku Graduate School.

My greatest thanks go to my supervisor, Professor Maija S. Peltola. Without your encouragement early on in my studies, I'd probably never have considered research as a career. You've (mostly) managed to keep me on track and focused on the work, and we've had quite a few laughs along the way as well. Not only would this thesis not exist without you, but it would also never have been finished. Thank you for all your guidance and support.

I would like to thank Associate Professor Natalia Kartushina and Senior Lecturer Katja Mäntylä for agreeing to review my thesis and for all their valuable comments. My thanks also go to Professor Emeritus Heikki Lyytinen for accepting the invitation to be my opponent, and to Professor Tapio Salakoski for being my research director.

All of my colleagues, past and present, at LAB-lab have of course been immensely important throughout the whole process. Maija S. Peltola, Katja Haapanen, Henna Tamminen and Kimmo Peltola have been and continue to be a daily source of laughs, academic support and lively discussions about everything under the sun. Thank you for putting up with my terrible jokes and the fact that my eagerness to rant about things seems to be inversely correlated with how interesting and important those things actually are. Thanks to Tomi Rautaoja, Elisa Reunanen,

Jemina Kilpeläinen, Laura Taimi and Katri Jähi for also being there throughout the years.

One person from academia deserves a special thank you, and that person is Leena Maria Heikkola. The last two studies in this thesis likely would not have happened in the same time frame, or at all, if it wasn't for you. Thank you for all the hours spent gathering data, poring over statistics, despairing over Reviewer 2's inability to read, and just spurring me on for the past few years. I look forward to all the work we'll be doing in the future.

Speaking of people without whom this thesis would not exist, my eternal gratitude to all my volunteer participants. Dozens of you dedicated hours of your time listening to strange things in the lab, but I assure you, it all made sense in the end. Thank you also to all of the phonetics students and other people who helped out in the lab, particularly with the EEG collection, and to Professor Paavo Alku for expertly crafting the stimuli for most of the studies in this thesis.

Various phoneticians, researchers and other academics have made me feel welcome in academia both as a researcher and as a person throughout the years. I'd especially like to thank Damon Tringham, Tommi Nieminen, Pertti Palo, Michael O'Dell, Minnaleena Toivola and Lisa Tulaja.

Dozens of people in "real" life have played a role in seeing this journey through, and I'm not going to even try to list all of you, but I'd like to especially thank Tommi Salovaara, Antti Riipinen, Alar & Tanja Kaskikallio, the Turku Whisky Club and Cosmic Comic Café for being some of the cornerstones of life in Turku. All your contributions to my mental health and overall well-being can't be overstated and again, it's unlikely I'd still be here with a finished thesis if it wasn't for all of you.

Thank you to my family: my mom Mari and to my siblings Maija, Leena and Ilkka, and all my aunts, uncles, cousins and in-laws. I feel very privileged to have the sort of family I have, and you've all been incredibly supportive.

Essi, you may have entered into the equation fairly recently, but you've made the last stretch of this thesis process and, consequently, my life so much better than I could've hoped for. Thank you so much for your endless joy, kindness and support every day. And of course I need to thank my sweet boys Piko for being there in spirit and Fermi for keeping me company and bringing joy to my days.

Turku, August 2022
Antti Saloranta

Table of Contents

Tables

Figures

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Saloranta, A., Tamminen, H., Alku, P. & Peltola, M.S. Learning of a non-native vowel through instructed production training. *Proceedings of the 18th International Congress of Phonetic Sciences*, 2015.

II Saloranta, A., Alku, P., Peltola, M.S. Learning and generalization of vowel duration with production training: behavioral results. *Linguistica Lettica*, 2017; 25: 67-87.

III Saloranta, A., Alku, P., Peltola, M. S. Listen-and-repeat training improves perception of second language vowel duration: Evidence from mismatch negativity (MMN) and N1 responses and behavioral discrimination. *International Journal of Psychophysiology*, 2020; 147: 72-82.

IV Saloranta, A., Heikkola, L-M, Peltola, M.S. Listen-and-repeat training in the learning of non-native consonant duration contrasts: influence of consonant type as reflected by MMN and behavioral methods. *Journal of Psycholinguistic Research*, 2022.

V Saloranta, A., Heikkola, L-M. Acquisition of non-native vowel duration contrasts through classroom education: perception and production affected differently. Accepted for publication in the *Journal of Second Language Pronunciation.*

The original publications have been reproduced with the permission of the copyright holders.

1 Introduction

Second language acquisition in adulthood poses several problems for learners, ranging from pronunciation mistakes to difficulties acquiring vocabulary to unfriendly attitudes from native speakers. At a phonetic level, the focus is typically on segmental errors, such as the widely studied problems native speakers of Japanese face when trying to differentiate between the English /l/ and /r/ in perception and production. However, learners also face issues when it comes to suprasegmental features, such as stress, intonation, and duration. The purpose of this thesis is to focus on one of these features, duration, and the ways in which its perception and production can be trained using short-term listen-and-repeat training, and how results acquired this way compare to more traditional language course education. The work draws from earlier studies using similar training, particularly those conducted by the Learning, Age & Bilingualism laboratory (LAB-lab) at the University of Turku, that successfully trained segmental features. The thesis begins with an overview of the use of duration in languages and how it can be problematic for second language learners. This is followed by some discussion on the language acquisition process and the effects the native language has on the second language learning process. This is followed by a discussion of models attempting to describe the problems associated with overcoming the native language effects, and various training methods that have been used in previous studies to train novel second language features, including the methodologies employed in the current work. The second half of the thesis describes the methods and results of the thesis, ending with a discussion and conclusions. To begin with, however, duration in languages and the problems it can present for second language learners will be discussed.

1.1 Duration in languages

Variation of the duration of individual speech segments is a phenomenon that exists in all the languages of the world to some extent. Individual segment duration is affected, for example, by the position of the segment in the phrase (e.g. Cho & McQueen, 2005; Fougeron & Keating, 1997) and the number of syllables in the word, with polysyllabic words exhibiting shorter vowel duration (e.g. White & Turk, 2010). Higher-level phenomena, such as stress, also affect segment duration, with stressed syllables typically showing longer vowel durations and unstressed syllables shorter ones (e.g. Braun, Lemhöfer, & Mani, 2011). In addition to vowel duration, consonant durations can also undergo variation, for example lengthening when the same consonant appears at the end of a morpheme and at the beginning of the following morpheme, or when consonants appear at the beginning of phrases (Cho & McQueen, 2005; Fougeron & Keating, 1997).

1.1.1 Phonological length

The type of duration variation being discussed in most of the studies of this thesis, however, is phonological length, i.e. the use of segment duration to differentiate meaning. Languages that employ phonological length contrasts are typically also known as quantity languages. Lengthening can occur in either vowels, such as in Japanese, Swedish, Finnish, or Czech, or consonants, such as in Italian, Japanese or Finnish. The extent of the quantity system can vary between languages: in Japanese and Finnish, for example, vowel duration can vary distinctively in any syllable, regardless of stress, while in Estonian duration is contrastive only in stressed syllables (Isei-Jaakkola, 2004; Meister, Nemoto, & Meister, 2015). The perceptual systems of the speakers of quantity languages are tuned to differentiate duration contrasts to a better degree than speakers of other languages (e.g. Ylinen, Shestakova, Huotilainen, Alku, & Näätänen, 2006). In Finnish, the context language for all of the duration studies comprising this thesis, vowels and consonants can show contrasting durations, and they are independent of each other and any stress phenomena (Suomi, Toivanen, & Ylitalo, 2008, p. 39). Long vowels are typically 2.2 to 2.4 times longer than short ones (Lehtonen, 1970; Wiik, 1965), and are phonetically very similar to short ones, with long vowels exhibiting slightly more peripheral formant values (Iivonen & Harnud, 2005; Wiik, 1965, p. 60). It is generally thought that native speakers also consider long and short vowels to be very similar, although O’Dell (2003) did find that categorization of the Finnish minimal duration pair tuli–tuuli (“fire–wind”) was affected by whether vowel qualities were extracted from the /u/ in tuli or the /u:/ in tuuli. It is unclear, however, to what extent this affects perception of quantity in everyday situations and whether similar quality differences exist for more open vowels as well; Iivonen & Harnud (Iivonen & Harnud, 2005) also found a clear difference between long and short qualities only for /u/.

1.1.2 Duration and second language acquisition

Phonological duration contrasts typically present problems for those second language learners of quantity languages that do not have contrastive duration in their L1, and the effect is the more noticeable the less important duration is in the L1 (McAllister, Flege, & Piske, 2002). Out of vowel and consonant durations, vowels seem to present fewer problems for learners (e.g. Bohn, 1995; Cebrian, 2006; Flege, Bohn, & Jang, 1997). Bohn (1995) found that adult German and Spanish learners of English were able to use vowel duration as to differentiate English tense-lax contrasts in cases where they were unable to rely on spectral differences. Similarly, Cebrian (2006) found that Catalan learners of English were also able to rely on duration to categorize an /i/-/I/-/e/ continuum, unlike English natives who categorized it mainly based on spectral differences. Finally, Flege et al. (1997) found that inexperienced Mandarin and Korean learners of English relied more on duration cues to distinguish English *bat-bet* and *beat-bit* continuums than experienced learners, who were able to make use of spectral differences. These findings that suggest the availability of vowel duration to those learners with no native duration contrasts were in fact formalized by Bohn into the Desensitization Hypothesis (Bohn, 1995), in which he posits that "whenever spectral differences are insufficient to differentiate vowel contrasts because previous linguistic experience did not sensitize listeners to these spectral differences, duration differences will be used to differentiate [the contrast]" (Bohn, 1995, pp. 294–295). The unavailability of spectral differences means, for example, a situation where the L1 of the learner only has the /i/ phoneme, into which the English /i/ and /I/ phonemes are both categorized, making them indistinguishable to the learner. It should be noted that this availability to detect duration differences can still be improved upon: experience in the L2 or training can improve the detection of non-native duration contrasts (e.g. Ylinen, Shestakova, Alku, & Huotilainen, 2005). Furthermore, learners with vowel duration contrasts in their native language typically fare even better at distinguishing L2 duration contrasts than those with no native duration. McAllister et al. (2002), for example, showed that Estonians, who have native vowel duration contrasts, were better able to identify Swedish short and long vowels than Spanish speakers, who do not use vowel duration contrastively.

As stated earlier, consonant duration contrasts seem to be more problematic for learners than vowel duration. Perhaps consequently, there is also less published literature regarding their L2 perception and production. Hayes (2002) tested different proficiency levels of English learners of Japanese on the perception of the Japanese /k-k:/, /t-t:/ and /s-s:/ duration contrasts in a same-different discrimination task of word-medial consonants. They found that proficiency in Japanese resulted in more accurate, but not native-like, performance. Hayes-Harb (2005) used a 13-step identification task to assess the perception of Japanese consonant duration contrasts by naïve English speakers and English learners of Japanese compared to native Japanese speakers. While the native Japanese displayed a clear category boundary, the identification curve for the naïve English was nearly linear, with neither end of the continuum fully recognized as short or long. The English learners of Japanese fared better, but not at a native-like level. Altmann (2012) compared native Italian and German speakers to advanced German learners of Italian in a same-different task of both vowel and consonant duration contrasts. Italian has consonant but not vowel duration contrasts, while the reverse is true for German. It was found that the Italians were easily able to perceive non-native vowel duration contrasts, while the German groups performed poorly in in perceiving non-native consonant duration contrasts, with the German learners of Italian faring slightly better than the naïve German group.

Taken together, these studies seem to suggest that vowel duration, but not consonant duration, is somewhat easily accessed by non-native speakers and learners of quantity languages. Crucially for this thesis, however, it seems that experience can improve the perception of at least vowel contrasts. In the following two sections, the acquisition process of the native language and its effect on second language acquisition and learning will be examined.

1.2 Acquisition of language and the DIVA model

The fact that humans perceive speech sounds categorically has been known since the mid-1950s. This was first empirically shown by Liberman et al. (1957), who found that listeners broke a synthetic continuum of vowel sounds into distinct categories that were separated by sharp transitions, rather than a gradual change. Furthermore, perception of sound differences that fall within a category was significantly worse than a difference of similar magnitude straddling a category boundary. This suggested that when listeners hear speech, speech sounds are classified based on some set of internal rules. The existence of phoneme categories has since become the basis of modern models of second language acquisition, and they have been extended to include information about both the perception and production of native speech sounds. They are, however, also the main reason for the difficulties experienced in second language acquisition in adulthood. In order to understand the reasons for this, and why training is needed, it is useful to briefly examine the process of native language acquisition in childhood.

1.2.1 Language acquisition in childhood

Children have traditionally been considered to be blank slates regarding language skills, although some recent neurological evidence has shown that newborns exhibit different overall brain responses (e.g. May, Byers-Heinlein, Gervain, & Werker, 2011) and lateralization patterns (e.g. Vannasing et al., 2016) to native vs. non-native language. This suggests that the ambient i.e. native language, heard during pregnancy, alters infants' language perception to some extent even before birth. Nevertheless, infants are typically able to distinguish speech sounds from all languages. Werker and Tees (1984) found that infants of six to eight months of age were able to distinguish between sounds from a variety of languages without any relevant experience with the languages in question. This ability had disappeared, however, at the age of one; at this age, the children performed similarly poorly to adults of their native language with no exposure to the languages being tested. Kuhl (1992) found that some of the perceptual reorganization necessary for segmental native language acquisition was already evident at six months of age. These findings suggested that infants were able to discriminate phonetic contrasts using general auditory processing mechanisms rather than any pre-existing linguistic ability; rather, children were committing existing neural resources to the detection of native language phonetic units (Kuhl et al., 2006). The loss of this processing ability is caused by perceptual warping, brought on by the detection of language-specific acoustic patterns and their statistical distributions and probabilities in ambient speech. This perceptual reorganization leads to a perceptual system that is fine-tuned to detect acoustic features relevant for understanding the speaker's native language and ignore those that are not needed (Iverson et al., 2003). This perceptual warping leads to the formation of phonetic prototypes, based on the most common sounds detected in ambient speech. Prototypes represent the ideal forms of native speech sounds and are used as the reference points to which phonetic categories are based on.

The existence of native prototypes has been shown, for example, for Swedish and English by Kuhl (1992), who tested six-month-old Swedish and American children on their perception of vowels on a continuum containing a prototypical example of a front vowel in both languages. For both languages, it was found that the children treated the other language's vowel as a non-prototypical exemplar of their native vowel. On the other hand, detection of features not relevant for the native language is diminished. Iverson and Kuhl (2003) tested adult speakers of Japanese on their perception of the English /r/-/l/ contrast by comparing monolingual Japanese speakers to native speakers of English in a syllable identification task. They were asked to identify syllables as either /ra/ or /la/, evaluate their goodness of fit in their category, and compare their similarity to the other stimuli. While the native English speakers clearly split the stimuli into two categories, displaying clear boundary effects and short perceptual distances for syllables judged as similar, no such effect arose for the native Japanese speakers. No boundary effects were found at all for the Japanese speakers, suggesting they perceived all stimuli to be members of the same phonetic category. This result demonstrates how perception is warped by exposure to the native language, resulting in the same acoustic information being processed in a completely different way by speakers of different languages.

1.2.2 The DIVA model

An attempt has also been made to model the speech acquisition process at a neural level. The DIVA (Directions into Velocities of Articulators) model (Guenther & Hickok, 2015) describes the native language acquisition process through the use of various neural feedback mechanisms, with a particular emphasis on explaining the formation and continued function of the speech production system. The model is based on the interaction of auditory, tactile and proprioceptive feedback and the roles they play in speech acquisition, and how the mechanisms are used to form an internal model of speech, containing information regarding both speech perception and production. Auditory feedback is the most straightforward of the three, consisting mainly of the heard vocalizations of the speaker, either through the ear or through bone conduction in the skull. As it can only be acquired after speech has been produced, its latency is quite high, on the order of hundreds of milliseconds. Tactile feedback consists of physical contact between different articulators, such as the tongue making contact with the hard palate during the production of certain palatal consonants. Proprioceptive feedback provides information about the positions of articulators by monitoring the shapes and positions of muscles, tendons and joints. The latency for the last two mechanisms, collectively known as somatosensory feedback, is very low, and can be used to correct speech production even at mid-utterance.

The DIVA model posits that the formation of the internal model of speech begins in early childhood as the child is exposed to ambient speech. This results in the development of perceptual representations of native language speech sounds. These representations are formed during the first year of life, at a stage where children do not yet produce speechlike sounds of their own. As the child begins to produce its first speechlike sounds in the babbling stage, the auditory system monitors the speech output and correlates it with the stored internal representations of native speech. Simultaneously, the somatosensory feedback mechanisms store the relative positions of the articulators used to produce each sound. As more and more sounds are produced, the more articulatory information is stored for each sound. The result of this process is a model containing detailed information about the articulatory pattern needed to produce each native sound. These targets are not necessarily positions for specific articulators, but rather entire configurations of the vocal tract, and they can contain a great deal of variability (Perkell et al., 1997).

The advantage of the model is that it allows the control of speech using so called feedforward commands that combine the strengths of auditory and somatosensory feedback (Perkell, 2012). As speech requires the simultaneous and coordinated function of dozens of parts of the vocal tract, it is not feasible to produce it by chaining together individual sounds in real time, as errors in output could only be detected through auditory feedback after they have been uttered. Feedforward commands are essentially plans that contain the vocal tract configurations needed to produce the desired acoustic goals in the required articulatory context. (Perkell, 2012, p. 401). This allows the monitoring of the plan's execution through the very fast somatosensory feedback mechanisms, instead of simply listening to the output. In this way, errors in speech production can be corrected almost mid-speech (Perkell, 2012, p. 394). Auditory feedback is used for long-term maintenance of the model: if the feedforward commands produce incorrect sounds despite being executed correctly, this is detected by auditory feedback and the commands are adjusted to again correspond with the native representations (Perkell, 2012, pp. 389–390). The motor patterns that arise from the process described by the DIVA model handily explain articulatory phenomena such as coarticulation: as the internal model allows for variability in the production patterns, the exact same configuration of the vocal tract is not used for every production of a specific speech sound. As long as the produced sound falls is in the correct phoneme category, which themselves contain acoustic variation, it does not matter in which way it was achieved. This means, for example, that the same consonant may be produced in a different way depending on what vowel follows it, resulting in measurable acoustic differences known as coarticulation phenomena.

The way in which the native language is acquired and affects perception and production has certain significant implications regarding second language acquisition. Because speech perception is warped and tuned to the native language, and the production system is based on internal perceptual representations of the native language's phonemes, second language sounds that are similar enough to fall into native phoneme categories will initially be treated like they were native sounds. Not only will they be heard incorrectly, but they will also be produced as if they were native. This means that they will most likely be produced wrong in most contexts, as the acoustic target the articulatory system is attempting to meet is based on the native prototype, with all the variation that phoneme category allows. This will result in accented speech. In order to produce second language sounds correctly, it is necessary to first form the internal perceptual targets, which can then be used as models for the correct vocal tract configurations. Some ways in which this may be achieved with the training provided in the current study will be examined in more detail in Section 1.4. First, however, a brief overview will be given to some prominent models of second language acquisition in order to examine how the learning problems induced by the previously described native language influence have been approached theoretically.

1.3 Models of second language acquisition

The existence of categorical perception and phoneme categories has long formed the basis of most modern models of second language acquisition[11]. The most relevant for this thesis are perhaps the Speech Learning Model (SLM) by Flege (Flege, 1995a; revised in Flege, Aoyama, & Bohn, 2021), and to some extent the Perceptual Assimilation Model (PAM) and its second language version, PAM-L2 (Best, 1995; Best & Tyler, 2007). In the following paragraphs, these models will be given a brief overview in order to illustrate how modern theories tackle the problem of native language-based warping of perception and other issues affecting the ability to learn the sounds new languages in adulthood.

1.3.1 The Speech Learning Model

The Speech Learning Model is a model of second language acquisition originally developed by James E. Flege in the early 1990s. It was spawned as a response to the mainstream views on language and second language acquisition of the early 1980s, particularly contrastive analysis and the critical period hypothesis that suggested that there was a sharp cutoff point in humans' ability to learn languages. Research had started to come out suggesting that this was not the case, and the Speech Learning Model sought to find the true factors and limitations to second language acquisition. The SLM treats second language acquisition as a single, dynamic system, where phoneme categories can shift and reorganize with exposure to new linguistic features (Flege et al., 2021, p. 64). The key factor in determining whether or not a novel second language sound will (eventually) be perceived correctly is whether or not a new category is formed for it. Category formation, according to the model, is dependent on the similarity of the new phoneme compared to the existing ones.

If it is too similar to native categories, it may get assimilated into one of them, leading to incorrect perception. Dissimilarity to existing phonemes, on the other hand, supports new category formation. Both assimilation and category formation

[1] It would perhaps be more accurate to refer to these as models of second language *sound* or *speech* acquisition, but the current naming was chosen, given the overall phonetics-based context of this thesis.

can lead to shifts in the native categories. If a novel sound is assimilated into a native category, the end result may be a category that is not fully representative of either the original or the novel phoneme's features (Flege et al., 2021, p. 64). This could happen, for example, when a native front vowel assimilates a more central second language vowel, resulting in a category that is more central than the native one but less so than the novel one. In the case of new category formation, native categories can shift in order to accommodate the new category and retain maximal discriminability. If in the previous example a category was formed for the novel central vowel, it could lead to the native category becoming more peripheral in order to facilitate discrimination. This category formation is driven by several factors, such as the phonetic similarity of the non-native sound being learned, the amount and quality of second language input received, and the state of the learner's native language categories, which act as reference points for category formation (Flege et al., 2021, p. 65). In the earliest versions of the SLM, the similarity of native and non-native phonemes was described using a three-category system, consisting of New, Identical and Similar. "New" represented non-native phonemes that did not resemble any native phonemes to the level that the same IPA symbol could be used. "Identical" phonemes, on the other hand, bore such a strong resemblance to some native category that no systematic differences could be found. Between them, the "Similar" sounds were those that closely resembled some native phoneme but differed from it in a systematic, measurable way (e.g. Flege, 1988). This classification, however, was abandoned in later versions as it is too difficult to determine which category each sound falls into. The revised SLM also incorporates the numerous findings that second language acquisition skills remain intact throughout adulthood (e.g. Flege, 1987; Flege & Eefting, 1987), and it suggests in its adoption of a "full access" hypothesis that the mechanisms used for the acquisition of the native language "remain intact and accessible for L2 learning" (Flege et al., 2021, p. 64). Given these findings, it seems entirely plausible that the adult participants of this thesis will be able to improve their perception and production of the non-native contrasts being trained, assuming the training is otherwise sound.

Given that the training method used in this thesis is listen-and-repeat, which simultaneously trains production and perception, the SLM's stance on the interplay of the two faculties is particularly relevant. The original version suggested that improvement in perception should typically precede improvement in production, formulated as "The production of a sound eventually corresponds to the properties represented in its phonetic category representation" (Flege, 1995a, p. 239), while the current version proposes that "segmental production and perception coevolve without precedence" (Flege et al., 2021, p. 64). Listen-and-repeat, as it is used in this thesis, provides a possibility to hear high-quality input for the new non-native contrast, immediately followed by an opportunity to practice its production. While the amount of input is admittedly low, it allows the participants to focus their attention fully on the contrast being trained and hone their perception and production of it simultaneously, which may support category formation, or at least begin the process of cue reweighing that is required for the correct detection of the new feature. From a difficulty perspective, the learning situations presented in this thesis are somewhat supportive of new category formation and therefore improved perception and production of the new contrasts. The most difficulty likely arises from phonetic similarity, which is extreme for most of the contrasts being trained. In the duration-focused studies, the contrasts are acoustically identical except for the duration of the target sound.

1.3.2 Perceptual Assimilation Model

The Perceptual Assimilation Model (PAM), by Catherine Best (Best, 1995), is a model originally developed to describe naïve non-native speech perception. The core of how PAM approaches the perception of non-native phonemes and sounds is its system of describing how non-native phones are placed, i.e. assimilated, into native categories. There are three main ways in which a sound can be heard: it can be either categorized, i.e. heard as good or poor version of a native phone, uncategorized, when it is considered unlike any native phonemes, or non-assimilated, when it is considered to be a non-linguistic and non-speech sound (Best, 1995, pp. 194–195). The assimilations are further described with patterns, which depict how easy two non-native sounds are to discriminate, based on how they fit into existing native categories (Best, 1995, p. 195). In a two-category assimilation, both sounds are assimilated into different categories, and are easily discriminated. In a category goodness difference, both sounds are assimilate into one category unequally, so that one more closely resembles the native prototype and the other deviates from it. This is thought to be somewhat difficult, depending on how large the difference between the non-native sounds is. In a single category assimilation both non-native sounds are assimilated into the same category, and are equally distant from the native prototype. Discrimination will be very difficult. The patterns for uncategorizable or nonassimilable sounds are not within the scope of this thesis, as the sounds used in the experiments are quite common and unlikely to remain uncategorized or unassimilable. PAM-L2 (Best & Tyler, 2007), the second language acquisition version, extends upon these themes by introducing the concept of phonological similarity on top of phonetic similarity that the SLM mainly deals with. This means that a second language sound that “has a similar contrastive relationship to surrounding categories in the phonological space (Best & Tyler, 2007, p. 24)” than a native sound may be considered to be similar, despite not being very close to it phonetically. This is the case, for example, when the French uvular /r/ is considered to be similar and equivalent to the English liquid /r/, despite there being little phonetic equivalence.

When it comes to the main contrast tested in this thesis, duration, the most likely assimilation pattern for a non-native speaker of a language with no quantity contrasts is a category goodness difference. The short member of a duration contrast is considered to closely resemble the native sound, while the longer member is a deviation, which, as per the pattern just described, would lead to a somewhat difficult learning situation. As per PAM-L2, this could eventually lead to the formation of a new phonetic and phonological category (Best & Tyler, 2007, p. 27) for the deviant member of the contrast, in this case a category for the long version of the sound. The more similar member is likely to assimilate to the native category quite well, and no separate new category is formed for it. For non-natives who do speak a quantity language, a two-category assimilation is also possible, in which case they would be easily able to discriminate the contrast (see Ylinen et al. (2005) for evidence of categorical processing of duration). This situation may present itself in Study V of this thesis, where a subgroup of the participants are speakers of a quantity language, and therefore already familiar with duration contrasts.

Overall, the models discussed in this section provide a good framework for the examination of the learning situations likely to occur in the studies comprising this thesis. Some of the predictions of the models will be revisited in the Conclusion section. The next section, however, will focus on the various training methods that have been used to overcome these difficulties and how well they have worked, with a special focus on the learning of second language duration contrasts and the use of listen-and-repeat.

1.4 Previous training studies

Different methods of training have been a mainstay of the study of second language acquisition for decades. The previously described models of second language acquisition and empirical findings related to them have shed some light on the learning capabilities of various groups of language learners through ambient learning in second language environments. Often, however, learning using natural acquisition is not realistic, due to time and location constraints, and more convenient and time-efficient training methods are desired. It is therefore of interest to examine methods with which the perceptual warping and associated production patterns caused by the native language can be modified in laboratory or classroom environments. This section will first introduce some of the most commonly used methodologies of perceptual and production training and results that have been achieved using them, and studies with the methodologies most closely resembling ones used in this thesis will then be examined in more detail.

1.4.1 Perceptual training

Training the perception of second language sounds and sound contrasts is perhaps the most common method in laboratory training studies. It is a theoretically reasonable approach, as it is perceptual categories that guide the formation of the correct production patterns in language acquisition (Flege et al., 2021; Guenther & Hickok, 2015). While not all findings and models support the view that improvement in second language perception must always precede improvement in production (Flege et al., 2021; Isbell, 2016), perceptual familiarity with the novel sounds of the language being learned is essential for understanding. The two main types of training are identification and discrimination training (e.g. Flege, 1995b), and both of these can also act as the method for measuring learning outcomes. In identification tasks, a stimulus is presented, and the participant is asked to classify it into one of the categories provided by the researcher, such as different vowel qualities or durations. In discrimination tasks, the participant typically hears two or more stimuli, and is asked whether or not they think they are the same or different (AX discrimination), or which one of two stimuli was the same as a third one (ABX/AXB discrimination). Furthermore, oddball discrimination refers to a task where the participant hears a continuous stimulus train and is asked to respond whenever they hear some sort deviation from a pattern, such as a non-native vowel in a chain of native ones.

Identification training tasks have been widely and successfully used to train, among others, vowel quality (e.g. Aliaga-Garcia & Mora, 2009; Carlet & Cebrian, 2015; Pederson & Guion-Anderson, 2010; Rato & Rauber, 2015; X. Wang & Munro, 1999; Ylinen et al., 2010), consonant quality (e.g. Aliaga-Garcia & Mora, 2009; Carlet & Cebrian, 2015; Lively, Logan, & Pisoni, 1993; Lively, Pisoni, Akahane-Yamada, Tohkura, & Yamada, 1994; Pederson & Guion-Anderson, 2010) and duration, for both vowels and consonants (e.g. Hirata, 2004; Hirata, Whitehurst, & Cullings, 2007; Okuno & Hardison, 2016; Tajima, Kato, Rothwell, Akahane-Yamada, & Munhall, 2008). Many studies also report generalization to novel speakers and/or contrasts. A particularly noteworthy and successful version of identification training, High Variability Phonetic Training (HVPT), gained popularity in the 1990s due to a series of studies in using it to train the English /r/-/l/ contrast to native Japanese speakers (Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999; Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Lively et al., 1993, 1994; Logan, Lively, & Pisoni, 1991). The methodology is effective in improving the perception of the /r/-/l/ category, with the improved skills generalizing to new talkers and untrained tokens (Lively et al., 1993) and transferring to production (Bradlow et al., 1997). Furthermore, the results achieved with the method were retained for at least three months (Bradlow et al., 1999). The downside of identification training is the large number of stimulus presentations and the long time needed for the training paradigm, with thousands of repetitions of the stimuli over several weeks (e.g. Bradlow 1999). This large time requirement may be difficult to fulfill in normal language teaching situations, making HVPT difficult to apply outside the research laboratory.

Discrimination training has not been used quite as widely as identification training, but there are still several studies that have successfully employed it for second language vowel and consonant contrasts (Carlet, 2017; e.g. Carlet & Cebrian, 2015; Flege, 1995b; Nozawa, 2015; Strange & Dittmann, 1984) and also second language duration contrasts (Menning, Imaizumi, Zwitserlood, & Pantev, 2002). Most of these studies employed a simple AX paradigm, where the participants had to decide whether the two stimuli they heard were the same or not. The training times were on average shorter than those used in most of the identification tasks, with 2-18 sessions of 30-90 minutes each, dependent on the study.

As stated, perceptual methods, particularly identification training, have been successful in improving perception and production of non-native duration contrasts, the main focus of this thesis. Most of these studies have been performed from a Japanese context, which is quite suitable for comparison with the Finnish context of the studies comprising this thesis, as the quantity systems of the two languages are quite similar. Hirata et al. (2007) used forced-choice identification training of Japanese vowel length contrasts at different speaking rates with adult English learners of Japanese. The training lasted for 11-17 days in four sessions and consisted of 540 naturally produced tokens. The training resulted in significant, though weak, perceptual improvement for the group that trained with the lowest speaking rate. No generalization was tested. Tajima et al. (2008) trained Japanese vowel and consonant duration contrasts to 19-25-year-old Canadian English speakers using minimal-pair forced-choice identification training with feedback. Training consisted of 15 35-60-minute sessions over 5 days, with 3600 training trials in total. The training improved identification performance for both vowels and consonants, but no generalization effects were found. Okuno (2014) used audio only and audiovisual training to train identification of Japanese vowel contrasts to English-speaking university students of various experience levels. The training consisted of eight sessions of forced-choice identification training with naturally produced disyllabic stimuli by native Japanese speakers with waveforms acting as the visual component of the audiovisual training. Both the audio only and audiovisual groups were able to improve their identification performance, and the training generalized to novel tokens and previously unheard talkers. Production accuracy also increased, despite there being no production training. Motohashi-Saigo and Hardison (2009) also used audio only and audiovisual forced-choice identification training to train Japanese consonant duration contrasts to beginner level English students of Japanese. The participants completed 10 sessions of self-paced training over two weeks, each session consisting of 120 stimulus presentations. The training tokens consisted of minimal triplets with singleton and geminate consonants and long vowels with the consonants /t, k, s/ and vowels /a, u/. The training resulted in improved perception of geminate consonants for both groups, with no change for the control group. The effects also transferred to novel stimuli and production skills, as assessed by native Japanese raters.

The study by Menning et al. (2002) is perhaps the most interesting, as it reports psychophysiological changes after discrimination training of Japanese vowel and consonant duration contrasts. German participants underwent 10 1.5-hour sessions of forced-choice AX discrimination training. The standard component of the AX training was a long sound, and the comparison stimuli were shorter versions of it, consisting of seven duration steps. The stimuli were originally naturally produced, and the shortened versions were artificially created with editing. Psychophysiological learning effects were measured with the magnetic equivalent of mismatch negativity, the mismatch field (MMF). Behaviorally, reaction times and discrimination accuracy were also measured. The results showed improvement both behaviorally and psychophysiologically: discrimination accuracy and reaction times both improved significantly, and the training was successful in enhancing the neural response, with a significant overall increase in MMF amplitudes and a decrease in MMF latency for two of the stimulus types.

Taken together, the results described above suggest that perceptual training can be successfully used to train non-native phoneme contrasts, and most importantly for this thesis, duration contrasts. Given, however, that the training methodology of this thesis employs both a perceptual and a production component, effects of production-based training will be examined next.

1.4.2 Production training

Production training studies are notably rarer in the second language training literature than those examining perceptual training. While purely production-based studies examining the learning of non-native duration contrasts were not found, production has been shown to be effective in training non-native contrasts in general. Here, a few examples of studies using production-based methodologies will be briefly examined.

An early study by Catford and Pisoni (1970) was able show that production training can be effective with native English participants. The study compared articulatory training with auditory training in the learning of “exotic” sounds, consisting of various consonant and vowel sounds not found in English. The articulatory training consisted of detailed instructions, in which the researcher first gave the participant a point of reference in the form of a familiar vowel, and then gave them guidance in how to move the articulatory organs from the familiar reference point to the configuration required for the new sound. This was done silently, i.e. no reference sound was given. This training was compared to auditory training, where the participants listened to the exotic sounds with frequent comparisons to more familiar ones. The results showed that the production training group not only improved more on their production of the novel sounds than the auditory group, but they also improved more in the perception of the novel sounds.

With more modern production training schemes, it is quite common to use visual aids or other kinds of feedback with the production task. Akahane-Yamada et al. (1998) used spectrograms as feedback in training the English /r-l/ contrast to 10 18-24-year-old native Japanese speakers with modified listen-and-repeat training. During training, the participants saw the spectrogram of their own productions side by side with the spectrogram of the model sound they heard, overlaid with formant tracks, and were able to compare their production directly to the model. The training resulted in significant improvement in intelligibility, as assessed by American English raters, coupled with a small but significant improvement in perceptual identification accuracy. Similar results were achieved by Kartushina et al. (2015) and Dowd et al. (1998), who used visual representations of articulator positions and a graph depicting mean resonances of the vocal tract, respectively, and were able to improve recognizability (Dowd et al., 1998) and accuracy (Kartushina et al., 2015) of non-native vowel contrasts. Kartushina et al. also found a small but significant transfer effect to the perception of the same vowel contrasts, using an ABX discrimination task.

Overall, the previously discussed literature shows that laboratory training can be effectively used to train both the perception and the production of non-native contrasts. In the next section, results from listen-and-repeat studies conducted by the Learning, Age & Bilingualism laboratory of the University of Turku will be examined in more detail, as this is the methodology used in the laboratory studies comprising this thesis.

1.4.3 Listen-and-repeat studies

Of the different types of production studies, the ones most relevant for this thesis are those that use listen-and-repeat training. Various such studies have been published in recent years, and I will now discuss some of them in order to shed light on the effectiveness of this type of training. Particular focus will be on studies by the Learning, Age & Bilingualism laboratory (LAB-lab) of the University of Turku, which largely use the same stimulus pair, with variations on training paradigms, participant populations and the methods used to measure learning results. The most commonly used stimulus pair consisted of the vowels /y/ and /ʉ/, embedded in the pseudoword pair /ty:ti/-/tʉ:ti/. The vowel /ʉ/ is found in Swedish, along with the /y/

and /u/ categories also present in Finnish. This stimulus pair was selected because it was thought it would "result maximal learning difficulties, as the vowel /ʉ/ is not phonological in Finnish but it is acoustically located near the categories /y/ and /u/ (Taimi, Jähi, Alku, & Peltola, 2014, p. 1231)". Furthermore, the stimulus pair was created using the Semi-synthetic Speech Generation method (SSG), which uses digital signal processing methodologies to model natural speech (Alku, Tiitinen, & Näätänen, 1999). This method produces stimuli that are acoustically exactly equal in their prosody, fundamental frequency, voice quality and other acoustic features, except during the target vowel where the words differ in their formant frequencies, while remaining otherwise as identical as possible. The creation of this particular stimulus set is explained in detail in Taimi (2014, p. 1231). This methodology was used in order to eliminate possible speaker-related artifacts and variation that could exist in fully natural stimuli produced by a real person.

The first of the LAB-lab studies, establishing the basic training and analysis protocol, was published by Taimi et al. (2014). They tested the effect of the training with 13 7-10-year old monolingual Finnish children, who had had very limited exposure to other languages. The children underwent four sessions of training over two days, with two sessions on each day. In each training session they heard the stimuli in an alternating long-short order for 30 times each, and after each stimulus they were asked to repeat it out loud as well as they could. Overall, this resulted in 120 repetitions of the long-short pair during the training sessions. In order to measure learning effects, each training session was paired with a recording session, with 10 repetitions of the stimulus pair. The first recording took place before the first training session, and the rest followed training sessions. The recordings were acoustically analyzed for their F1 and F2 formants in order to detect differences in the productions of /y/ and /ʉ/, and the formant values were subjected to statistical analysis.

The study found that as the training progressed, the children changed their production of the non-native /ʉ/ vowel, but not the native /y/. By the third recording session, i.e. the beginning of the second day, the average production of /ʉ/ was statistically significantly different from the pre-training production recorded on the first day. The change was due to the lowering of the F2 formant, making the vowel more central and thus more like the Swedish /ʉ/. The finding was in line with previous literature, suggesting that children had higher motor plasticity than adults and were able to quickly adapt to the production of the new vowel.

A follow-up to the previous study (Taimi, Alku, Kujala, Näätänen, & Peltola, 2014) measured the effects of the training on both the discrimination and production of the novel vowel in children. Five 7-11-year-old monolingual Finnish participants, returning from the previous study, underwent the same listen-and-repeat training paradigm as they had previously done. In addition to production recordings, however, they also were also tested on their preattentive and behavioral discrimination before and after training. The preattentive testing was performed by measuring their MMN responses using EEG. Both discrimination tests used an oddball paradigm. The stimuli were the same as in Taimi et al. (2014) and the same stimuli were used in all three tasks. While initial analysis of the behavioral results suggested that three of the five participants did manage to alter their production, increase their discrimination sensitivity and lower their reaction times, these results did not reach statistical significance. This was possibly due to the low number of participants. The EEG results, however, showed that an MMN response was elicited for the non-native deviant stimulus on the second day of the study, suggesting that the training affected the participants' preattentive discrimination.

Jähi et al. (2015) tested the effects of the same basic training paradigm on senior participants. The 20 monolingual Finnish participants, aged 62-73, were divided into a language group and a control group, based on whether or not they had studied languages during their retirement. The purpose was to investigate "whether participating in foreign language courses preserve[s] the ability to learn new production in elderly learners (Jähi et al., 2015)". The participants underwent the same training with the same stimuli as the children in the two previous studies, and they were tested on their production with analyses of formant frequencies and individual standard deviations of the formants. Formant frequency analysis did not produce any conclusive results, but analysis the individual standard deviations suggested that the productions of the non-native vowel became more consistent with training in the language group, but not in the control group. The change occurred after only two training sessions.

Peltola et al. (2015) changed the basic structure of the paradigm by adding visual hints to the training sessions. These were presented as either orthographic representations or phonetic transcriptions. The orthographic representations were purposefully misleading to the monolingual Finnish participants: the forms "*tyyti*" and "*tuuti*" were used for /ty:ti/and /tʉ:ti/, respectively. In Finnish, written "u" always represents the /u/ vowel, and the purpose of the study was to find out whether this misleading visual cue alters the production of the non-native vowel being trained. The 20 monolingual Finnish participants were between 20 and 34 years of age and were equally divided into orthographic and instruction groups. The training sessions were otherwise identical to the previous studies, but both groups saw their visual cues simultaneously with the sounds they were training. The results showed that both groups were able to alter their productions to be more like the target vowel, but the orthographic group changed more, suggesting that their productions became more /u/-like than the other group. This indicated that visual cues could "affect the outcome of production learning even more strongly than auditory information (Peltola et al., 2015, p. 4)".

Peltola (2017) tested the effectiveness of the unmodified paradigm versus a passive listening task with 18-32-year-old monolingual Finnish participants. The participants were again divided into two groups, the listen-and-repeat group and the passive listener group. The paradigm the groups underwent was otherwise identical, but the passive listener group only listened to the training stimuli instead of repeating them. The purpose of this was to examine whether or not learning new production requires explicit articulatory training, or whether production learning can occur even without practicing. No differences between the groups emerged in the statistical analysis of the formant values, or the individual standard deviation values. The standard deviations did, however, change during the training, suggesting that the productions of the two groups became more consistent as the training progressed. The change was apparent at the third recording, i.e. the second day of the experiment. Further analysis revealed that the change occurred in the non-native vowel and the native vowel remained unchanged. The findings were thought to imply that the simple act of listening to speech activates the motor patterns in the speech areas of the brain, supporting a strong connection between speech perception and speech production.

However, Peltola et al. (2020) compared listen-and-repeat training with active listening and observed quite a different pattern. A total of 22 18-38-year-old participants, divided into listener and producer groups, underwent a two-day paradigm, where the producer group listened to and repeated the training stimuli, while the listener group counted all instances of the new, non-native vowel and reported their findings to the researcher. The participants' learning outcomes were measured with a behavioral identification test and a production task. The results showed that for the producer group, changes occurred in the category boundary steepness for the identification task and the target vowel formant values and standard deviations in the production task. More importantly, neither of these changes were present for the active listener group. These findings were interpreted to suggest that the learning of novel speech sounds requires input from both the auditory and motoric domain, and that motor activation is of great importance for the acquisition of non-native speech sounds.

Besides the studies concerned with vowel quality, LAB-lab has also published research on the use of listen-and-repeat training with voice onset time contrasts (Tamminen & Peltola, 2015; Tamminen et al., 2015). These studies used a similar training paradigm as the studies discussed above, but the stimuli consisted of the English word pair /fi:l – vi:l/ ("feel"-"veal"), which is a minimal pair with regards to consonant voicing. Monolingual Finnish speakers took part in the studies, as consonant voicing is not distinctive in Finnish. Both studies utilized the same training paradigm, and measured learning results in both behavioral discrimination and identification and preattentive discrimination (with mismatch negativity). In

Tamminen & Peltola (2015) participants were Finnish university students of English, and in Tamminen et al. (2015) they were monolingual Finnish speakers with no advanced language studies, and in both studies they underwent four sessions of training over three days, consisting of 120 repetitions of the /fi:l – vi:l/ pair. Both studies reported increased sensitivity and decreased reaction times in the discrimination task, and changes in category boundary steepness in the identification tasks. In addition, stimulus goodness ratings changed in Tamminen & Peltola (2015), and the location of the category boundary changed in Tamminen et al (2015). Learning also occurred in the preattentive domain: an existing MMN response to the novel voicing contrast was further strengthened by the training in the former study, and an entirely new MMN response was elicited in the latter.

Taken together, these studies suggest that listen-and-repeat training can be effective in producing learning results for novel vowel and consonant contrasts. Learning results were observed over a very short time period in both perception and production of the novel contrasts, with significant results emerging in just two days at the fastest. The participants were able to improve their production consistency and produce new vowel qualities, and it also significantly improved their perception of the novel contrasts, with results visible in both the behavioral and the preattentive domains. In this thesis, similar methodologies will be applied to vowel and consonant duration contrasts, in order to determine whether this method is also effective with suprasegmental features.

2 Materials and Methods

2.1 The current study

The main purpose of the current study is to examine the learning of non-native duration contrasts through training and classroom instruction. Other objectives are the examination of generalization effects to untrained contrasts, the differences between learning in vowels and consonants, and comparing the results of listen-and-repeat training in the laboratory to learning results achieved in the language classroom. Finally, the study also provides an opportunity to extend the use of the methodologies used in earlier LAB-lab studies to previously untested contrasts. In the following paragraphs, the justification behind the selection of the training methods used in the thesis will be described in more detail, followed by an overview of the research questions that the work hopes to answer.

As stated in the Introduction section, variation of speech sound duration is a natural part of all languages, and some kind of systematic variation in duration is also very common, normally related to phenomena such as stress, speech rate and moraic structure (Altmann et al., 2012, p. 389). Phonological duration contrasts, however, are rarer, with duration change often being a secondary feature to some type of quality change. This is the case in the tense-lax oppositions of English, for example, where lax vowels are typically shorter in duration than their tense counterparts. In quantity languages, such as Japanese and Finnish, sound durations vary independently of other factors, and speakers of these types of languages are highly tuned to the detection of duration differences, with evidence of categorical and prototypical representations of duration (Ylinen, Shestakova, et al., 2005; Ylinen et al., 2006) and of separate processing mechanisms for phoneme quality and duration (Ylinen, Huotilainen, & Näätänen, 2005). According to current models of second language acquisition, such as the Speech Learning Model (Flege, 1995a) and Perceptual Assimilation Model (Best, 1995), this can present a significant learning problem for a student in whose native language duration is not phonologically relevant.

The DIVA model (Guenther & Hickok, 2015) suggests that perceptual representations are essential for the formation of the control system for speech, as they are what are used to create the motoric feedforward commands that are responsible for speech production. This means that in order for correct sounds to be produced, they first need to be learned perceptually and perceptual categories need be formed. Once this is done, the correct motor commands, and by extension feedforward commands, can then be formed, using the perceptual categories as production targets (Guenther & Bohland, 2002). Once the commands have been formed, they are controlled by the speech motor control system, and maintained through perceptual monitoring, i.e. speakers hearing themselves speak and modifying erroneous production until they match the perceptual categories. The DIVA model is focused on the acquisition of native language in childhood, but it seems reasonable to suggest that a similar process is at play in the learning of new languages as well. A comparable viewpoint is also put forward by Flege's Speech Learning Model, which suggests that correct production requires correct perception, which in turn means new category formation. The speech production system is therefore based on the interplay between speech perception and speech production, and it would make sense to employ training methods that activate both of these faculties. This would serve to simultaneously strengthen perceptual representations for the new speech sounds and provide the learner with an opportunity to use those representations to train the correct productions. This can be achieved with listen-and-repeat training.

The literature discussed so far regarding various training methods and, on the other hand, different types of non-native contrasts that were trained using them, show that laboratory training can be an effective tool in second language speech sound acquisition. Particularly good results have been achieved in the perception of novel non-native contrasts, typically using a variation of identification or discrimination training, with some learning extending to production as well. The trained features include non-native quality contrasts (e.g. Iverson, Pinet, & Evans, 2011; Kingston, 2003; Lively et al., 1993), voice onset time (e.g. Pisoni, Aslin, Percy, & Hennessy, 1982), lexical tones (e.g. Y. Wang, Jongman, & Sereno, 2003), and duration contrasts (e.g. Hirata et al., 2007; Okuno, 2014). Production-based training has seen much less use, though some examples of this were also discussed in the previous section (e.g. Catford & Pisoni, 1970; Dowd et al., 1998; Kartushina et al., 2015).

While production-based training is overall used much less than perceptual methods, a particular gap can be found in the literature regarding production training of non-native duration contrasts. Production learning of duration has been reported, but using perceptual methods (e.g. Motohashi-Saigo & Hardison, 2009; Okuno, 2014). Given the good results achieved by pure production training in other types of contrasts, as well as listen-and-repeat training in vowel quality and VOT contrasts, it seems warranted to examine the usefulness of production-based training. However, given the findings that a purely production-based approach does not produce learning results that are comparable to those achieved with perceptual training, an approach combining perception and production training could be the best approach, and as stated, listen-and-repeat training offers this combination.

As stated in the previous section, listen-and-repeat training has been used to great effect in studies conducted by LAB-lab in the training of vowel quality (e.g. Jähi et al., 2015; Peltola et al., 2015; Taimi, Alku, et al., 2014; Taimi, Jähi, et al., 2014) and voice onset time contrasts (Tamminen & Peltola, 2015; Tamminen et al., 2015). Learning results have been seen with psychophysiological event-related potentials, behavioral discrimination and identification, and production tasks. Listen-and-repeat training is used in four of the five studies comprising this thesis, and all of the four methodologies used in the aforementioned studies are used to gauge the learning results. This is done in particular to ensure maximal comparability of the results to the earlier studies by LAB-lab. The use of event-related potentials is particularly important, as it can shed light on the development of learning results from a very early stage in the process, where the learners themselves are not consciously aware of any changes having taken place.

The main ERPs that were measured in the current thesis were the N100 and the MMN. The N100 is an early component with a peak at around 100 ms after the onset of a stimulus or a change in its energy (Näätänen and Picton 1987). Its amplitude can increase when a stimulus deviates from an established sequence, for example a tone with a higher frequency than ones preceding it (Näätänen and Picton 1987, p. 388). Furthermore, Näätänen (1992) suggested that N1 amplitudes may be enhanced for "relevant stimuli" due to improvement in general sensory sensitivity (Näätänen 1992, p. 132). This feature may be useful in investigating auditory discrimination sensitivity. More crucially for this thesis, it has been shown that N100 responses can be affected by training (Brattico et al. 2003). Furthermore, Tremblay et al (2001) found a training-induced increase in the N1-P2 peak-to-peak latency, thought to reflect improved perception through changes in neural activity. The other ERP, the MMN, occurs later than the N100 at around 150-250 ms in the ERP complex. It is a response to "any discriminable change in the stream of auditory stimulation" (Näätänen et al. 2019, p. 1). It is elicited even when participants are performing an unrelated task and are not attending to the stimuli (Näätänen et al. 2007). What makes MMN particularly useful for examining second language acquisition is that it is language specific, with the same stimuli eliciting different responses for speakers of different languages (e.g. Näätänen et al. 1997; Ylinen et al. 2006; Chandrasekaran 2009; Dehaene-Lambertz 1997). Most importantly for this thesis, MMN responses can also be affected by training (e.g. Tremblay et al. 1997; Menning et al. 2002), and specifically listen-and-repeat training (Tamminen et al. 2015; Tamminen & Peltola 2015).

Both of the studies by Tamminen (Tamminen & Peltola, 2015; Tamminen et al., 2015) showed that listen-and-repeat training can be effectively used to improve preattentive perception of non-native contrasts, and the changes were also visible with behavioral perception methods. Neither of these studies, however, measured changes in production. The LAB-lab studies that did measure and report production changes largely did not report any psychophysiological results, positive or negative, with the exception of Taimi et al. (2014), who saw tentative changes in MMN responses in children who had undergone listen-and-repeat training of a non-native vowel contrast. These results, however, can only be considered preliminary due to the low participant count. Furthermore, none of these studies were conducted with duration contrasts as the training target. Therefore, studies that both a) use listen-and-repeat training accompanied with psychophysiological measurements and b) use the method to train duration contrasts are missing from the literature. This thesis aims to fulfill this gap in the literature, both to provide more information about the efficacy of listen-and-repeat and to extend the overall training literature on non-native duration contrasts to include production training results. Furthermore, the findings from listen-and-repeat training in the laboratory are compared to learning achieved in a language classroom environment. This will shed light on the timeframe necessary to achieve improvements in the learning of non-native contrasts: if similar results can be achieved in days in the laboratory that are achieved over weeks in the classroom, listen-and-repeat training can be considered to be a good supplement to language education for particularly difficult learning situations.

Based on a review of existing literature on the training of non-native phonological contrasts, the learning of non-native duration contrasts and in particular earlier studies conducted by LAB-lab, the following research questions have been formed for the current thesis:

1. Can non-native duration contrasts be successfully learned using listen-and-repeat, similarly to vowel quality and voice onset time contrasts?
2. Are the possible learning results limited to perception or production, or are both faculties affected by training?
3. Can duration be trained as a general process that the learner can apply to untrained speech sounds or non-linguistic sounds, or is it specific to certain phonemes?
4. Does the learning of duration differ between vowels and consonants? Or between different types of consonants?
5. How does listen-and-repeat training compare with an intensive language course with a communicative approach and no specific focus on pronunciation?

In order to shed light on these questions, five studies have been conducted. The participants of all studies are 18-35-year-old adults whose native languages have been controlled in advance to not contain phonological features relevant to the contrasts being trained, with the exception of Study V where they were divided into groups based on whether or not they had duration contrasts in their native languages. Study I, where vowel quality is the target contrast, can be considered a pilot of the methodology used in the other listen-and-repeat studies: it uses the same amount of training, but enhanced with production instructions, and learning outcomes are only measured from production. The rest of the studies are all focused on non-native duration contrasts, and use a wider variety of methods to assess learning. Studies II and III use a three-day training paradigm with vowel duration contrasts, where training takes place on the first two days with a final measurement of the outcomes on the third day, a week after the training is completed. Both studies measure outcomes using behavioral discrimination and production tasks, and in Study III the psychophysiological event-related potentials N1 and MMN are also used. In addition to the trained contrast, generalization of duration processing to untrained sounds is also studied. Study IV uses a two-day training paradigm with two consonant duration contrasts that differ in manner of articulation, the other one containing stops and the other sibilants. Learning outcomes are measured with event-related potentials (MMN and P3), behavioral discrimination and a production task. Finally, Study V examines the effects of an intensive four-week language course on the perception and production of several non-native vowel duration contrasts. Learning outcomes are measured with an identification task and a production task with read aloud sentences. Furthermore, the role of native phonological duration contrasts in the learning of non-native ones is a key focus of analysis.

In the following section, the various methodologies used in the studies comprising the thesis will be described in detail for each study. This will include descriptions of, for example, the participants, stimulus creation, test structures and statistical analyses. Next, results from each individual study will be described in detail, compared to the other studies, and briefly summarized together. Following this, in the Discussion section the findings from each study will be examined and discussed in relation to the research questions laid out above and any emergent phenomena not accounted for in the questions. Finally, in the Conclusion section an attempt will be made to draw all of the findings together to present a coherent view of the learning results gained from listen-and-repeat training, its effectiveness on the perception and production of different speech sound types, and how it compares to education received in the more traditional language classroom.

2.1.1 Overview of methodology

Assessing possible learning results in second language training studies can be done in numerous ways that offer different insights into the learning process. Depending on the research questions, in some cases it may be useful to employ native raters to assess the native-likeness of the participants' speech, while in other studies it may be desirable to remove human interpretation almost entirely with the use of more objective measurements. In this thesis, as the research is focused on the acquisition of segment-level phenomena that could have been difficult for human raters to accurately assess, several direct measurements of learning were used instead. The purpose was to assess the training-related development of representations for the novel features by using both preattentive and behavioral measurements. This allowed us to examine the progression of learning, as initial learning effects are often, though not always, seen in preattentive perceptual processes (e.g. Tremblay, Kraus, & McGee, 1998), moving on to behavioral effects and then production. In this chapter, the various methods used in this thesis will be presented in more detail, including a basic outline of the analyses. The statistical analyses used will be covered briefly, as they will be discussed in more detail in the Results section.

2.1.2 Participants

All of the studies in the thesis had a roughly similar participant profile. The overall purpose was to recruit people with little knowledge of phonetics or linguistics. This was done in order to reduce the possibility of the participants figuring out the purpose of the experiment, as this would have affected their decision-making and possibly their results. The participants of the studies are 18-30-year-old healthy, normally hearing, neurologically typical adults. In Studies I-IV the participants' native languages, or other languages they spoke well, did not contain the phonological features that were to be trained in the study they were taking part in. In Study I, this meant not having a close central rounded vowel in their native language, and in Studies II-IV it meant no phonological duration contrasts. They also could not have spent more than two months in a country where these features were used. In Study V, however, the participants' linguistic backgrounds were not controlled beforehand, and they were assigned to groups based on whether they had phonological duration contrasts in their native language. Unlike the other studies, a native control group of Finnish speakers was used in the perception and production tasks of Study V in order to determine the native-likeness of the participants' productions.

In Study I, the participants were recruited from the general student population of the University of Turku and were all monolingual Finnish speakers. For the other studies they were recruited from Finnish for foreigners courses, organized either by the Center for Language and Communication Studies of the University of Turku (studies II and III) or the Finnish department at Åbo Akademi (studies IV and V). Their linguistic backgrounds were highly varied, but they were carefully controlled to fulfill the previously stated requirements. All participants volunteered to take part in the studies with no monetary compensation, and they all signed written consent forms, consenting to the use of their data in these and possible follow-up studies. The Finnish participants were spoken to in Finnish, and all others in English.

In all of the studies, participants were asked to self-evaluate their language skills before the actual experiment started. They were asked to assess their overall skill level, frequency of use, and how often they heard the languages in the media. In Studies III and IV that employed EEG measurements, the participants were asked to fill in a questionnaire about their handedness using the Edinburgh Handedness Inventory (Oldfield, 1971) as only right-handed volunteers could take part in the study. This is done in order to minimize individual variation and ensure the homogeneity of the data (see e.g. Picton et al., 2000, p. 130); left-handed participants may exhibit different lateralization patterns for the ERP components being studied (e.g. Polich & Hoffman, 1998). The participants were required to be normally hearing, but in Studies I and V they were only asked to self-evaluate their hearing and confirm that they did not have diagnosed hearing loss. In the studies containing EEG measurements, however, the participants' hearing was tested with an audiometer, and in these studies the participants had normal hearing in the 100-4000 Hz range at 5-25 decibels.

As stated above, in studies I-IV the participants were all treated as a single experimental group, but in study V they were divided into two groups: those who spoke quantity languages ($n = 29$) and those who did not ($n = 39$), collectively called the quantity language groups. This was done in order to examine if speaking a quantity language other than Finnish aided them in acquiring Finnish quantity contrasts.

2.1.3 Stimuli

Due to the different focuses of interest in the various studies of this thesis, a number of different speech stimulus types was used in them. The three main types are semisynthetic stimuli, edited natural stimuli, and fully natural stimuli. The following paragraphs will detail their selection criteria, synthetization or recording, and properties.

Semisynthetic pseudoword stimuli were used in studies I, II and III. The semisynthetic method refers to a technique (Alku et al., 1999) where natural human glottal excitations, extracted from a real speaker, are used in combination with a digital vocal tract model in order to produce vowel stimuli. The advantage of the method is that it produces speech that sounds natural but whose phonetic features can be carefully controlled. Pseudoword, on the other hand, refers to words that follow the phonotactic rules of a specific language, in this case Finnish, but do not mean anything. All three studies used pairs of disyllabic pseudowords with a CV(V)CV structure. The phonetic features of the stimuli can be seen in Table 1.

Study I, where vowel quality was the feature being trained, used the pseudoword pair /ty:ti/-/tʉ:ti/. This particular pair was chosen as it represents a very difficult learning situation for Finnish speakers. In Finnish, the close front rounded vowel /y/ is contrasted only with the close back rounded vowel /u/. /ʉ/, however, is central, and therefore likely falls within one or both of the existing Finnish categories, /y/ or /u/. As per the models of second language acquisition discussed in Section 1.3 (PAM: Best, 1995; SLM: Flege, 1995a; Flege et al., 2021), this results in it being miscategorized as a poor exemplar of the native vowels, which results in a maximally difficult learning situation. Study I consisted only of a production task, and the stimuli were used in both the training and the recording phases.

Studies II and III used the same stimuli and focused on the learning of vowel duration and its generalization to new vowels. The pseudoword pairs used were /tite-ti:te/ and /tote – to:te/. The first pair was used as the training pair, and the second one as the generalization test pair. These vowels were chosen due to their different places of articulation, but also because vowels similar to them are among the 10 most common vowels in the world's languages (Maddieson & Disner, 1984, p. 125). They were therefore likely to be familiar to the multilingual participant group of the study. This was important in order to ensure that vowel quality did not pose additional difficulties on top of the duration contrast. All of the stimuli had an identical CVCV structure, and the linguistic stimuli were identical in aspects other than the target vowel. All three stimulus pairs were used in the EEG, behavioral discrimination, and production parts of the study, but only /tite – ti:te/ was used in the production training task.

Study IV used modified natural stimuli that were created from natural speech produced by a 31-year-old male native Finnish speaker. There were two pseudoword stimulus pairs, /tete/-/tet:e/ and /tese/-/tes:e/. /t/ and /s/ were chosen as the consonants to be tested, as they have the same place but different manner of articulation, allowing for comparison of the difficulty of duration processing between silence and frication. In order to ensure maximal similarity between the stimuli, apart from the manner of articulation, the following process was used to create them: "the /tes:e/ token was recorded, and its frication period was then adjusted to the desired length for the short member of the pair, /tese/. Then, /tet:e/ was created by removing the frication from /tes:e/, leaving only the formant transitions to and from it. Finally, the length of the gap was matched to the length of frication for /tese/, creating /tete/". This method resulted in stimuli that consisted of natural speech but were perfectly identical all other ways except for the target consonant. Within the first 170 ms all four words were identical, as the difference between them only started at that point. Both stimulus pairs were used in all parts of the task, including training, meaning that the participants were all trained with both stimuli, and learning effects were evaluated for both stimuli as well. The phonetic features of the stimuli can be seen in Table 1.

Table 1. Mean values of the stimuli in studies I-IV.

Word	Duration (ms)	Target segment duration (ms)	Mean f0 (Hz)	1. Vowel f1 (Hz)	1. Vowel f2 (Hz)
/ty:ti/	624	--	126	269	1866
/tʉ:ti/	624	--	126	338	1258
/tite/	392	154	110	330	2129
/ti:te/	428	194	110	330	2129
/tote/	392	154	110	452	805
/to:te/	428	194	110	452	805
/tese/	301	73	110	479	1663
/tes:e/	347	119	110	479	1663
/tete/	301	73	110	479	1663
/tet:e/	347	119	110	479	1663

While the previous studies used only 2-3 pairs of stimuli for the entire experiment, a much larger set was created for the identification task in Study V. It consisted of 50 pairs of duration minimal pairs with a CV(:)C:V structure, where the duration of the vowel in the first syllable was the feature that was varied. The vowels used in the stimuli were /y/, /æ/ and /ø/ and they were chosen due to their relative difficulty to many Finnish learners. The stimuli were created by first forming a list of all possible combinations with the aforementioned structure, with /l/, /m/, /n/, /r/, /s/ or /t/ as the initial consonants and the stops /t/, /k/ or /p/ in the word-medial position. This resulted in 74 minimal pairs, such as /syp:y/ - /sy:p:y/. All pairs containing meaningful Finnish words were then removed to avoid any effects of word recognition during the task. The final 50 pairs were recorded by a 33-year-old male native Finnish speaker. The only modification was that the average amplitude of each file was normalized to 65 dB. Mean values of all identification stimuli can be seen in Table 2.

For the production task in Study V, the stimulus list consisted of 60 sentences to be read aloud by the participants. The sentences were simple, three-word declarative sentences that contained a short or long /y/, /æ/ or /ø/ in the first syllable of one of the words, followed by a stop consonant.

Table 2. Mean values of the identification stimuli in Study V

	Short words	**Long words**
Word duration (ms) (stdev)	458 (30)	491 (35)
Vowel duration (ms) (stdev)	81 (12)	174 (15)
Long/short ratio, word (stdev)	1.07 (0.06)	
Long/short ratio, vowel (stdev)	2.18 (0.28)	

2.1.4 EEG recording and ERP analysis

All EEG recordings described in this thesis were performed in the same way. A Brain Products ActiCHAmp EEG recording system running on Brain Products Recorder software, version 1.20.0801. The system consists of 32 active electrodes on the scalp, and separate electrodes above and below the left eye for monitoring vertical eye movement, such as blinks. Horizontal eye movement was monitored with frontal electrodes F7 and F8, located at the sides of the head. The impedance of the electrodes did not exceed 10 kΩ. During all EEG recordings, participants were instructed to sit comfortably on an armchair while focusing on watching a film with no sound or subtitles. The stimuli were presented to them using Presentation software (version 16.3) by Neurobehavioral Systems and Sennheiser HD 25-1 II stereo headphones. Stimuli were presented in an oddball paradigm, where the probability of the deviant stimulus was 0.13, with 874 standard stimuli and 140 deviants and with an interstimulus interval of 650 ms. In both studies, short members of the stimulus pairs used acted as the standards, and the long member as deviants. Short members of the stimulus pairs were used as the standards (as well as in the discrimination task), as they were thought to be more phonologically familiar to the participants than the long ones, and therefore more likely to be perceived correctly.

EEG analyses were performed following the typical methodology used at LAB-lab (e.g. Tamminen et al., 2015). The recorded signal was offline referenced to the averages of the left and right mastoid electrodes, and then filtered with a 1-30 Hz bandpass filter. Artifact rejection sensitivity was set at ± 100 µV. Analysis epochs started at 100 ms before stimulus onset and ended at 500 ms after it. A baseline correction was then applied: for Study III it consisted of the 100 ms prestimulus period and for Study IV it consisted of the 100 ms prestimulus and the immediate 170 ms after stimulus onset that was identical in both the long and short stimuli, for a total of 270 ms. The first and second standard stimulus following each deviant was excluded from the analysis, as the change from deviant to standard could elicit ERPs that would distort the data. A separate average waveform was calculated for valid standard and deviant responses, and the standard responses were then subtracted from the deviants in order to create difference waveforms where the ERP amplitudes could be measured. Finally, 30 ms time windows were chosen for each stimulus and each ERP, with their center on the estimated peak amplitude observed in the difference waveform. In Study III, with three stimulus types and two measured ERPs, time windows were different for each one. For the trained linguistic stimuli, N1 windows were set at 195–225 ms and 225–255 ms, and MMN windows at 310–340 ms and 340–370 ms. Two consecutive time windows were used due to two observed amplitude peaks in the difference waveform for both responses. Single time windows were used for the untrained linguistic stimuli as no double peaks were observed: the N1 window was set at 220–250 ms and the MMN 330–360 ms. For the non-linguistic stimuli a single MMN window was set at 350–380 ms as no N1 response could be discerned. In Study IV, all stimuli used the same time windows as peak amplitudes were situated at roughly the same times: the MMN window was set at 360-390 ms and the P3 window at 425-455 ms. For the final statistical analyses, mean amplitudes from electrodes C3, C4, Cz, F3, F4 and Fz were used.

2.1.5 Oddball discrimination task

An oddball discrimination task was used in studies II, III and IV. The basic structure of the task was the same across all three studies: the short members of the stimulus pairs acted as the standards, and the long members as the deviants. There were 130 standards and 20 deviants, resulting in a deviant probability of 0.13. The ISI was 1000 ms. The task was presented diotically using Presentation software (version 16.3) by Neurobehavioral Systems and Sennheiser HD 25-1 II stereo headphones. Participants were instructed to press a response button as quickly as they could whenever they detected a change in the stream of stimuli. They were not told any specific details about what they would hear, apart from there being either words or sounds. No feedback or any other performance-related comments were given either during the task or after it. Detections and reaction times were the metric used to evaluate performance.

For analysis, mean reaction times for each correct identification of the deviants were recorded. Any reactions that were three standard deviations higher or lower than the mean were discarded. If a participant had not responded to any deviants, the stimulus onset asynchrony value, i.e. the difference between the beginnings of stimuli (1428 ms), was used for that particular block. A discrimination accuracy score, d’, was also calculated, based on the number of correct identifications of deviants, missed deviants, “false alarms” where the participant responded when there was no deviant, and correctly ignored standards. The formula used was d’ = z(H) - z(F), where H is the hit rate, calculated by dividing the number of hits by the number of deviants, and F is the false alarm rate, calculated by dividing the number of false alarms by the number of standards. If the number of hits or false alarms was zero for, the value 0.5 was used in order to avoid the d’ value becoming infinite (MacMillan & Creelman, 2005, pp. 6–9). The ceiling level for the score achieved in this way was 4.62, and a participant who did not respond at all would score 0.7. Mean reaction times and mean discrimination accuracy scores were calculated for pre- and posttest, and these were subjected to statistical analysis.

2.1.6 Identification task

A simplified identification task was used in Study V. Unlike in some identification tasks taking place in more typical laboratory settings, no category boundaries or goodness ratings were measured in the task, and the purpose was simply to calculate the number of correct identifications of either long or short stimuli, instead of a continuum. The stimuli were presented to all participants at once through loudspeakers in a lecture hall. They were asked to listen to the stimuli, and circle the word they think they heard on a form, where the short and long version (i.e. syppy or syyppy) was presented for each token. The long and short members of all stimulus pairs acted as the correct identification target twice, resulting in four presentations of each pair and 200 stimulus events in total. Stimulus presentation order was pseudorandomized so that members of the same pair never appeared consecutively, and they were delivered in 4 blocks of 50 with an ISI of 3 seconds.

For the analysis of the identification task, several different variables were formed from the participants’ responses, all consisting of the mean percentage of incorrect identifications. In order to examine overall identification accuracy in the pre- and posttest, variables containing responses for both the long and short stimuli pooled together, and then separately for only the short or long vowels were created. Then, variables were created for each vowel, again for long and short stimuli separately, in order to compare identification accuracy between the different vowels. All of the variables were created for both pre- and posttest, and they were used statistical analyses. A repeated measures ANOVA with a basic structure of Group(2) X Time(2) was used, where the Time variables were the pre- and posttest identification scores, and the Group variables were either the two language proficiency groups or the two quantity language groups.

2.1.7 Production task and production training

Two different types of production tasks were used in the studies of this thesis. Studies I-IV had a listen-and-repeat task, whereas in Study V participants read aloud sentences from a list. The listen-and-repeat task again followed the same basic structure in all of the studies it was used in. The stimuli were presented diotically with a Sanako SLH-07 headset and Sanako Lab 100 language lab software and hardware. The interstimulus interval was 3 seconds, and all stimuli were presented in an alternating pattern, either /ʉ:ti/-/ty:ti/ in Study I or "short-long" in studies II-IV. Participants were instructed to carefully listen to what they heard, then repeat to the best of their abilities in a normal, calm voice. Listen-and-repeat was used as both the production training method and the production task in all of these studies; the only difference between them was the number of stimuli and that the production tasks were recorded for analysis. In the training blocks, the participants heard and repeated each stimulus pair 30 times, and in the recording blocks 10 times. Four training blocks were used in each study for a total of 120 repetitions of each trained pair.

In Study I, each training block was preceded by instructions meant to help the participants modify their production of the target vowel. They were designed so that on the first day they provided a linguistic context and increased awareness of the different vowels in the stimuli, and on the second day they were articulatory advice for fine-tuning the production of the non-native vowel. The instructions were as follows (Saloranta, Tamminen, Alku, & Peltola, 2015, p. 3, translated from Finnish):

1. "1. You will hear two words alternately. The other one has a Finnish vowel, and the other has a Swedish one.
2. The Swedish vowel can be described as a mixture of the Finnish "y" and "u".
3. Try keeping your mouth otherwise in the same position as you do for /y/, but move your tongue slightly back in your mouth.
4. There are minor differences in the roundedness of the lips. The lips are more tightly rounded in /y/ than they are in /ʉ/."

In studies II-IV, instructions were only given related to the performance of the task, not the stimuli themselves. No feedback or any other performance indicators were provided in any of the studies.

The number of recordings differed between the experiments. Study I used four recordings, situated between the training blocks. In studies II and III, the trained linguistic block was recorded three times: at pretest, immediately after all training blocks were completed and on the third day of the experiment. The untrained linguistic block was recorded once on the third day. In Study IV, both stimulus pairs were recorded once at pretest and once at posttest. All training and recording blocks were performed in a sound attenuated laboratory.

In Study V, an entirely different production task was used. Rather than listening and repeating, the participants read aloud a list of 60 sentences. Participants were instructed to read each sentence aloud at their own pace in a normal voice, but without rushing. The recordings were performed with Audacity (2.3.2) software either in an acoustically treated studio using a Røde Podcasting microphone, or in a conference room using a Zoom H2n microphone. Each participant was recorded once at pretest and once at posttest.

As Study I was focused on vowel quality, and studies II-V on segment duration, acoustic analysis of the recordings was different between them. In Study I, each individual word produced by the participants was analyzed for formant values F1 and F2 at the midpoint of the first vowel. Mean formant values were then calculated for both vowels in all sessions, and these were used in statistical analyses.

In Studies II-IV, each production was analyzed for total word duration and the duration of either the first syllable vowel or the word-medial consonant, depending on the focus of the study. In order to minimize differences caused by variations in speaking rate, the durations were normalized by dividing the values of the repetitions of the long stimuli by the values of the repetitions of the short ones. This resulted in ratios that described how much longer the long words or vowels were in comparison to the short ones. A ratio of 1.0, for example, would mean that the participant produced long and short sounds with equal durations, while a ratio of 2.0 would mean that long sounds were twice as long as the short ones. The ratios were used in the statistical analyses.

In Study V, 12 sentences were first selected for analysis. Six of the sentences contained short exemplars of the three target vowels, and six contained long ones. The sentences were matched into pairs so that the members of the pair had a short and a long production of the vowel in a similar phonetic context. Full minimal pairs could mostly not be achieved, as meaningful sentences were used, and minimal pairs did not exist in the Finnish lexicon. The duration of each target vowel was then measured, and the durations were again normalized pairwise by diving the long ones with the short ones. The ratios achieved in this way were used in statistical analyses.

2.1.8 Test structures

The different studies comprising this thesis employed different test structures. What all five studies have in common is that pre- and posttest measurements were performed, but the length of time between them varied. Some studies also had elements that were missing from others. Study I consisted only of a production training task and performance recordings, conducted over two days. The first day began with a pretest recording, followed by the first training session, a recording and the second training session. The second day was the same, but in reverse, so that it began with a training session and ended with a final recording, the posttest. The total duration of the experiment was under one hour.

Studies II and III consisted of EEG recordings, discrimination tasks and production tasks and training. They used a three-day structure. The first day began with pretest EEG recordings and discrimination tasks for all three stimulus types, followed by the pretest production task for the trained linguistic stimuli. These were followed by two production training sessions with the trained linguistic stimuli. On the second day, the order was reversed, and the day began with two production training sessions. These were then followed by the production task, discrimination task and EEG recording, but only for the trained linguistic stimuli. Finally, the third day had an identical structure as the first, except that there was no production training, and the production task was performed for both the trained and untrained linguistic stimuli. The first and second were always consecutive, and the third day was 1-2 weeks after the second one. In total, the experiment lasted six to seven hours.

Study IV was similar in structure to studies II and III, except that the middle day was eliminated, resulting in a two-day experiment. It began with the EEG recordings, followed by the discrimination task and the production tasks. The day ended with four training sessions, two for each stimulus pair. All of these were performed with both stimulus pairs consecutively, so that each task was performed with both stimulus types before moving on to the next one. The order of the training and test blocks was counterbalanced so that half of the participants always started with the stop stimuli, and the other half with the sibilants. The second day was a reverse version of the first, starting with the training sessions and followed by the production task, discrimination task and the EEG recordings. The experiment days were consecutive, and the experiment lasted 4-5 hours.

The structure of Study V was entirely different to the other four. The identification and production tasks were performed on consecutive days both at pretest and posttest, with the identification task preceding the production task. All participants took part in the identification task simultaneously, while the production task was performed individually. The pretest and posttest measurements were separated by the three-week language course that was the intervention in this study. The measurements lasted around 45 minutes for each individual participant. The test structures are presented in Table 3. Test structures of the studies

Table 3. Test structures of the studies.

Study	First day	Second day	1-2 weeks later	4 weeks later
Study I	Production training, production task	Production training, production task		
Study II	Production training, EEG recording, oddball discrimination, production task (all stimuli)	Production training, EEG recording, oddball discrimination, production task (trained stimuli only)	Production training, EEG recording, oddball discrimination, production task (all stimuli)	
Study III	Production training, EEG recording, oddball discrimination, production task (all stimuli)	Production training, EEG recording, oddball discrimination, production task (trained stimuli only)	Production training, EEG recording, oddball discrimination, production task (all stimuli)	
Study IV	Production training, EEG recording, oddball discrimination, production task	Production training, EEG recording, oddball discrimination, production task		
Study V	Identification task, production task	Intensive language course		Identification task, production task

3 Results

In the following section, the results from each of the studies comprising the thesis will be summarized, first individually and then grouped together based on the findings and the learning effects or lack thereof they suggest. More detailed analysis of the findings, however, will be done in the Discussion section.

3.1 Study I

In Study I, only production tasks were performed, and the analysis focused on the F1 and F2 formants of the target and non-target words and their change over time. The average formant values for both words are presented in Figure 1.

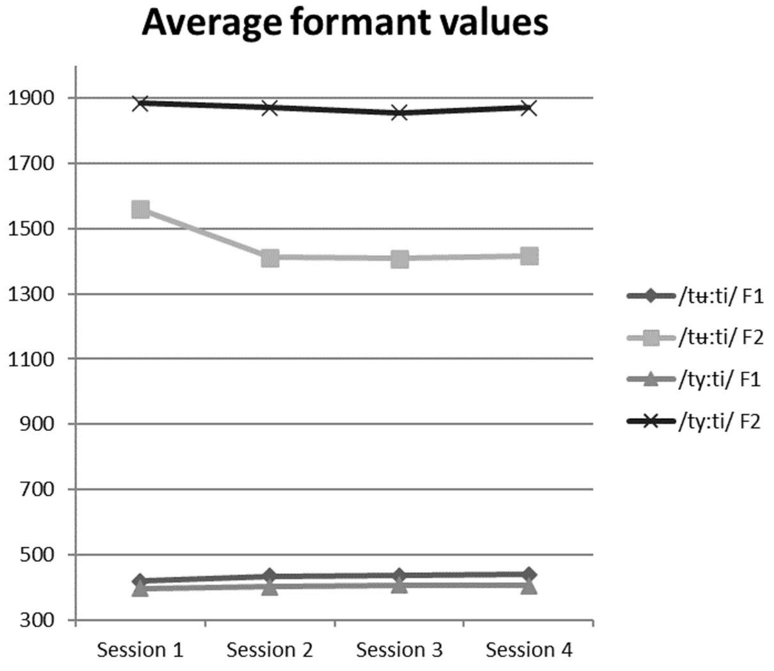

Figure 1. Mean formant values of the production recordings.

It can be seen that the F1 values for each word are very similar and remain unchanged throughout the experiment. Both of these findings were to be expected, as any vowels that the participants were likely to produce, be it /u/, /y/ or /ʉ/, are close, and closeness is mainly seen in a low F1 value. The seeming lack of change in F1 is also expected, as the differences between the vowels are mainly realized through a frontedness contrast, which mainly affects the F2. If the participants heard the vowels as anything /y/ or /u/ like, there would be no reason for articulations that affect F1. F2, however, presents a very different pattern to F1. In the first session, both words are somewhat similar, though already clearly separate, but from the second session on the F2 for /tʉ:ti/ deviated notably from its original values and remained there throughout the rest of the experiment.

In order to see whether these findings are significant, an omnibus repeated measures ANOVA with the structure Word (2) X Session (4) X Formant (2) was performed in order to find any overall effects and interactions. A significant main effect of Word ($F(1,8) = 135.110$; $p < 0.001$) and a Word X Formant interaction ($F(1,8) = 175.714$; $p < 0.001$) were observed, suggesting that overall, the words were different throughout the experiment, and that they differed in their formant structures. No effects or interaction of Session were found. However, given the strong suggestion of change in /tʉ:ti/, indicated by the data, a Word (2) X Session (2) X Formant (2) repeated measures ANOVA was performed between sessions 1 and 2. This resulted in significant main effect of Word ($F(1,8) = 36.012$; $p < 0.001$) and Session ($F(1,8) = 6.343$; $p = 0.036$), followed by a significant Session X Formant interaction ($F(1,8) = 7.900$; $p = 0.023$). This indicated that the two words had been produced differently in the two sessions, but more importantly, a change had occurred, and the change had affected the F1 and F2 formants differently. A similar analysis was conducted between sessions 2-3 and 3-4, but no session effects emerged, suggesting that all change that took place as a result of the training happened already between sessions 1 and 2. Finally, in order to find out if the significant change was a result of the change in F2, as suggested by Figure 1, a Word (2) X Session (2) repeated measures ANOVA was performed with the F2 values alone between sessions 1 and 2. This resulted in a significant main effect of Word ($F(1,8) = 42.283$; $p < 0.001$) and Session, ($F(1,8) = 42.283$; $p = 0.026$), confirming that the change was indeed focused on F2. The same analysis between sessions 2-3 and 3-4 revealed no session effects.

Overall, these results suggest that as a result of listen-and-repeat training with linguistic and articulatory instructions, the participants were able to modify their production of a novel vowel contrast. While the instructions were designed to first increase their awareness of the contrast on the first day, and then provide articulatory instructions on the second, all changes occurred already after the first training session. At this point, the participants had only been made aware that there were two different vowels. These results have some implications for orienting attention during training tasks and will be discussed in more detail in the Discussion section.

3.2 Study II

In Study II, a behavioral oddball discrimination task and a production task were used to gauge the effectiveness of a three-day listen-and-repeat training paradigm on a novel vowel duration contrast. The purpose was to see whether the contrast could be trained, and whether any learning effects in the trained contrast (TL) would be transferred to an untrained one (UT), or even a non-linguistic sine tone contrast (NL), mimicking the structure of the vowel stimuli (see Stimuli section in this thesis).

Average values for the discrimination task can be seen in Figures 2 and 3. The task was performed on each of the three test days for the trained contrast, and on the first and last day for the untrained linguistic and non-linguistic contrasts. Observation of the data suggests that discrimination sensitivity increased across the board, with perhaps the largest increase seen in the trained linguistic pair between days 1 and 2. Statistical analysis, however, did not corroborate this: a repeated measures ANOVA with a Word (3) X Session (2) structure was performed with all stimuli between the first and last sessions with no effects reaching significance. Given the strong signal of improvement for the trained stimuli alone, a repeated measures ANOVA with a Session (3) structure was run with just the trained stimuli. This resulted in a significant main effect of Session (*F*(2,5) = 9,907; *p* = 0,018), indicating that discrimination sensitivity indeed increased as a result of the training. However, post hoc paired sampled t-tests between individual sessions showed a significant difference between the first and second session (*t*(6) = -4,280; *p* = 0,005), but not the first and third, suggesting that the increase in sensitivity that was observed immediately after the training had ended may not have been retained in full by the end of the experiment.

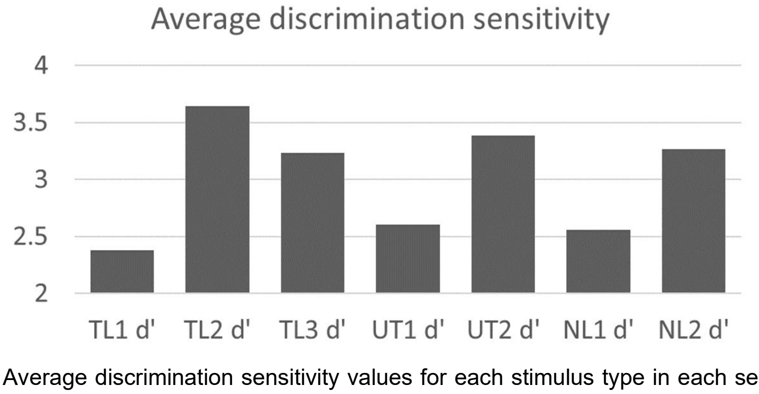

Figure 2. Average discrimination sensitivity values for each stimulus type in each session. TL = trained linguistic, UT = untrained linguistic. NL = non-linguistic.

Suggestion of overall improvement can also be seen in the reaction time data, with all stimulus types showing decreased reaction times between baseline and the end of the experiment. A repeated measures ANOVA with a Word (3) X Session (2) structure between the baseline and the final session resulted in a significant main effect of Word (*F*(2,5) = 8,399; *p* = 0,025) and Session (*F*(1,6) = 9,001; *p* = 0,024), suggesting that the reaction times were indeed lower between sessions, and also that they were different between the stimulus types. Post hoc paired samples t-tests were performed to explore these results further, beginning with between session comparisons for sessions 1 and 3 for each stimulus type. The only significant finding came from the non-linguistic stimulus pair (*t*(6) = 3,144; *p* = 0,02), suggesting that decrease observed for the other stimulus types was not significant. In order to explore the Word main effect, paired samples t-tests were also conducted within the sessions to compare the reaction time between the stimulus types. From these, a significant difference was found between the trained linguistic and non-linguistic stimuli in sessions 1 (*t*(6) = -2,687 ; *p* = 0,036) and 3 (*t*(6) = -3,144; *p* = 0,02), suggesting that the reaction times to the trained linguistic stimuli were consistently faster than those to the non-linguistic ones. No other post hoc tests reached significance.

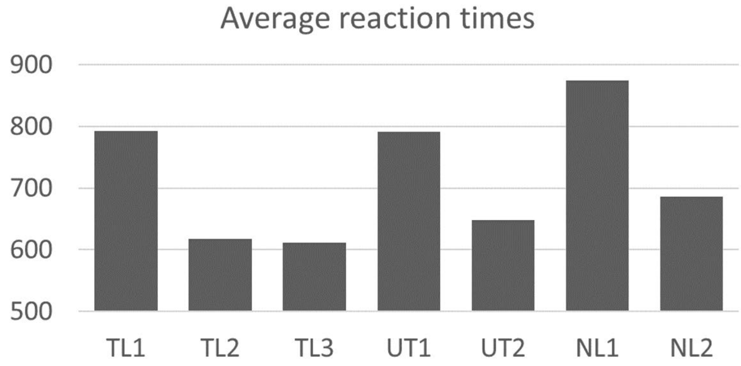

Figure 3. Average discrimination reaction times for each stimulus type in each session. TL = trained linguistic, UT = untrained linguistic. NL = non-linguistic.

The results from the production task can be seen in Figure 4. As stated in the Stimuli section of the thesis, the production ratios used in the analysis were achieved by dividing the durations of the first syllable vowels in the productions of the long members of the stimulus pairs by the same durations from the productions of the short ones. Examination of the data reveals a striking difference between the two stimulus types. It was originally hypothesized that the production ratios would be somewhat similar for the trained and untrained stimulus pairs before the training, but this does not appear to be the case. The ratios for the trained stimuli are clearly higher than those of the untrained ones already at baseline, and the difference increases throughout the experiment. A Session (3) repeated measures ANOVA was first conducted with just the values for the trained stimulus in order to see whether the increase in ratios is statistically significant, but it was not. Next, the ratios for each session were individually compared to the single value for the untrained stimuli with paired samples t-tests. This resulted in a significant difference in sessions 2 ($t(6)$ = 1,824; p = 0,024) and 3 ($t(6)$ = 3,776; p = 0,009), suggesting that despite the seemingly large difference, the ratios between the trained and untrained stimuli did not differ significantly in the first session, but did so in the latter two, suggesting that the training may have had an effect on the production of the trained vowel contrast.

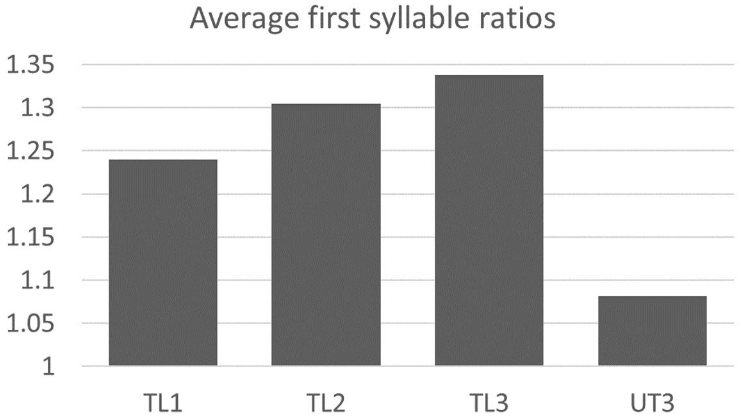

Figure 4. Average long/short ratios for the first syllable vowels for each linguistic stimulus type.

The results from Study II are somewhat undermined by the low participant count; only seven people took part in the study, making it more of a pilot rather than a full-fledged experiment, but the results still point in the direction of perception and production changes being achievable with the quite short training paradigm that was used.

3.3 Study III

Study III is the first of the two studies in the thesis in which ERPs were used to measure learning outcomes in addition to the behavioral measures used in the earlier studies. Study III is based on the data from Study II, with the addition of five more participants. The structure and stimuli are therefore identical, except for the ERP recordings that were excluded from Study II altogether. As explained in the Materials and methods section, the analyses for the ERP recordings are performed separately for each stimulus type, as the observed peaks in the difference waveforms did not match up between them, and therefore the time windows for analysis are different for all of them. Time windows are the same between pre- and posttest, however. The grand average difference waveforms for each stimulus type can be seen in Figures 5, 6 and 7.

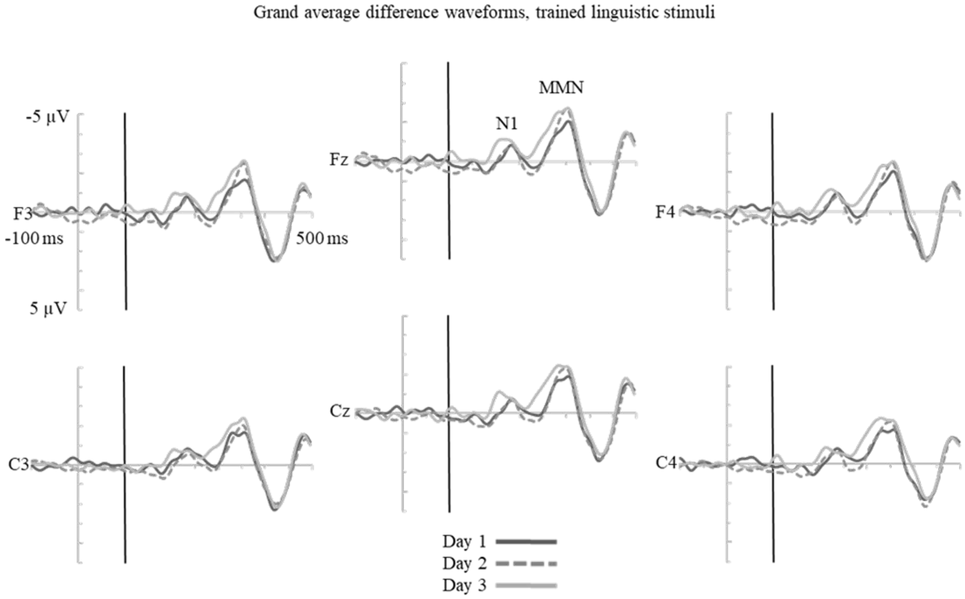

Figure 5. Grand average difference waveforms for the trained linguistic stimuli.

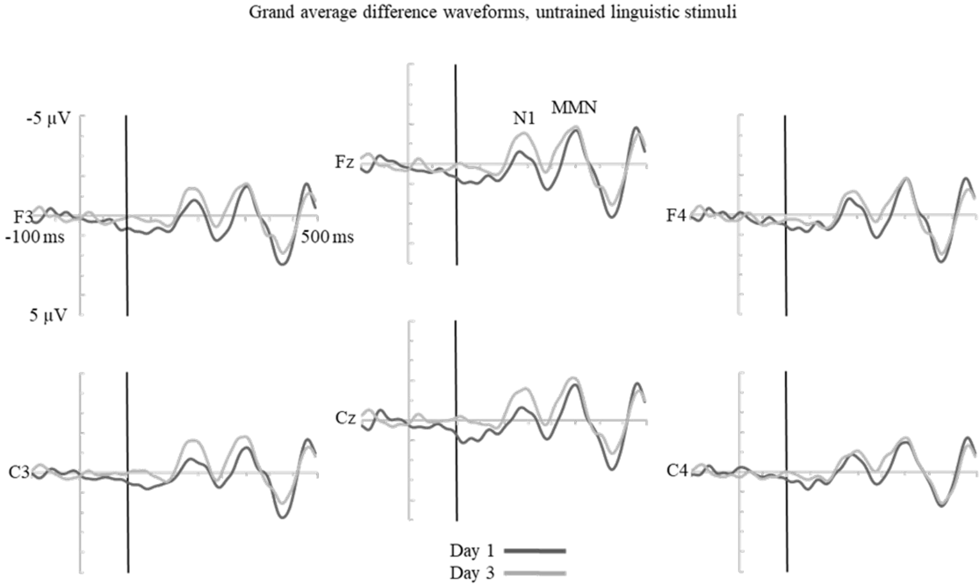

Figure 6. Grand average difference waveforms for the untrained linguistic stimuli.

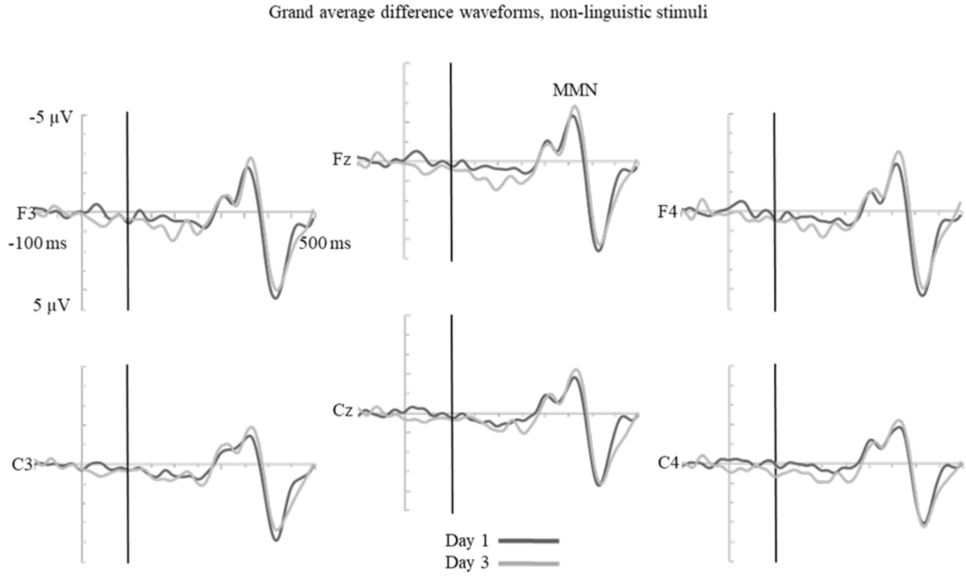

Figure 7. Grand average difference waveforms for the non-linguistic stimuli.

To start the ERP analysis, one-sample t-tests were performed for the Cz and Fz electrodes to determine whether or not statistically significant ERPs were elicited by the different stimulus types. This was done by comparing the mean amplitude values to zero for each stimulus type at each time window. All MMN responses were statistically significant in all sessions, all time windows and all stimulus types, and were therefore considered to be elicited. N1, however, displayed a different pattern. For the trained linguistic stimuli, N1 was not significant at either electrode site in the first time window in the first and second session, but it was significant in both time windows and both electrodes in the third session, suggesting that its latency had decreased as a result of the training. For the untrained linguistic stimuli, N1 was not significant in the first session at either electrode site, but it was significant in the third session at both, suggesting that a N1 response was elicited for the deviant stimuli as a result of training. Mean amplitude values for each electrode can be seen in Table 4; responses that differ significantly from zero are marked with an asterisk (*) in Cz and Fz.

Table 4. Time windows (ms), mean amplitudes (µV) and standard deviations (in brackets, µV) for the psychophysiological measurements for each stimulus in each session for each electrode.

	trained				**untrained**		**non-linguistic**
	N1-1	**N1-2**	**MMN1**	**MMN2**	**N1**	**MMN**	**MMN**
	195-225	**225-255**	**310-340**	**340-370**	**220-250**	**330-360**	**350-380**
Fz							
Session 1	-0,13 (0,9)	-0,61* (0,84)	-1,23* (1,68)	-1,78* (1,26)	-0,47 (0,88)	-1,44* (1,94)	-1,79* (1,14)
Session 2	-0,23 (1,11)	-0,60* (0,79)	-1,40* (1,35)	-2,28* (1,16)	-	-	-
Session 3	-1,03* (0,74)	-0,86* (0,78)	-2,21* (1,73)	2,49* (1,38)	-1,40* (0,65)	-1,72* (1,65)	-2,36* (1,89)
Cz							
Session 1	-0,1 (0,71)	-0,57* (0,59)	-1,37* (1,4)	-1,68* (1,16)	-0,49 (0,82)	-1,54* (1,96)	-1,49* (1,09)
Session 2	-0,11 (0,97)	-0,46* (0,66)	-1,51* (1,07)	-1,94* (1,11)	-	-	-
Session 3	-0,96* (0,8)	-0,66* (0,63)	-2,23* (1,52)	-2,20* (1,32)	-1,46* (0,81)	-1,99* (1,21)	-1,97* (1,74)
C3							
Session 1	-0,34 (0,62)	-0,71 (0,61)	-1,35 (1,06)	-1,49 (1,15)	-0,6 (0,44)	-1,09 (1,69)	-1,1 (1)
Session 2	-0,09 (0,94)	-0,45 (0,6)	-1,14 (0,95)	-1,72 (1,1)	-	-	-
Session 3	-0,7 (0,64)	-0,64 (0,54)	-1,88 (1,53)	-2,1 (1,22)	-1,55 (0,86)	-1,73 (1,1)	-1,62 (1,46)
C4							
Session 1	-0,13 (0,58)	-0,67 (0,64)	-1,52 (1,22)	-1,62 (1,03)	-0,73 (0,75)	-1,4 (1,56)	-1,6 (1,01)
Session 2	-0,28 (0,9)	-0,48 (0,64)	-1,64 (0,9)	-1,85 (0,97)	-	-	-
Session 3	-0,71 (0,48)	-0,58 (0,74)	-2,25 (1,3)	-2,08 (1,25)	-0,92 (0,83)	-1,61 (1,19)	-1,96 (1,35)
F3							
Session 1	-0,16 (0,7)	-0,56 (0,73)	-1,12 (1,46)	-1,51 (1,1)	-0,59 (0,84)	-1,19 (1,84)	-1,6 (0,99)
Session 2	-0,15 (1,08)	-0,57 (0,74)	-1,12 (1,27)	-2,12 (1,19)	-	-	-
Session 3	-0,9 (0,8)	-0,79 (0,7)	-1,85 (1,73)	-2,34 (1,38)	-1,31 (0,64)	-1,49 (1,53)	-2,25 (1,71)
F4							
Session 1	-0,17 (0,92)	-0,77 (0,73)	-1,24 (1,52)	-1,83 (1,14)	-0,63 (0,97)	-1,48 (1,68)	-1,84 (1,25)
Session 2	-0,21 (0,99)	-0,57 (0,74)	-1,38 (1,32)	-2,18 (0,95)	-	-	-
Session 3	-0,89 (0,63)	-0,83 (0,62)	-2,11 (1,58)	-2,36 (1,09)	-1,1 (0,76)	-1,61 (1,5)	-2,62 (1,93)

- = no recordings were made on the second day for the untrained and non-linguistic stimuli. * = responses that statistically differ from zero (only Fz and Cz).

Next, the analysis on N1 continued with a Session(2) X Time window(2) X Electrode(6) repeated measures ANOVA for the trained linguistic stimuli between the first and third sessions. This resulted in a significant Session X Time Window interaction ($F(1,11) = 6.855$; $p = 0.024$; $\eta_p^2 = 0.384$) , implying that different time windows had different mean amplitudes depending on the session. Together with the mean amplitude data and the elicitation analyses showing that N1 was non-existent in the first time window in the first session, this analysis confirms that the latency of the N1 response decreased as a result of training. The same repeated measures analysis for the trained linguistic stimuli was conducted between the first and second and the second and third sessions, but no significant main effects or interactions emerged.

A Session(2) X Electrode(6) repeated measures ANOVA for N1 with the untrained linguistic stimuli between the first and last sessions resulted in a significant main effect of Session ($F(1,11) = 7.889$; $p = 0.017$; $\eta_p^2 = 0.418$). Together with the elicitation analysis this supports the finding that N1 was elicited by the untrained duration contrast as a result of training-induced generalization. N1 analysis was not performed for the nonlinguistic stimuli, as there was no discernible N1 peak in the grand average difference waveform.

MMN analysis started with the trained linguistic stimuli. A Session(2) X Time window(2) X Electrode(6) repeated measures ANOVA was performed for the mean MMN amplitudes between the baseline and the final session, resulting in a significant main effect of Session ($F(1,11) = 5.794$; $p = 0.035$; $\eta_p^2 = 0.345$). This suggests that the mean amplitude of the MMN increased as a result of training, but the latency did not change, as there were no effects of Time window. A Session(2) X Electrode(6) analysis for the mean MMN amplitudes for either the untrained linguistic or the nonlinguistic stimuli did not result in any significant main effects or interactions. A Session(2) X Time window(2) X Electrode(6) repeated measures ANOVA for the trained linguistic stimuli between the first and second sessions resulted in a Session X Time Window X Electrode interaction ($F(5,55) = 2.593$; $p = 0.035$; $\eta_p^2 = 0.191$) and a Time Window X Electrode interaction ($F(5,55) = 6.548$; $p < 0.001$; $\eta_p^2 = 0.373$), and the same analysis between the second and third sessions resulted in a significant main effect of Session ($F(1,11) = 5.361$; $p = 0.041$; $\eta_p^2 = 0.328$). This suggests that the increase in MMN amplitudes for the trained stimuli did not occur immediately after the training, but rather between the break between the second and third sessions.

A combination chart of results of the discrimination task can be seen in Figure 8. In the chart, the vertical position of the markers depicts reaction times, and the horizontal position depicts discrimination accuracy. The lower the marker is, the lower the reaction, and the more right it is, the higher the discrimination accuracy. It can clearly be seen that both the discrimination accuracy and reaction times seem to have improved with completion of the experiment, with both the highest discrimination accuracy and the lowest reaction time achieved for the trained linguistic stimuli on the second day, right after the completion of the training. Whether or not all changes are significant and whether there are differences between the stimuli will be next examined statistically.

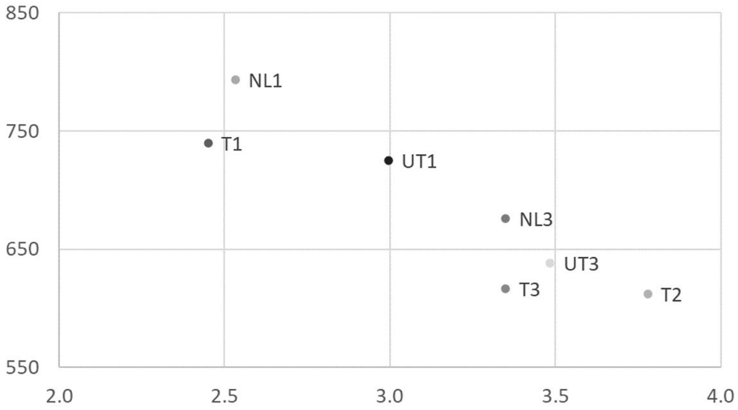

Figure 8. Behavioral discrimination reaction times (vertical axis, ms) and sensitivity scores (horizontal axis) for each stimulus pair in each session. Proximity to bottom right corner indicates improved performance, i.e. lower reaction times and higher discrimination sensitivity.

The discrimination task was statistically analyzed in the same way as in Study II, starting with a Session(2) X Stimulus (3) repeated measures ANOVA with all three stimuli between Session 1 and Session 3. A significant main effect of Session (*F*(1,11) = 6.545; *p* = 0.027; η_p^2 = 0.373) was found, suggesting that the overall, reaction times decreased between after completion of the experiment. A Session(2) repeated measures ANOVA was then carried out for each stimulus type separately between the first and last sessions in order to see if all of them had changed significantly. This resulted in a significant main effect of Session for the trained linguistic (*F*(1,11) = 5.168; *p* = 0.044; η_p^2 = 0.320) and the non-linguistic stimuli (*F*(1,11) = 6.633; *p* = 0.026; η_p^2 = 0.376). No effect emerged for the untrained linguistic stimuli. Paired samples t-tests were then carried out between the trained linguistic and non-linguistic stimuli within the same sessions in order to find out if observed difference in reaction times between the two is also statistically significant. The difference was nonsignificant in the first, but significant in the third session (*t* (11) =−2.468; *p* = 0.031; *d* = 0.656), suggesting that at the end of the experiment, the participants were able to respond to the trained linguistic stimuli faster than the non-linguistic ones. Finally, Session(2) repeated measures ANOVAs were conducted only for the trained linguistic stimuli between first and second sessions, and the second and third sessions, resulting in a significant main effect of Session (*F*(1,11) = 5.986; *p* = 0.032; η_p^2 = 0.352) and no main effects, respectively. This shows that the decrease in reaction times for the trained linguistic stimuli was already present immediately after the training had ended and did not undergo further significant changes by the end of the experiment.

Discrimination accuracy was analyzed similarly to the reaction times, starting with a Session (2) X Stimulus (3) repeated measures ANOVA between the first and third sessions. This resulted in a significant main effect of Session (*F*(1,11) = 6.030; *p* = 0.032; η_p^2 = 0.354), suggesting that overall discrimination sensitivity increased, as indicated by Figure 8. No other effects or interactions were significant. Each stimulus type was then analyzed separately with a Session (2) repeated measures ANOVA between sessions 1 and 3. This resulted in a significant main effect of Session (*F*(1,11) = 11.842; *p* = 0.006; η_p^2 = 0.518) for the trained linguistic stimuli, but no significant effects for the other stimulus types. Session (2) Repeated measures ANOVAs between sessions 1-2 and 2-3 for the trained linguistic stimuli resulted in a significant main effect of Session (*F*(1,11) = 21.157; *p* = 0.001; η_p^2 = 0,658) between the first and second days, but not the second and third. The discrimination sensitivity therefore followed a similar pattern to the reaction times, where peak performance for the trained stimuli was achieved immediately after the training in the second measurement session, with no further significant changes occurring between that and the end of the experiment.

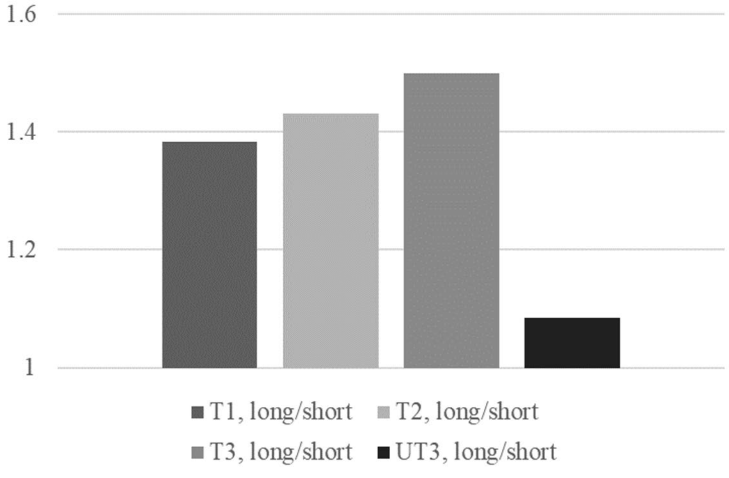

Figure 9. Average long/short production ratios of the first syllables of both stimulus pairs, calculated by dividing the vowel durations of the repetitions of the long members of the pairs by the duration of the short ones. Values above 1 indicate that repetitions of the long vowels were longer than the short ones.

Initial examination of the production ratios (Figure 9) suggests that the participants produced the long members of the trained linguistic pair with a duration that was 38% higher than that of the short members. From then on, the difference increased steadily to 50% in the third session. The only measurement for the untrained linguistic pair, however, shows that the participants only produced an 8% difference between the short and long vowels. In order to gauge these findings statistically, a Session(3) repeated measures ANOVA was performed for the production ratios of the trained linguistic stimuli. No significant findings emerged, indicating that the training did not have a systematic effect on the participants' production. Next, paired samples t-tests were performed session by session between the ratios of trained and untrained linguistic stimuli. This analysis structure was necessary as only one production measurement was performed for the untrained stimuli. The t-tests showed that the production ratios for the trained linguistic pair were significantly higher than those for the untrained pair on all test days: Day 1 = $t(11) = 4.567$; $p = 0.001$; $d = 1.802$; Day 2= $t(11) = 4.604$; $p = 0.001$; $d = 1.458$; Day 3 = $t(11) = 4.871$; $p < 0.001$; $d = 1.354$.

Overall, the results from Study III strengthen the mixed preliminary findings of Study II in suggesting that listen-and-repeat training can be used to achieve significant learning effects in the perception of sound duration. The perception of the trained linguistic stimuli improved both psychophysiologically and behaviorally, as evidenced by the increased MMN amplitude, shortened latency of the N1, higher discrimination accuracy and lower reaction times. Furthermore, some generalization effects were found in the N1 response for the untrained linguistic stimuli, although no changes were detected in the MMN. These results hint at a lower detection threshold for general sound duration. The discrimination reaction times also decreased for the untrained linguistic stimuli, although reactions to the trained linguistic stimuli were faster on the final day of the experiment. The suggestions of production improvements seen in Study II were not confirmed by these results, with no significant changes over time in the duration ratios between the long and short members of the trained stimulus pair.

3.4 Study IV

In Study IV, the focus of the research shifted to the duration differences of consonants rather than vowels, but the study still followed the same basic structure as studies II and III, with the exception of the removal of the third test day. Unlike in Studies II and III, both stimulus types, sibilants and stops, were tested and trained equally much. The methodology used was otherwise the same as in Study III, with ERPs (MMN and P3), behavioral discrimination, and production tests being the main indicators of possible learning effects, with the first two training sessions performed on the first day and the last two on the second. The grand average difference waves for both stimulus types can be seen in Figure 10 and the mean amplitudes for each electrode in Table 5.

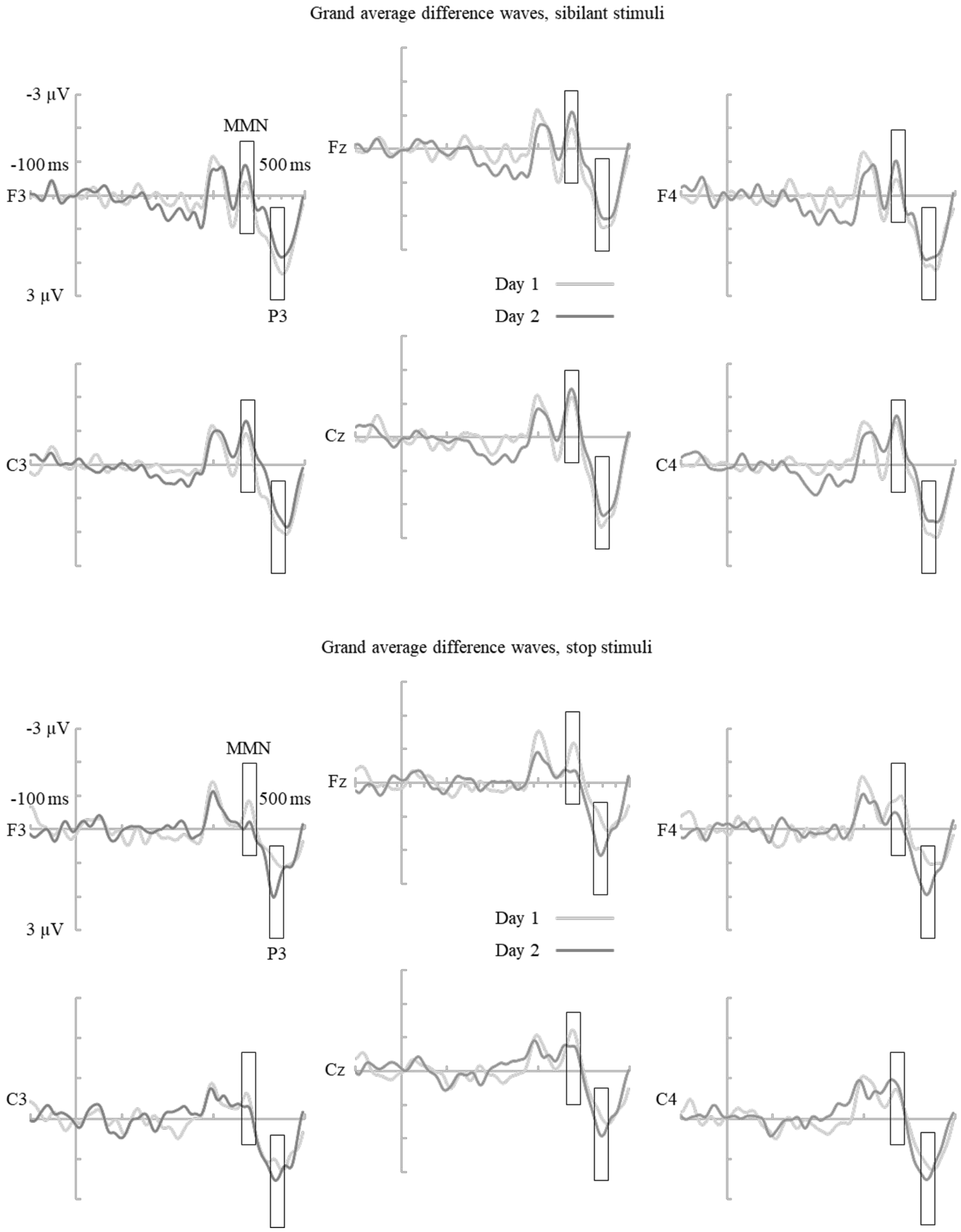

Figure 10. Grand average difference waveforms for the sibilant and stop stimuli for the C3, C4, Cz, F3, F4 and Fz electrodes. Boxes indicate the time windows for each ERP. NB: the difference between the standard and deviant stimuli begins at 170 ms for both stimulus pairs.

Statistical analysis of the EEG recordings started with one-sample t-tests on the mean amplitudes of the MMN response at Fz and Cz electrodes to determine whether the response differed from zero. The analysis revealed mostly significant responses, with the exception of the Fz electrode in the first session for the sibilants, and the Fz electrode in the second session for the stops. These amplitudes can be seen in Table 5.

Table 5. Mean EEG amplitudes.

Fz	sibilant		stop	
	MMN	P3	MMN	P3
Pre-test	-0.29 (1.19)	2.20 (1.17)	-0.91 (0.95)*	1.13 (0.72)
Posttest	-0.75 (1.07)*	1.92 (1.40)	-0.32 (0.85)	1.93 (1.23)
Cz	sibilant		stop	
	MMN	P3	MMN	P3
Pre-test	-0.89 (1.27)*	2.46 (1.30)	-0.97 (0.85)*	1.34 (0.85)
Posttest	-1.11 (1.04)*	2.11 (1.49)	-0.67 (0.76)*	1.74 (0.96)
C3	sibilant		stop	
	MMN	P3	MMN	P3
Pre-test	-0.64 (1.15)	1.75 (0.97)	-0.40 (0.69)	1.12 (0.64)
Posttest	-1.02 (1.01)	1.31 (0.99)	-0.23 (0.64)	1.40 (0.88)
C4	sibilant		stop	
	MMN	P3	MMN	P3
Pre-test	-0.90 (1.03)	1.89 (1.10)	-0.56 (0.69)	1.11 (0.77)
Posttest	-1.13 (1.05)	1.49 (1.18)	-0.72 (0.68)	1.40 (0.76)
F3	sibilant		stop	
	MMN	P3	MMN	P3
Pre-test	-0.15 (1.23)	1.98 (1.02)	-0.56 (0.96)	0.96 (0.84)
Posttest	-0.55 (1.00)	1.54 (0.98)	-0.09 (0.90)	1.69 (0.91)
F4	sibilant		stop	
	MMN	P3	MMN	P3
Pre-test	-0.24 (0.94)	2.04 (1.18)	-0.89 (0.92)	0.88 (0.67)
Posttest	-0.57 (1.01)	1.84 (1.13)	-0.34 (0.92)	1.77 (0.89)

Mean EEG amplitudes (μV) and their standard deviations (in brackets) for each stimulus type for the electrodes Cz, Fz, C3, C4, F3 and F4. Responses that differ significantly from zero are marked with an asterisk (only analyzed for Cz and Fz).

Next, the mean amplitudes were subjected to a Word(2) X Session (2) X Electrode (6) repeated measures ANOVA, with both stimulus types in Word, first and last sessions in Session and the electrodes Cz, C3, C4, Fz, F3 and F4 in Electrode. This resulted in a significant main effect of Electrode ($F(5,55) = 6.350$; p

<0.001; $\eta_p^2 = 0.366$) and a significant Word X Electrode interaction ($F(5,55) = 3.993$; $p = 0.004$; $\eta_p^2 = 0.266$), suggesting differences in the mean amplitudes between the electrode sites between stimulus types, and a significant Word X Session X Electrode interaction ($F(5,55) = 4.999$; $p = 0.001$; $\eta_p^2 = 0.312$). In order to examine the interactions further, a Word(2) X Electrode(6) repeated measures ANOVA with both stimulus pairs was performed for the first and last sessions separately. For the first session, this resulted in a significant main effect of Electrode ($F(5,55) = 3.782$; $p = 0.005$; $\eta_p^2 = 0.256$) and a Word X Electrode interaction ($F(5,55) = 7.153$; $p < 0.001$; $\eta_p^2 = 0.394$). The same analysis for the final session resulted in a significant main effect of Electrode ($F(5,55) = 6.251$; $p < 0.001$; $\eta_p^2 = 0.362$). These results suggest that the MMN amplitudes differ between electrode sites in both sessions, and additionally between the two stimulus types in the first session. When the same electrodes were compared between the two stimulus types within the same session with paired samples t-tests (i.e. C3 electrode for the sibilants against the C3 electrode for stops at pre-test, then C4 for the same etc.), no significant findings emerged. This suggests that the topography of the responses for each stimulus type was similar. To further examine Word X Session X Electrode interaction, a Session(2) X Electrode(6) repeated measures ANOVA was performed for both stimulus types separately. For the sibilants, this resulted in a significant main effect of Electrode ($F(5,55) = 2.169$; $p < 0.001$; $\eta_p^2 = 0.339$), and for the stops a significant main effect of Electrode ($F(5,55) = 4.852$; $p = 0.001$; $\eta_p^2 = 0.306$) and significant Session X Electrode interaction ($F(5,55) = 4.044$; $p = 0.003$; $\eta_p^2 = 0.269$) were found. This shows that no training-related effects are found in the responses for the stimuli, but for the stops, there were differences between the electrodes between the first and last sessions. However, paired samples t-tests for each electrode between the two sessions (i.e. C3 in Session 1 vs. C3 in Session 2 etc.) resulted in no significant findings.

Initial examination of the discrimination reaction times (Figure 11) suggests that reaction times for both stimulus types decreased with training, and that they were lower for the sibilant stimuli both at the end and the beginning of the experiment. Statistical analysis was started with a Session (2) X Word (2) repeated measures ANOVA, resulting in a significant main effect of Session ($F(1, 17) = 11.004$; $p = 0.004$; $\eta_p^2 = 0.257$)) and Word ($F(1, 17) = 9.238$; $p = 0.007$; $\eta_p^2 = 0.269$), supporting both initial observations. In order to confirm the causes behind the main effects, paired samples t-tests were performed, comparing the reaction times for each stimulus type individually between the first and last sessions. Significant differences were found for both stimulus types: $t(17) = 3.260$; $p = 0.005$; $d = 0.626$ for the sibilants and $t(17) = 2.519$; $p = 0.022$; $d = 0.660$ for the stops, meaning that the participants were able to respond to both stimulus types significantly faster as a result of training. Next, within-session paired samples t-tests were performed between the two stimulus types to examine the main effect of Word. The reaction times were not significantly different in the first session, but were so in the second ($t(17) = -3.377$; $p = 0.004$; $d = 0.747$), suggesting that at the end of experiment, the participants were able to respond to the sibilant stimuli faster than the stops.

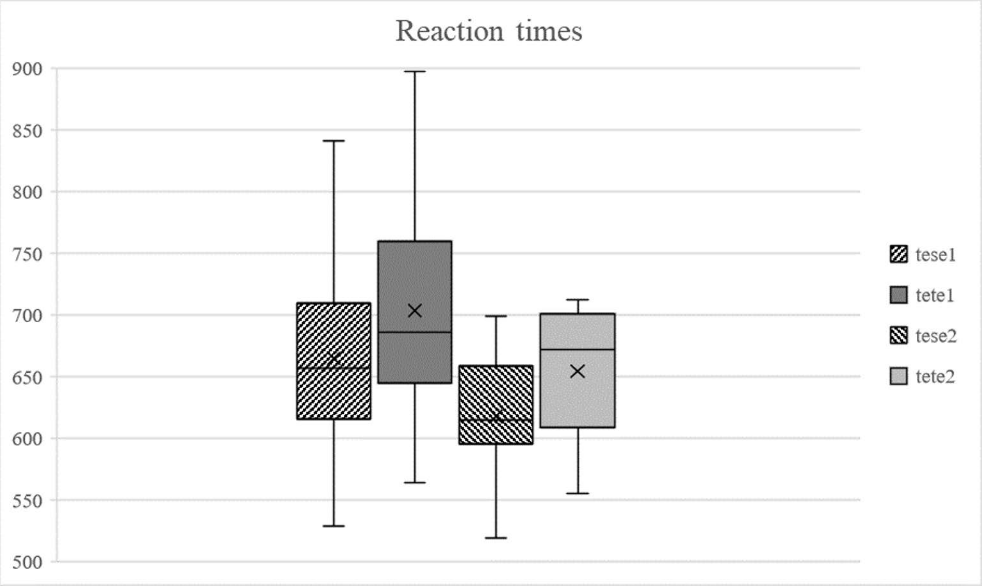

Figure 11. Average discrimination reaction times in milliseconds (Y-axis) for the sibilant (tese) and stop (tete) stimuli on each day (1 = pre-test, 2 = posttest). The box represents approximately 50% of the participants, and the whiskers mark upper and lower ranges. The horizontal line marks the median value.

Next, discrimination accuracy scores (Figure 12) were analyzed. Similarly to the reaction times, initial examination suggests that accuracy improved and variation decreased as a result of training, particularly for the sibilant stimuli. Analysis was started with a Session (2) X Word (2) repeated measures ANOVA, resulting in a significant main effect of Session ($F(1,17) = 8.725$; $p = 0.009$; $\eta_p^2 = 0.339$), confirming an overall improvement in discrimination accuracy. In order to examine this further, paired samples t-tests were run for both stimulus types between the first and last sessions, resulting in significant differences for both the sibilants ($t(17) = -2.709$; $p = 0.015$; $d = 0.426$) and the stops ($t(17) = -2.738$; $p = 0.014$; $d = 0.607$). The training therefore resulted in improved discrimination accuracy with no difference between the stimulus types.

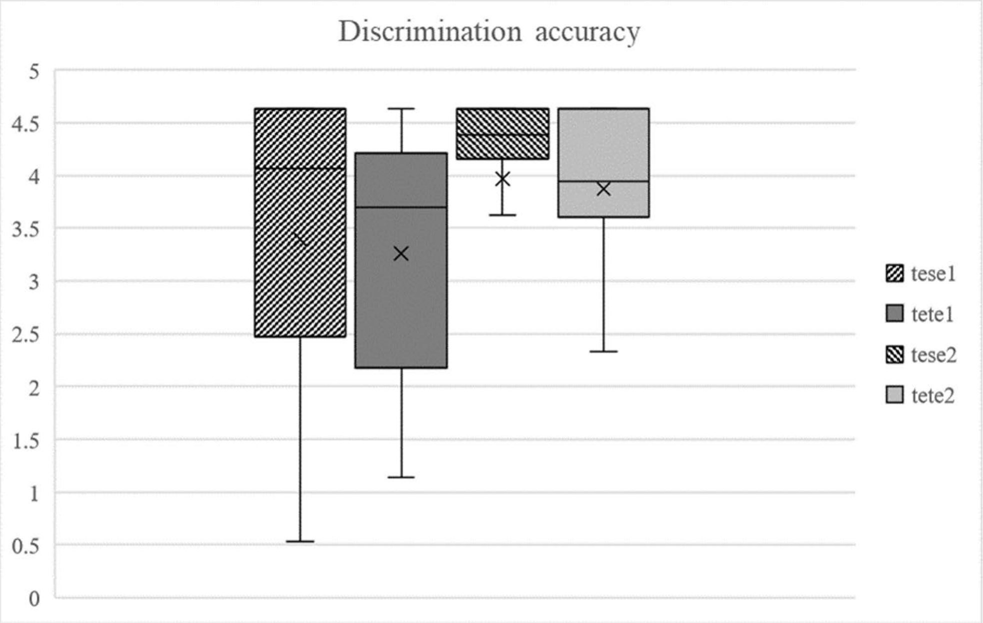

Figure 12. Average discrimination accuracy scores (Y-axis) for the sibilant (tese) and stop (tete) stimuli on each day (1 = pre-test, 2 = posttest). The box represents approximately 50% of the participants, and the whiskers mark upper and lower ranges. The horizontal line marks the median value

Finally, production results can be seen in Figure 13. Initial examination suggests that overall, the sibilant stimuli were produced with lower ratios than the stops, and that there was considerably less variation in their production. A Session (2) X Word (2) repeated measures ANOVA of the long/short ratios was performed, resulting in a significant main effect of Word ($F(1,17) = 15.459$; $p = < 0.001$; $\eta_p^2 = 0.476$), confirming the difference between the stimulus types, seen in Figure 13. Paired samples t-tests were then performed within each session, comparing the ratios between the stimulus types. The difference was significant in the first ($t(17) = 2.543$; $p = 0.021$; $d = 0.547$) and the second ($t(17) = 3.096$; $p = 0.007$; $d = 0.704$) session, indicating that the long and short sibilants were produced with a smaller and more consistent difference than the stops throughout the experiment. No effects of training emerged in the production results.

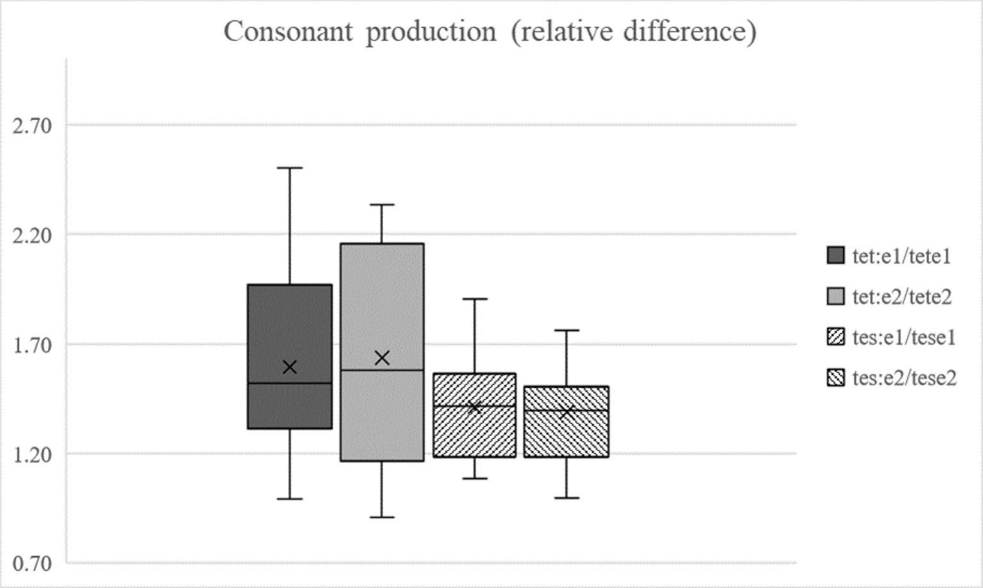

Figure 13. Average long/short ratios (Y-axis) for the consonants of the sibilant (tese/tesse) and stop (tete/tette) stimuli on both days (1 = pre-test, 2 = posttest). The box represents approximately 50% of the participants, and the whiskers mark upper and lower ranges. The horizontal line marks the median value.

Overall, the results differ somewhat from studies II and III, in which a very similar training paradigm was used with vowel durations. No training-related changes at all were seen in the psychophysiological measurements, in contrast to the changes in MMN amplitudes and N1 latency observed for the trained stimulus pair in Study III. While improvements were seen in the behavioral discrimination task in both reaction times and accuracy scores, the lack of psychophysiological development may mean that they need to be interpreted differently to Study III. Clear differences emerged between the stimulus types, however, most clearly seen in the production task, where the sibilant duration contrast was produced with much a smaller and more consistent difference than the stop contrast throughout the experiment. The reaction times to the stop consonants were also higher than the ones to the sibilants at the end of the experiment, suggesting that the participants had to process the two contrasts in a different manner.

3.5 Study V

Study V represents an entirely different type of experiment compared to the other four. The number of stimuli was very high, comprising of 50 stimulus pairs in the identification task and 6 sentence pairs in the production task. Furthermore, the participants were divided into two different groups, speakers of quantity or non-quantity languages, based on whether their native languages contained phonological duration contrasts. Furthermore, a native control group was used for both of the tasks. The analyses were in particular complicated by this control group: as their sole purpose was to provide a way assess the native-likeness of the participants' performance they did not take part in the intervention. Therefore, they were not expected to develop in any way and they were only recorded taking part in the tasks once. This meant that they did not fulfill the assumptions of a repeated measures ANOVA, and they had to be analyzed separately. This essentially meant that for both tasks, three separate analyses were required: first, an omnibus ANOVA containing all the variables for each participant group. Second, a repeated measures ANOVA with the participant groups to gauge intervention-related effects over time. Third, one-way ANOVAs for pretest and posttest with all three groups, followed by post hoc tests, to assess how native-like the participants' performance was. The structure of the analyses was therefore more complicated than in the previous studies, and the analyses will here be presented in table format, rather than individual statistical tests, and the findings will be discussed on a slightly more general level.

The mean identification error rates for the identification task for the various variables can be seen in Figure 14. Overall, the non-quantity group made more errors than the quantity group at both pretest and posttest, and the quantity group approached the native control group quite closely on several of the variables. Error rates between the short and long vowels show that the participants mistakenly identified long vowels as short ones more often than the other way around. Both groups generally achieved lower error rates at posttest.

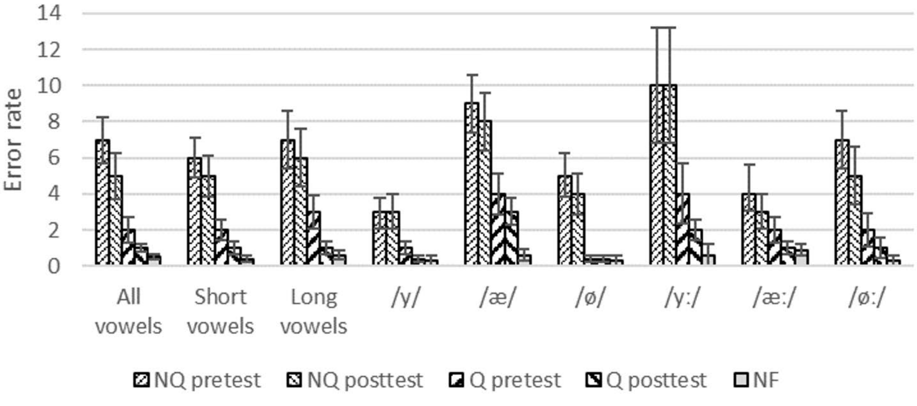

Figure 14. Percentage of incorrect responses for short and long vowels in the identification task for each group. The error bars indicate standard error.

Statistical analysis of the identification results begun with a four-way ANOVA with a (Group(2) x Time (2) x Vowel type(3) x Vowel length(2)) structure, where Time represented pre- and posttest measurements, Group consisted of the two non-native groups, Vowel type of the three vowels /y/, /æ/ or /ø/, and Vowel length of short and long durations. This resulted in the significant main effects of Time (F(1, 395) = 14.8, $p < 0.001$, $\eta_p^2 = 0.04$) and Group (F(1, 395) = 16.4, $p < 0.001$, $\eta_p^2 = 0.04$), suggesting an overall difference in performance over time and the two non-native groups. Furthermore the Vowel type * Vovel length interation was significant (F(2, 395) = 7.1, $p = 0.001$, $\eta_p^2 = 0.04$), suggesting a difference in performance with short and long duration for different vowels. Further analysis was performed by carrying out repeated measures ANOVAs with a structure of Group(2) X Time(2), where Time represented pre- and posttest measurements, and Group always consisted of the two quantity language groups. The findings, both significant and non-significant, are presented in Table 6.

Table 6. Group comparisons of the identification error rates.

	Sentence	Main effect & interaction	Details
Quantity groups	All vowels	Time	F(1,65) = 9.1, **p = 0.004**, η_p^2 = 0.1
		Group	F(1,65) = 7.48, **p = 0.008**, η_p^2 = 0.1
		Group * Time	F(1,65) = 0.08, p = 0.79, η_p^2 = 0.001
	Short vowels	Time	F(1,65) = 2.0, p = 0.16, η_p^2 = 0.03
		Group	F(1,65) = 8.2, **p = 0.006**, η_p^2 = 0.1
		Group * Time	F(1,65) = 0.02, p = 0.90, η_p^2 < 0.001
	Long vowels	Time	F(1,66) = 6.7, **p = 0.01**, η_p^2 = 0.1
		Group	F(1,66) = 14.0, **p = 0.049**, η_p^2 = 0.1
		Group * Time	F(1,66) = 0.06, p = 0.80, η_p^2 = 0.001
	/y/	Time	F(1,66) = 1.4, p = 0.24, η_p^2 = 0.02
		Group	F(1,66) = 4.3, **p = 0.042**, η_p^2 = 0.1
		Group * Time	F(1,66) = 0.44, p = 0.51, η_p^2 = 0.01
	/æ/	Time	F(1,66) = 1.8, p = 0.19, η_p^2 = 0.001
		Group	F(1,66) = 0.04, p = 0.09, η_p^2 = 0.001
		Group * Time	F(1,66) = 0.4, p = 0.85, η_p^2 = 0.001
	/ø/	Time	F(1,65) = 1.3, p = 0.26, η_p^2 = 0.02
		Group	F(1,65) = 9.6, **p = 0.003**, η_p^2 = 0.1
		Group * Time	F(1,65) = 1.3, p = 0.26, η_p^2 = 0.02
	/yː/	Time	F(1,66) = 4.1, **p = 0.048**, η_p^2 = 0.1
		Group	F(1,66) = 5.2, **p = 0.025**, η_p^2 = 0.1
		Group * Time	F(1,66) = 0.1, p = 0.77, η_p^2 = 0.001
	/æː/	Time	F(1,66) = 4.1, **p = 0.047**, η_p^2 = 0.1
		Group	F(1,66) = 1.8, p = 0.18, η_p^2 = 0.03
		Group * Time	F(1,66) = 0.2, p = 0.70, η_p^2 = 0.002
	/øː/	Time	F(1,66) = 4.1, **p = 0.046**, η_p^2 = 0.1
		Group	F(1,66) = 3.2, p = 0.08, η_p^2 = 0.1
		Group * Time	F(1,66) = 0.4, p = 0.54, η_p^2 = 0.01

Note. *p*-values in bold indicate statistically significant findings.

The statistical analyses revealed that identification error rates lowered significantly overall, and more specifically for all the long vowel variables, as evidenced by the significant main effects of Time. This improvement occurred despite the error rates being quite low to begin with. The lack of any Group X Time interactions suggests that the groups improved roughly equally. However, significant main effects of Group were found for both short and long vowels, and analysis of individual vowels showed that they occurred in /y/, /ø/ and /yː/. In all of these cases the quantity group outperformed the non-quantity group, suggesting that they were at an advantage in the perception of the contrasts.

Next, the participant groups were compared to the native control group separately for pretest and posttest using one-way ANOVAs with Group(3) structure (Table 7). Bonferroni corrected post hoc tests are listed separately for each significant ANOVA (Table 8).

Table 7. Statistical analysis of identification error rates between all groups (one-way ANOVA).

	Variable	**Main effect**	**Details**
NQ, Q and NF	All vowels	Pretest: Group	*F*(2,76) = 5.4, ***p* = 0.006**, η_p^2 = 0.2
		Posttest: Group	*F*(2,77) = 6.0, ***p* = 0.004**, η_p^2 = 0.14
	Short vowels	Pretest: Group	*F*(2,76) = 6.4, ***p* = 0.003,** η_p^2 = 0.2
		Posttest: Group	*F*(2,77) = 5.5, ***p* = 0.006**, η_p^2 = 0.13
	Long vowels	Pretest: Group	*F*(2,77) = 2.8, *p* = 0.068, η_p^2 = 0.1
		Posttest: Group	*F*(2,77) = 3.4, ***p* = 0.039**, η_p^2 = 0.1
	/y/	Pretest: Group	*F*(2,77) = 2.5, *p* = 0.09, η_p^2 = 0.1
		Posttest: Group	*F*(2,77) = 3.0, *p* = 0.055, η_p^2 = 0.1
	/æ/	Pretest: Group	*F*(2,77) = 5.5, ***p* = 0.006**, η_p^2 = 0.1
		Posttest: Group	*F*(2,77) = 5.2, ***p* = 0.008**, η_p^2 = 0.1
	/ø/	Pretest: Group	*F*(2,76) = 6.4, ***p* = 0.003**, η_p^2 = 0.1
		Posttest: Group	*F*(2,77) = 5.5, ***p* = 0.006**, η_p^2 = 0.13
	/yː/	Pretest: Group	*F*(2,77) = 3.6, ***p* = 0.03**, η_p^2 = 0.1
		Posttest: Group	*F*(2,77) = 4.2, ***p* = 0.019**, η_p^2 = 0.1
	/æː/	Pretest: Group	*F*(2,77) = 1.0, *p* = 0.38, η_p^2 = 0.03
		Posttest: Group	*F*(2,77) = 1.7, *p* = 0.18, η_p^2 = 0.04
	/øː/	Pretest: Group	*F*(2,77) = 3.2, *p* = 0.08, η_p^2 = 0.1
		Posttest: Group	*F*(2,77) = 2.2, *p* = 0.11, η_p^2 = 0.1

Note. NQ = non-quantity group. Q = quantity group. NF = native Finnish control group. *p*-values in bold indicate statistically significant findings.

Table 8. Post hoc analyses of identification error rates, pretest and posttest.

	Variable	Post hoc tests
NQ, Q and NF	**/æ/**	Pretest: NQ > Q *p* = 0.056. NQ > NF *p* = **0.016**. Q > NF *p* = 0.737 Posttest: NQ > Q *p* = **0.033**. NQ > NF *p* = **0.037**. Q > NF *p* = 1.00
	/ø/	Pretest: NQ > Q ***p* = 0.005**. NQ > NF *p* = 0.066. Q > NF *p* = 1.00 Posttest: NQ > Q *p* = **0.010**. NQ > NF *p* = 0.102. Q > NF *p* = 1.00
	/y:/	Posttest: NQ > Q ***p* = 0.037**. NQ > NF *p* = 0.132. Q > NF *p* = 1.00

Note. Post hoc analyses were conducted in the instances where a significant main effect of Group was found in the one-way ANOVA for individual vowel analysis (Table 6). Furthermore, entirely non-significant post hoc findings have been omitted. Statistically significant findings are in bold. NQ = non-quantity group. Q = quantity group. NF = native Finnish control group. The < and > symbols indicate which group had the larger error rate in each comparison.

These analyses show that for all of the variables tested, the quantity group was able to perform at a native-like level in identifying long and short vowel contrasts. It also further confirms that whenever there were differences between the quantity language groups, it was a case of the quantity group performing better than the non-quantity group. The non-quantity group performed significantly worse than the native group in the identification of /æ/ at both pretest and posttest, and new significant differences emerged at posttest between the quantity language groups in the identification of /æ/ and /y:/. These results suggest that the non-quantity group did not benefit from taking part in the course with regards to identification.

Moving on to the production task, mean ratios produced by the participants and the native control group can be seen in Figure 15. Initial examination of the ratios suggests that the non-quantity produced slightly higher ratios after they had taken part in the language course, while for the quantity group they were lower for each vowel type. The native control group had the highest ratios for all vowel types, and the non-quantity group had higher ratios than the quantity group for /æ/ and /ø/ at pretest and posttest. What is noteworthy is that as the quantity group's ratios lowered, this meant that they were moving away from native-like ratios.

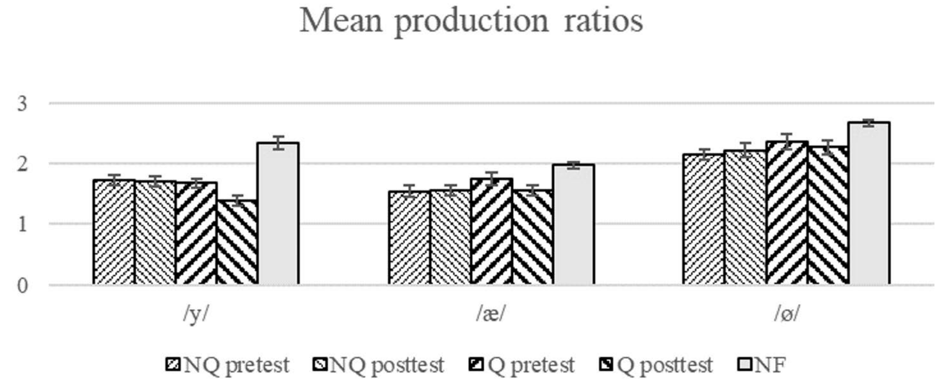

Figure 15. Mean production ratios for each vowel type in the production task for each group. The error bars indicate standard error.

Statistical analysis was started by performing a three-way ANOVA with a (Group(2) x Time(2) x Vowel Type(3)) structure, resulting in a statistically significant effect of Vowel type ($F(2,\ 348) = 5.5$, $p = 0.004$, $\eta_p^2 = 0.03$) and significant interactions between Time * Vowel Type ($F(2, 348) = 42.5$, $p < 0.001$, $\eta_p^2 = 0.2$) and Group * Time * Vowel Type ($F(2, 348) = 3.6$, $p = 0.003$, $\eta_p^2 = 0.02$). In the post hoc tests, a statistically significant difference was found between /y/ and /æ/ (p = 0.006) and between /y/ and /ø/ (p = 0.003).

Further analysis was carried out by performing repeated measures ANOVAs with a Group(2) X Time(2) structure for each vowel type in order to find out whether there were developments over time, and whether the participant groups differed from each other. Group always consisted of the two quantity groups, and Time consisted of pre- and posttest measurements. Results for these analyses can be seen in Table 9.

Table 9. Repeated measures ANOVAs of the production ratios with the quantity groups.

	Vowel type	**Main effect & interaction**	**Details**
Quantity groups	/y/	Time	*F*(1, 63) = 5.5, **p = 0.02**, η_p^2 = 0.08
		Group	*F*(1, 63) = 8.9, p = 0.11, η_p^2 = 0.04
		Group * Time	*F*(1, 63) = 3.7, p = 0.06, η_p^2 = 0.06
	/æ/	Time	*F*(1, 63) = 2.1, p = 0.15, η_p^2 = 0.03
		Group	*F*(1, 63) = 0.64, p = 0,43, η_p^2 = 0.01
		Group * Time	*F*(1, 63) = 2.9, p = 0.09, η_p^2 = 0.04
	/ø/	Time	*F*(1, 63) = 0.06, p = 0.81, η_p^2 = 0.001
		Group	*F*(1, 63) = 0.64, p = 0,43, η_p^2 = 0.01
		Group * Time	*F*(1, 63) = 1.8, p = 0.19, η_p^2 = 0.03

Note. p-values in bold indicate statistically significant findings.

This analysis revealed that the only development over time was also for /y/. Examination of Figure 15 suggests that this may mostly have been due to a notable decrease in the ratios for the quantity language group, suggesting that the group moved away from native-like ratios after completing the language course, instead of improving. For the other two vowel types, no change one way or the other was observed.

In order to statistically compare the participants to the native control group, one-way Group(3) ANOVAs were next performed with all three groups separately at pretest and posttest for each vowel type. Bonferroni corrected post hoc results are also reported wherever significant main effects were found in the ANOVAs. The results from these analyses are presented in tables Table 10 and Table 11.

Table 10. Statistical analysis of production ratios between all groups (one-way ANOVA).

	Vowel type	**Main effect**	**Details**
NQ, Q and NF	/y/	Pretest: Group Posttest: Group	$F(2,71) = 4.8$, $p = 0.009$, $\eta_p^2 = 0.12$ $F(2,71) = 9.3$, $p = 0.001$, $\eta_p^2 = 0.21$
	/æ/	Pretest: Group Posttest: Group	$F(2,74) = 2.6$, $p = 0.08$, $\eta_p^2 = 0.07$ $F(2,74) = 3.03$, $p = 0.06$, $\eta_p^2 = 0.08$
	/ø/	Pretest: Group Posttest: Group	$F(2,74) = 2.7$, $p = 0.08$, $\eta_p^2 = 0.07$ $F(2,74) = 1.9$, $p = 0.16$, $\eta_p^2 = 0.05$

Note. NQ = non-quantity group. Q = quantity group. NF = native Finnish control group.

Table 11. Post hoc analyses of duration ratios in production, pretest and posttest.

	Vowel type	**Post hoc tests**
NQ, Q and NF	/y/	Pretest: NQ < Q *p* = 1.0. NQ < NF ***p* = 0.015**. Q < NF ***p* = 0.011.** Posttest: NQ > Q *p* = 0.07. NQ < NF ***p* = 0.014**. Q <. NF ***p* < 0.001.**

Note. Post hoc analyses are provided only in the instances where a significant main effect of Group was found in the one-way ANOVA (Table 9). Statistically significant findings are in bold. NQ = non-quantity group. Q = quantity group. NF = native Finnish control group. The < and > symbols indicate which group had the larger ratios in each comparison.

The group comparison shows that the participants did not differ from the native speakers for the vowel types /æ/ and /ø/ at either pretest or posttest. For the vowel type /y/, where a significant main effect of Group was found, the post hoc test shows that both participants groups differed significantly from the native speakers, but not from each other. As hinted at by the repeated measures ANOVAs, both quantity groups produced the /y/ contrasts with a lower duration ratio than the native Finnish speakers.

Overall, despite the completely different intervention, the results for Study V are similar to studies II and III, in which vowel duration contrasts were also the target of examination. The participants were able to improve their perception of the non-native duration contrasts in a statistically significant way, as measured by the identification task. Improvement occurred even though the participants' performance level was already quite good at the beginning of the course. However, similarly to studies II and III, little change was seen in the production task, with only one vowel type showing intervention-related changes. Furthermore, these changes actually saw the participants move away from native-like productions. The lack of production changes may, however, be explained by the fact the participants produced two of the three vowel types with native-like long/short ratios already at the pretest stage, and there may therefore have been little room for further development.

Comparing the two participant groups, a clear advantage was found for speakers of quantity languages, who outperformed non-quantity language speakers consistently in the identification task and did not differ from the native control group in any of the tested variables. In the production task, however, no such advantage seemed to exist, as no statistically significant differences suggesting superiority one way or the other were found between the participant groups.

3.6 Overall results

Summarizing the results does show some clear trends that are visible across the different studies. Four out of the five studies in the thesis were performed with non-native vowels or their duration as the feature being examined, and in all of these the training paradigm or the intervention produced some kind of learning results. In Study I, where only production was used to measure learning results with listen-and-repeat training of a non-native a vowel quality contrast, the participants were able to modify their production in statistically significant way immediately after the training started, and the effect lasted until the end of the experiment. In studies II and III, where a three-day listen-and-repeat training paradigm was used to train a vowel duration contrast, improvement was seen in the psychophysiological perception (only Study III), behavioral discrimination accuracy and discrimination reaction times. Furthermore, psychophysiological generalization effects in the N1 response were also seen in Study III, suggesting that the brain became more sensitive to duration contrasts overall, even though the effect did not transfer to the MMN response for the untrained vowel contrast. Finally, in Study V, where the intervention was an intensive language course and improvement was tested on several different vowel duration contrasts, clear improvements were seen in the identification accuracy of long vowels despite high accuracy already at the pretest phase.

In Study IV, however, where sibilant and stop consonant duration contrasts were the target of listen-and-repeat training, improvement was not seen to a similar lever as it was with vowel contrasts. Psychophysiological measurements with MMN showed no significant improvement whatsoever, and for the stop contrast the elicitation of the MMN response actually weakened on the second day of the study after all the training had been completed. While behavioral discrimination did show improvement for both of the trained contrasts, given the total lack of psychophysiological development it is not clear whether or not this is an indication of phonological learning or something else, such as a task familiarization effect.

Another clear finding was the lack of production improvement in all of the duration studies. As stated, studies II-IV used listen-and-repeat training as the intervention, in which perceptual and production training are combined. Of these, only Study II suggested any changes in the production of the trained duration contrast, and even this finding is questionable, given that in it was not seen in Study III that used the same training paradigm with more participants. Study IV with consonant contrasts showed no indication of production changes whatsoever, and in Study V the production of one of the vowel types used in the production did change significantly, but the change saw the non-native participants perform worse, i.e. less native-like after they had completed the intervention. This suggests that there is a fundamental difference between in the way production develops between (vowel) quality and duration differences, as Study I saw the participants change their productions very fast and in a manner that suggested change towards more correct pronunciation of the novel contrast.

Finally, in Study IV there was indication that the two different consonant types were processed differently, with the sibilants being easier for the participants. This was evident in lower discrimination reaction times, more consistent production ratios, and the fact that the MMN response for the stop contrast actually got weaker after the training.

In the next section, all of these findings will be discussed in more detail, with respect to the research questions posed at the beginning of Section 2.1, and an attempt will be made to explain the underlying reasons for the findings.

4 Discussion

In the previous sections, the theoretical background, methodology and results of the studies comprising this thesis were explained in detail. In this section, the results will now be summarized and discussed, particularly in relation to the five research questions presented in Section 2.1. As stated, the overall focus of this thesis is to examine the effects of listen-and-repeat training on the learning non-native duration contrasts, and to compare them to classroom learning. To this end, five questions were formed. Suggested answers to these questions will be provided in the following paragraphs, based on the findings of each study applicable to each specific question.

4.1 Can non-native duration contrasts be successfully learned using listen-and-repeat, similarly to vowel quality and voice onset time contrasts?

For question one, the broadest of the five, the overall answer seems to be “Yes”. Improvement of some sort was observed in each of the duration-related listen-and-repeat studies in the thesis. Studies II and III, in which vowel contrasts were used as the training and testing stimuli, showed clear signs of improvement in both the behavioral and the psychophysiological measures of perception, with most effects concentrated on the stimulus pair that was trained. For the trained stimuli in Study III in particular, the MMN amplitude increased, latency for the N1 decreased, the sensitivity of behavioral discrimination increased, and reaction times lowered. All of these changed are indicative of improved perceptual performance, and they are also in support of each other: improvement in psychophysiological discrimination is often accompanied by behavioral improvement as well (e.g. Tremblay et al., 1998). Improvement was also observed in Study IV, where consonants were used as the training and testing stimuli, with behavioral discrimination sensitivity and reaction times both improving for both the stop and the sibilant stimulus pairs, both of which were trained an equal amount. Psychophysiological measures did not change, this will be discussed further regarding question 4. This type of perceptual improvement is well in line with previous LAB-lab listen-and-repeat training studies, particularly those by Tamminen et al. (2015) and Tamminen & Peltola (2015), in both of which listen-and-repeat training was used to train the perception of VOT contrasts in labiodental fricatives and which found improvement in both MMN amplitudes and behavioral identification. They also reflect some of the results from other training studies focused on duration contrasts (e.g. Hirata et al., 2007; Okuno, 2014; Tajima et al., 2008) that used different methodologies over varying amounts of time. What is noteworthy about listen-and-repeat studies overall is the low amount of training sessions and the consequently short time period during which the effects are achieved: all of the previously discussed studies of vowel quality (Jähi et al., 2015; Peltola et al., 2020, 2015; Taimi, Alku, et al., 2014; Taimi, Jähi, et al., 2014), voice onset time (Tamminen & Peltola, 2015; Tamminen et al., 2015) and studies I-IV in this thesis achieved their results in just four sessions and two days of training, with less than 200 repetitions of the trained contrast. This is in stark contrast to popular methods such as high variability phonetic training (e.g. Bradlow et al., 1999; Hirata et al., 2007) where thousands of repetitions of the target stimuli are used in several sessions over weeks of time.

4.2 Are the possible learning results limited to perception or production, or are both faculties affected by training?

For this question, differences begin to emerge between the studies in the thesis and between earlier training studies. As stated above, perceptual learning effects were found in all of the listen-and-repeat duration studies where they were measured. Behavioral changes were observed in studies II, III and IV, and psychophysiological ones in Study III. However, production improvement for the listen-and-repeat studies in the thesis was only observed in Study I, the pilot study for the project, where listen-and-repeat and pronunciation instructions were used to train a non-native vowel quality contrast. The amount of training per contrast was the same as in studies II, III and IV, four sessions over two consecutive days, with a total of 120 repetitions of the training contrasts. However, while in Study I the participants were able to significantly modify their production of the trained contrast after being given just one session of training and some very broad instructions, no training-related production changes occurred at all in the other three studies. The only indication of production changes in these studies came in Study II, where the production ratios for the trained stimulus pair did not differ from the untrained pair in the pretest session, but differed in the following ones. When the same paradigm and stimuli were used again with more participants in Study III, however, this effect was no longer observed. These results are at odds with previous findings from listen-and-repeat training studies with vowel quality contrasts using a highly similar paradigm (Jähi et al., 2015; Peltola et al., 2020, 2015; Taimi, Alku, et al., 2014; Taimi, Jähi, et al., 2014), which have all shown that this training method can alter the productions of child, adult and senior participants in two days or less. Other studies, using entirely perceptual methods, have also been able to induce production changes (Bradlow et al., 1999, e.g. 1997; Lambacher, Martens, Kakehi, Marasinghe, & Molholt, 2005; Lopez-Soto & Kewley-Port, 2009; Y. Wang et al., 2003). One of the main reasons for the lack of change in the duration studies of this thesis may be that the productions of the participants, at least for the trained contrasts, were already quite close to an acceptable level at the beginning of each experiment. In the listen-and-repeat studies with duration contrasts, the participants were not given any feedback or other clues about their performance level, and it was up to them to determine both what the relevant feature was and if they were performing well enough. All participants in the duration studies were either students of Finnish or had previous experience with Finnish through residence, and it may be that they were able to produce a contrast that they thought was good enough and employed a strategy of simply using that contrast throughout the experiments. In fact, in Study V, where a native control group was used to assess the participants' production, the participants' productions were mostly native-like in the three vowel types. The participants' performance also remained largely unchanged between pre- and posttest. Study V will be discussed in more detail for question 5.

4.3 Can duration be trained as a general process that the learner can apply to untrained speech sounds or non-linguistic sounds, or is it specific to certain phonemes?

For question 3, the answer is slightly mixed. Studies II and III dealt with generalization questions, by using one vowel pair as the training stimulus, and then examining learning effects in another, untrained vowel pair and an entirely non-linguistic sinusoidal tone pair, mimicking the durational structure of the vowel pairs. While the perception of the trained pair improved in all of the measured ways, no significant changes occurred in the perception of the untrained linguistic pair that were suggestive of any linguistic improvement. There was, however, an increase in the amplitude of the N1 response to the untrained pair, thought to reflect a response to the physical features of the stimulus, rather than being a linguistic component. This could mean training did indeed have an effect on the overall processing of duration, even if the linguistic memory trace for the untrained vowel was not affected. Some support for this interpretation can be found in the findings that duration and quality are processed separately in the brain (Ylinen, Huotilainen, et al., 2005) and that the temporal processing of sounds, speech or not, seems to derive from a specific location in the brain (Liégeois-Chauvel, De Graaf, Laguitton, & Chauvel, 1999). It may be that the training allowed for duration to be detected as the correct cue for the difference between the sounds the participants heard, and that the detection became more sensitive as a whole. This interpretation is further supported by the findings that the latency of the N1 response decreased for the trained linguistic stimuli, and that the discrimination reaction times decreased for the non-linguistic pair. Näätänen (1992) suggested that N1 amplitudes could indeed be selectively enhanced for stimuli that were deemed relevant, and that this effect could be due to a “general increase in sensory sensitivity” (Näätänen, 1992, p. 132). The reason why this processing difference was not transferred to an improved MMN response could be that the memory trace responsible for eliciting the MMN is by nature acoustically complex, and it consists of a spectral as well as a temporal component. As the untrained vowel pair was, by definition, untrained, the memory trace was not activated as much as the one for the trained vowel, and there was therefore limited spectral information available to connect the duration cue to the correct vowel. Spectral as well as temporal information may therefore be necessary to properly train the neural representations.

Previous studies with duration, using other types of training than listen-and-repeat, have both succeeded and failed in finding generalization to novel contrasts after training. For example, Hirata et al. (2007) used a high-variability approach in training vowel length contrasts using different speaking rates as the source of variability. Participants completed four sessions of training over 11-17 days, each consisting of 540 stimuli for a total of 2160 tokens, and it was found that those who trained with both slow and fast spoken stimuli improved the most in their perception of duration contrasts that they did not train with. Okuno (2014), trained vowel duration using perceptual training with either an audio-only or an audiovisual paradigm. They found perceptual generalization to novel stimuli and unfamiliar speakers after the completion of the training paradigm, which consisted of eight sessions, approximately 3.5 hours in total. Conversely, Tajima et al. (2008) used an identification training paradigm consisting of 3600 trials over five days to train perception of Japanese vowel contrasts, and did not find generalization to untrained contrasts, though an effect was found to new talkers with the trained contrasts. Based on these findings, it seems reasonable to assume that more training would be required to achieve generalization effects than was provided in the studies comprising this thesis, perhaps also spread over a longer time period than was used. While Tajima et al. (2008) used quite a large number of stimuli, similarly to the other studies, the training was kept short with just five days in total, compared to the 11-17 days used by Hirata et al. Given that generalization to novel duration stimuli seems to be possible, although not achieved in this thesis, more study may be required. Future studies with listen-and-repeat training and duration contrasts should examine the effects of training with different stimulus types, i.e. ones that contain spectral information vs ones that do not, and the effects of simply adding more training with the current stimulus sets. This could shed light on whether or not generalization is achievable with listen-and-repeat training.

4.4 Does the learning of duration differ between vowels and consonants? Or between different types of consonants?

Regarding question 4, there were both similarities and differences between the studies focusing on vowel duration, and Study IV that focused on the training of consonant duration. Behavioral perception for the trained stimuli improved significantly in all of them, with both stimulus types showing improvement in discrimination accuracy and reaction times in Study IV. Furthermore, no clear improvement was observed in the production results in any of the studies. In Study IV, however, behavioral improvement was not coupled with psychophysiological improvement. Statistical analysis revealed no change over time in the MMN (or P3, for that matter), and there was indirect evidence of the MMN response actually getting weaker with training for the stop stimuli, indicated by the fact that the response stopped being elicited in the Fz electrode in the second session. This suggests a major difference in the way vowel and consonant duration can be trained, with vowels seemingly being much more susceptible to this training paradigm. The fact that there was no psychophysiological development while behavioral performance improved also calls into question whether the behavioral results are suggestive of linguistic learning or something else.

In addition to the lack of psychophysiological development for the consonants, it is noteworthy that the amplitude of the responses for the consonant stimuli was clearly lower than those for the vowels. Examination of the Cz and Fz electrodes reveals that in Study III, the mean amplitudes, measured around the highest peaks of the MMN, were consistently over 1.5 μV in the first session for all stimulus types, and increased to up to 2.49 μV for the trained stimulus pair in the final session. In Study IV, by comparison, the mean amplitudes were consistently under 1 μV except for the Cz electrode in Session 2 for the sibilant stimuli, which measured 1.11 μV. The lowest amplitudes were 0.29 μV and 0.32 μV for the sibilants and stops, respectively. This means that the lowest MMN amplitude for the vowel stimuli was 35% higher than the highest for the consonants, suggesting at significant differences in how well they were perceived by the participants.

At least part of the reason the overall perception and the development patterns between the vowels and the consonants were so obviously different in the duration studies may lie in their saliency. Bohn (1995) suggests in his Desensitization Hypothesis that non-native speakers may be able to use duration contrastively with non-native contrasts even if it is not a distinctive feature in their native language. This would take place when spectral features are not salient enough to be of use in the distinction of the contrast, typically due to the influence of the native language. A typical example of this would be distinction of the English tense-lax contrasts, that have a primary spectral cue and secondary temporal one. If the learner is unable to distinguish between the two vowels spectrally, they will use duration instead. This was confirmed in Bohn's study with Spanish, German and Mandarin learners of English hearing the English /i/ - /ɪ/ and /ɛ/ - /æ/ contrasts, and the hypothesis has been corroborated by a number of other studies (e.g. Cebrian, 2006; Kondaurova, 2008; Rato & Rauber, 2015). In the case of the primarily duration-based contrasts in quantity languages, spectral cues are often very limited. This is indeed the case for the vowel stimuli in this thesis: in studies II-IV, all vowels are identical in quality, and in Study V the identification stimuli are produced by a native Finnish speaker and display the typical minor quality differences of Finnish, with the longer vowels showing slightly more prototypical qualities than the short ones. This may have forced the non-native participants to rely on the duration cue to differentiate the vowels. Evidence for this can be seen in the fact that MMN responses of relatively high amplitudes already existed for the vowel duration contrasts at the beginning of Study III and the very low error rates observed in the identification task in Study V. As stated in the previous paragraph, the initial MMN amplitudes for the consonants were notably lower across the board than those for the vowels, suggesting that their initial perception was not subject to a similar effect. Due to this effect, it may therefore be that consonant duration contrasts require more input and, i.e. more training, than vowels in order to achieve similar learning results. It should be noted this does apply to all types of consonant contrasts: the findings by Tamminen et al. (2015) and Tamminen & Peltola (2015) show that the perception of consonant contrasts, VOT in this case, can be trained in a short time frame with listen-and-repeat. It may be that duration contrasts are simply less accessible for consonants than other contrast types.

Interesting differences emerged when examining the results between the consonant types, sibilants and stops, in Study IV. While their development patterns over time were the same, the participants both perceived and produced the duration contrasts in different ways. In the discrimination task, both contrasts had similar reaction times in the first session, but after the training the participants were able to respond to the sibilant stimuli significantly faster than the stops. An even clearer difference emerged in the production task, where the stop stimuli were produced with a larger short-long difference than the sibilants both before and after the task. The variation in the productions of the participants is particularly noticeable and displays an inverse pattern between the stimulus types: for the sibilants, the initially quite low variation is even lower after the task, while for the stops variation increases in both directions, with some participants producing a smaller difference after training and some larger, even though the mean difference stays roughly the same. Some indirect evidence of a difference was also seen in the psychophysiological measurements. When the elicitation of the MMN was measured by comparing the responses to 0, the stimulus types again displayed opposite patterns: MMN was not significant for the sibilants in the Fz electrode in the first session, but became significant in the second. For the stops, the reverse is true, as MMN was elicited in the Fz in the first session, but not the second. Similarly to the production results, this suggests that something about the stops became more difficult for the participants as the study progressed. These results are somewhat aligned with the findings of Hardison and Motohashi-Saigo (2009) who have studied the learning of Japanese geminates with perceptual training, and found differences in the way stops and sibilants are perceived and learned. In their studies, the sibilant /s/ was found to be more difficult for the participants to perceive initially than the stops /t/ or /k/, although the sibilants went on to show the most perceptual improvement after training. In the current study, there was some weak evidence of a similar pattern, as the MMN response was less uniform for the sibilants in the first session than it was in the second. This finding is far from conclusive, however, and overall, it seems that the stop was more difficult for the participants than the sibilant.

What could be behind these between-consonant differences? Both consonant types were presented in identical acoustic and morphological contexts, with exactly the same stimulus lengths and target region durations, meaning that any differences should stem directly from the consonants themselves. Furthermore, why did behavioral perception improve for both consonant types, while there was a distinct lack of any psychophysiological improvement? The answer to the latter question may be similar to the findings of Study III: it is possible that duration was correctly identified as the feature separating the two members of the stimulus pairs, but the participants were not able to connect the duration cue to the memory trace responsible for the detection of /s/ or /t/. The difference waveforms acquired for the stimuli in studies III and IV are quite similar in their shape, in that the MMN is preceded by distinct N1 peak. N1, thought to be non-linguistic and elicited by the detection of the physical properties of the stimuli, did not to develop over time, but its existence in the difference waveform, at an amplitude similar to the MMN, suggests that some kind of physical difference was detected between the long and short stimuli for both the sibilants and the stops. Furthermore, the oddball discrimination task does not force the participants to focus on any specific feature of the stimuli, they simply need to detect that there is *something* different about the deviant and react to it. With this being the case, the improvement in the behavioral discrimination task could be explained by the participants becoming better over time at detecting the duration difference, despite it not being linguistically relevant to them. These results bear resemblance to the ones achieved in Study III, where the N1 was elicited for the untrained stimuli after training, suggesting improved general duration detection. In the case of Study IV, the detection ability may already have been there, and the participants simply became better at behavioral detection as well.

The difference between the consonant types, however, may be dependent on the actual stimuli and the way they were interpreted by the participants. As they were provided with virtually no information about the words they would hear during the study, nor feedback about their performance, it was up to the participants to find the relevant cues differentiating the short and long members of the stimulus pairs. While this resembles the natural way to acquire non-native contrasts, it also opens up the possibility for completely erroneous interpretations. Quené (1992) studied the role of variations in the duration of intervocalic consonants on the interpretation of word boundaries in Dutch. In the study, whole word stimuli were presented in semantically ambiguous contexts, and the duration of intervocalic consonants was varied artificially. Acoustic context was kept otherwise identical throughout. It was found that the participants' interpretation of the stimuli was significantly affected by the modification of the duration, leading to the conclusion that "acoustic-phonetic cues contribute to word segmentation, at least under conditions where no other information is available (Quené, 1992, p. 345). The latter part of the sentence is particularly relevant for the studies of this thesis, as the use of individual, semantically meaningless pseudowords meant that there was quite literally no other information to judge word boundaries on. It is therefore possible that at least some participants interpreted the stimuli consisting of two words rather than two syllables. This is particularly likely in the case of the stop stimuli, as there is a clear boundary between the two syllables, caused by the occlusion phase of the consonant, instead of continuous frication noise. Furthermore, while the stimulus words are meaningless as individual words, the stop stimuli could be interpreted as a sequence of the Finnish words "te" ("you"). Given that all participants are students of Finnish and they have not been provided with any context about the words, this is not an entirely unplausible interpretation. If this were the case, the larger ratios for the stop stimuli could be explained by attempts to produce two words in sequence with a distinct gap, rather than two syllables. This could also explain the slower reaction times in the discrimination task, if some participants thought the stimulus they were listening for consisted of two words, rather than one. Unfortunately, the design of this study does not allow to for the confirmation of this interpretation, and the participants were not questioned on what they thought the stimuli meant.

4.5 How does listen-and-repeat training compare with an intensive language course with a communicative approach and no specific focus on pronunciation?

Studies I-IV in this thesis all followed the same basic paradigm of four sessions of listen-and-repeat training over two consecutive days, with various additions to this basic system and various methods used to measure learning outcomes. However, what they all have in common is that they were performed in a laboratory setting one person at a time, and do not therefore resemble a typical language instruction situation. Therefore, Study V examined the results of a four-week intensive Finnish language course, where the focus was on overall communicative abilities and interaction with other students. There was no focus on any specific area of pronunciation, and indeed no targeted pronunciation instruction to begin with. The target feature was vowel duration, and learning outcomes were measured with a two-alternative forced-choice identification task and production task consisting of read aloud sentences with relevant duration contrasts embedded in them.

Interestingly, the pattern of perceptual improvement accompanied by a lack of production improvement was also observed in Study V. In this study, production ratios only changed in one of the three tested vowel types, and this change saw the participants actually perform worse than they did before starting the course. Meanwhile, perceptual performance in the two-alternative forced choice identification test improved significantly, despite very good performance already at the pretest phase. This perceptual improvement is particularly interesting for two reasons: first, the error rates for the participants were mostly very low to begin with, with the quantity language group performing at a native-like level at both pretest and posttest. Second, an identification task, rather than discrimination, was used. The latter is significant, because it is thought that identification tasks are better at measuring actual linguistic ability than discrimination task, as the task involves conscious decision-making in the categorization, rather than a simple same-different evaluation, based on non-phonological properties of the stimulus (e.g. Hallé, Chang, & Best, 2004). This means that the perceptual improvement observed here may be more grounded in actual linguistic improvement than the discrimination improvement in Study IV, for example, that was not accompanied by any change in psychophysiology. This is impossible to investigate with the current methodologies, however, and would require more comprehensive testing of the participants' language skills, which falls outside the scope of this thesis.

The lack of production improvement is most likely explained by the fact that the participants mostly performed at a native-like level already at the start of the experiment, similarly to the quite high ratios in the duration production tasks in the other studies. This meant that there was no room for improvement. The finding that the only significant change over time saw the participants perform less native-like at posttest than they did at pre-test, however, requires more examination. In addition to the one significant change, non-significant changes in the participants' productions also mainly indicated changes away from native-like performance, rather than towards it, particularly for the group that consisted of native speakers of quantity languages. It is likely that this is a reflection of an incomplete learning process. All participants were at a somewhat early stage of their Finnish studies, and the slight worsening of the production ratios could indicate a U-shaped learning curve, where learners initially perform well, then abandon the correct behavior, and finally return to it as their studies advance (Rogers, Rakison, & McClelland, 2004). This process reflects the transfer from patterns and behaviors learned by heart to a more productive system that applies language rules productively. The language course environment may have been particularly conducive to this development, as it may have been the first time many of the students were subjected to regular native speaker input, in addition to hearing a wide variety of non-native accents with their language-specific mistakes. Flege et al. (2021) suggested in their revision of the Speech Learning Model that the quality of second language input is important, and that exposure to other non-native learners' speech may result in non-native like production. While the four-week duration of the language course was shorter than the "months" suggested by Flege et al. (e.g. Flege et al., 2021, p. 28), it still seems reasonable to assume that the somewhat varied quality of the input may have had an effect on the participants' category representations, and by extension their productions.

5 Conclusions

Overall, both the listen-and-repeat training studies and the language course were able to induce improvement in the perception of non-native vowel duration contrasts, but not in their production. The reasons for this are partly similar, in that the production levels for the participants were already quite high at the beginning of all of the experiments, and partly dissimilar, in that the language course study saw the participants perform worse in production. It seems, however, that both methods can produce learning results for non-native contrasts and given that the listen-and-repeat training took only two days, compared to the four weeks for the language course, it seems reasonable to suggest that it can be used as a somewhat effective and rapid way to support second language education in adulthood. More research is needed to determine how quickly production changes can be induced; previous listen-and-repeat studies have achieved production changes in vowel quality in two days or less, but for consonant contrasts, be they VOT or duration, improvement has so far proved elusive. Given that production learning for consonant duration has previously been achieved even with purely perception-based training, there is no reason to think that listen-and-repeat could not work, but care has to be taken in stimulus selection to ensure that they provide enough of a challenge to the participants so as to not render improvement impossible. Adaptive approaches using continua of stimuli of varying difficulties, for example, could be used to provide a more fine-grained view of the learners' skill level than that achieved with pseudoword pairs. Furthermore, listen-and-repeat seemed to work better for vowel than consonant duration contrasts. Further research is also needed to shed light on this phenomenon, starting by examining the learning of vowels and consonants in the same study in order to eliminate individual differences between participants, and any effects of the study conditions. Again, there is no reason to think that consonants are particularly resistant to learning with this method, but aspects of the study design may have masked results; direct comparison with the same learners would allow for more accurate assessment of differences between vowels and consonants.

Abbreviations

μV	microvolt
ANOVA	analysis of variance
DIVA	Directions Into Velocities of Articulators
EEG	electroencephalography
ERP	event-related potential
Hz	hertz
IPA	International Phonetic Alphabet
L1	native language/first language
L2	second language
LAB-lab	Learning, Age and Bilingualism laboratory
MMF	mismatch field
MMN	mismatch negativity
ms	millisecond
NF	native Finnish speakers
NL	non-linguistic stimuli
NLM	Native Language Magnet
NQ	non-quantity language speakers
PAM	Perceptual Assimilation Model
Q	quantity language speakers
SLM	Speech Learning Model
SSG	Semi-synthetic Speech Generation
TL	trained linguistic stimuli
UT	untrained linguistic stimuli
VOT	voice onset time
η_p^2	partial eta squared

List of references

Akahane-Yamada, R., McDermott, E., Adachi, T., Kawahara, H., & Pruitt, J. S., 1998. Computer-based second language production training by using spectrographic representation and HMM-based speech recognition scores. Retrieved from http://www.isca-speech.org/archive/icslp_1998/i98_0429.html

Aliaga-Garcia, C., & Mora, J. C., 2009. Assessing the effects of phonetic training on L2 sound perception and production. *Recent Research in Second Language Phonetics/Phonology: Perception and Production.*, (January), 2–31.

Alku, P., Tiitinen, H., & Näätänen, R., 1999. A method for generating natural-sounding speech stimuli for cognitive brain research. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, *110*(8), 1329–1333. <https://doi.org/10.1016/S1388-2457(99)00088-7>

Altmann, H., Berger, I., & Braun, B. 2012., Asymmetries in the perception of non-native consonantal and vocalic length contrasts. *Second Language Research*, *28*(4), 387–413. <https://doi.org/10.1177/0267658312456544>

Best, C. T., 1995. A direct realist perspective on cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research,* (pp. 167-200.). York Press, Baltimore.

Best, C. T., & Tyler, M. D., 2007. Non-native and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O.-S. Bohn (Eds.), *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege* (Vol. 10389). Amsterdam: John Benjamins Publishing Company.

Bohn, O.-S., 1995. Cross-language speech perception in adults: first language transfer doesn't tell it all. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 279–304). York Press, Baltimore.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y., 1999. Training Japanese listeners to identify English /r/ and /l/: long-term retention of learning in perception and production. *Perception & Psychophysics*, *61*(5), 977–985. <https://doi.org/10.3758/BF03206911>

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y., 1997. Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*(4), 2299–2310.

Brattico, E., Tervaniemi, M., & Picton, T. W. (2003). Effects of brief discrimination-training on the auditory N1 wave. *Neuroreport*, 14(18), 1–4. <https://doi.org/10.1097/01.wnr.0000098748.87269.a1>

Braun, B., Lemhöfer, K., & Mani, N. (2011). Perceiving unstressed vowels in foreign-accented English. *The Journal of the Acoustical Society of America*, *129*(1), 376–387. <https://doi.org/10.1121/1.3500688>

Carlet, A., 2017. L2 perception and production of English consonants and vowels by Catalan speakers: The effects of attention and training task in a cross-training study. PhD Thesis. University of Barcelona.

Carlet, A., & Cebrian, J., 2015. Identification vs. discrimination training: Learning effects for trained and untrained sounds. *In The Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK: the University of Glasgow.*

Catford, J. C., & Pisoni, D. B., 1970. Auditory vs. articulatory training in exotic sounds. *The Modern Language Journal*, *54*(7), 477–481.

Cebrian, J., 2006. Experience and the use of non-native duration in L2 vowel categorization. *Journal of Phonetics*, *34*(3), 372–387. <https://doi.org/10.1016/j.wocn.2005.08.003>

Chandrasekaran, B., Krishnan, A., & Gandour, J. T., 2009. Sensory processing of linguistic pitch as reflected by the mismatch negativity. *Ear And Hearing*, 30(5), 552–558.

Cho, T., & McQueen, J. M., 2005. Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics*, *33*(2), 121–157. <https://doi.org/10.1016/j.wocn.2005.01.001>

Dehaene-Lambertz, G., 1997. Electrophysiological correlates of categorical phoneme perception in adults. *Neuroreport*, 8(4), 919–924.

Dowd, A., Smith, J., & Wolfe, J., 1998. Learning to Pronounce Vowel Sounds in a Foreign Language using Acoustic Measurements of the Vocal Tract as Feedback in Real Time. *Language and Speech*, *41*(1), 1–20.

Flege, J. E., 1987. The production of " new " and " similar " phones in a foreign language : evidence for the effect of equivalence classification. *Journal of Phonetics*, *15*(1), 47–65.

Flege, J. E., 1988. The production and perception of foreign language speech sounds. *Human Communication and Its Disorders: A Review – 1988*, 224–401.

Flege, J. E., 1995a. Second-language Speech Learning: Theory, Findings, and Problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 233–276). York Press, Baltimore.

Flege, J. E., 1995b. Two procedures for training a novel second language phonetic contrast. *Applied Psycholinguistics*, *16*, 425–442.

Flege, J. E., Aoyama, K., & Bohn, O.-S., 2021. The Revised Speech Learning Model (SLM-r) Applied. In *Second Language Speech Learning*. <https://doi.org/10.1017/9781108886901.003>

Flege, J. E., Bohn, O.-S., & Jang, S., 1997. Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, *25*(4), 437–470. <https://doi.org/http://dx.doi.org/10.1006/jpho.1997.0052>

Flege, J. E., & Eefting, W., 1987. Production and perception of English stops by native Spanish speakers. *Journal of Phonetics*, *15*, 67–83.

Fougeron, C., & Keating, P. A., 1997. Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, *101*(6), 3728–3740. <https://doi.org/10.1121/1.418332>

Guenther, F. H., & Bohland, J. W., 2002. Learning Sound Categories: A Neural Model and Supporting Experiments. *Acoustical Science and Technology Journal of the Acoustical Society of Japan*, *23*(587), 213–221. <https://doi.org/10.1250/ast.23.213>

Guenther, F. H., & Hickok, G., 2015. Role of the auditory system in speech production. In G. Celesia & G. Hickok (Eds.), *Handbook of Clinical Neurology* (1st ed., Vol. 129, pp. 161–175). Elsevier B.V. <https://doi.org/10.1016/B978-0-444-62630-1.00009-3>

Hallé, P. A., Chang, Y. C., & Best, C. T., 2004. Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics*, *32*(3), 395–421. <https://doi.org/10.1016/S0095-4470(03)00016-0>

Hayes-Harb, R., 2005. Optimal L2 speech perception: Native speakers of English and Japanese consonant length contrasts. *Journal of Language and Linguistics*, *04*(01).

Hayes, R. L., 2002. The perception of novel phoneme contrasts in a second language: A developmental study of native speakers of English learning Japanese singleton and geminate consonant contrasts. *Coyote Papers*, *12*, 28–41.

Hirata, Y., 2004. Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts. *The Journal of the Acoustical Society of America*, *116*, 2384–2394. <https://doi.org/10.1121/1.1783351>

Hirata, Y., Whitehurst, E., & Cullings, E., 2007. Training native English speakers to identify Japanese vowel length contrast with sentences at varied speaking rates. *The Journal of the Acoustical Society of America*, *121*(6), 3837–3845. <https://doi.org/10.1121/1.2734401>

Iivonen, A., & Harnud, H., 2005. Acoustical comparison of the monophthong systems in Finnish, Mongolian and Udmurt. *Journal of the International Phonetic Association*, *35*(1), 59–71. <https://doi.org/10.1017/S002510030500191X>

Isbell, D., 2016. The Perception - Production Link in L2 Phonology. *MSU Working Papers in Second Language Studies*, *7* (October 2016), 57–67.

Isei-Jaakkola, T., 2004. *Lexical quantity in Japanese and Finnish*. PhD Thesis. University of Helsinki.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C., 2003. A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57. <https://doi.org/10.1016/S0010-0277(02)00198-1>

Iverson, P., Pinet, M., & Evans, B. G., (2011. Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, *33*(01), 145–160. <https://doi.org/10.1017/S0142716411000300>

Jähi, K., Peltola, M. S., & Alku, P., 2015. Does interest in language learning affect the non-native phoneme production in elderly learners? *In The Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK: the University of Glasgow.*

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N., 2015. The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *The Journal of the Acoustical Society of America*, *138*(2), 817–832. <https://doi.org/10.1121/1.4926561>

Kingston, J., 2003. Learning foreign vowels. *Language and Speech*, *46*(2–3), 295–349. <https://doi.org/10.1177/00238309030460020201>

Kondaurova, M. V., 2008. Training to ignore vs. training to attend: The distribution of selective attention in the acquisition of a foreign phonetic contrast. *Proceedings from the Annual Meeting of the Chicago Linguistic Society 44*, Number 1, 169-177.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P., 2006. Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*(2). <https://doi.org/10.1111/j.1467-7687.2006.00468.x>

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B., 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*, 606–608.

Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., & Molholt, G., 2005. The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics*, *26*(2), 227–247.

Lehtonen, J., 1970. *Aspects of quantity in standard Finnish*. University of Jyväskylä.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C., 1957. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368. <https://doi.org/10.1037/h0044417>

Liégeois-Chauvel, C., De Graaf, J. B., Laguitton, V., & Chauvel, P., 1999. Specialization of left auditory cortex for speech perception in man depends on temporal coding. *Cerebral Cortex*, *9*(5), 484–496. <https://doi.org/10.1093/cercor/9.5.484>

Lively, S. E., Logan, J. S., & Pisoni, D. B., 1993. Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3), 1242–1255.

Lively, S. E., Pisoni, D. B., Akahane-Yamada, R., Tohkura, Y., & Yamada, T., 1994. Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *The Journal of the Acoustical Society of America*, *96*(4), 2076–2087. <https://doi.org/10.1121/1.410149>

Logan, J. S., Lively, S. E., & Pisoni, D. B., 1991. Training Japanese listeners to identify English /r/ and /l/: a first report. *The Journal of the Acoustical Society of America*, Vol. 89, pp. 874–886. <https://doi.org/10.1121/1.1894649>

Lopez-Soto, T., & Kewley-Port, D., 2009. Relation of perception training to production of codas in English as a second language. *Proceedings of Meetings on Acoustics*, *6*(2009). <https://doi.org/10.1121/1.3262006>

MacMillan, N. A., & Creelman, C. D., 2005. *Detection theory: A user's guide* (2nd ed.; C. D. Creelman, Ed.) [Book]. Mahwah, N.J: Taylor & Francis Group.

Maddieson, I., & Disner, S. F., 1984. *Patterns of sounds*. London: Cambridge University Press.

May, L., Byers-Heinlein, K., Gervain, J., & Werker, J. F., 2011. Language and the newborn brain: Does prenatal language experience shape the neonate neural response to speech? *Frontiers in Psychology*, *2*(SEP), 1–9. <https://doi.org/10.3389/fpsyg.2011.00222>

McAllister, R., Flege, J. E., & 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*(2), 229–258. <https://doi.org/10.1006/jpho.2002.0174>

Meister, E., Nemoto, R., & Meister, L., 2015. Production of Estonian quantity contrasts by Japanese speakers. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *6*(3), 330–334. International Speech and Communication Association. <https://doi.org/10.12697/jeful.2015.6.3.03>

Menning, H., Imaizumi, S., Zwitserlood, P., & Pantev, C., 2002. Plasticity of the human auditory cortex induced by discrimination learning of non-native, mora-timed contrasts of the Japanese language. *Learning and Memory*, *9*(5), 253–267. <https://doi.org/10.1101/lm.49402>

Motohashi-Saigo, M., & Hardison, D. M., 2009. Acquisition of L2 Japanese Geminates: Training with Waveform Displays. *Language Learning & Technology*, *13*(2), 29–47. <https://doi.org/10.1007/s00167-015-3787-1>

Näätänen, R. (1992). *Attention and brain function*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Näätänen, R., Kujala, T., & Light, G., 2019. *Mismatch Negativity: A Window to the Brain*. Oxford: Oxford University Press.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., et al., 1997. Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, *385*(6615), 432–434. <https://doi.org/10.1038/385432a0>

Näätänen, R., & Picton, T. W., 1987. The N1 wave of the human electric and magnetic response to sound: a review and an analysis of component structure. *Psychological Science*, 24(4), 375–425.

Nozawa, T., 2015. Effects of Attention and Training Method on the Identification of American English Vowels and Coda Nasals. *In The Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK: the University of Glasgow.*

O'Dell, M., 2003. Intrinsic timing and quantity in Finnish. PhD Thesis. University of Tampere.

Okuno, T., 2014. Acquisition of L2 Vowel Duration in Japanese by Native English Speakers. PhD Thesis. Michigan State University.

Okuno, T., & Hardison, D. M., 2016. Perception-production link in L2 Japanese vowel duration: Training with technology. *Language Learning and Technology*, *20*(2), 61–80.

Oldfield, R. C., 1971. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

Pederson, E., & Guion-Anderson, S., 2010. Orienting attention during phonetic training facilitates learning. *The Journal of the Acoustical Society of America*, *127*(2), 54–59. <https://doi.org/10.1121/1.3292286>

Peltola, K. U., Alku, P., & Peltola, M. S., 2017. Non-native speech sound processing changes even with passive listening training. *Linguistica Lettica*, *25*, 158–172.

Peltola, K. U., Tamminen, H., Alku, P., Kujala, T., & Peltola, M. S., 2020. Motoric Training Alters Speech Sound Perception and Production — Active Listening Training Does Not Lead into

Learning Outcomes. *Journal of Language Teaching and Research*, *11*(1), 10. <https://doi.org/10.17507/jltr.1101.02>

Peltola, K. U., Tamminen, H., Alku, P., & Peltola, M. S., 2015. Non-native production training with an acoustic model and orthographic or transcription cues. *In The Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK: the University of Glasgow*.

Perkell, J. S., 2012. Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of Neurolinguistics*, *25*(5), 382–407. <https://doi.org/10.1016/j.jneuroling.2010.02.011>

Perkell, J. S., Matthies, M., Lane, H., Guenther, F. H., Wilhelms-Tricarico, R., Wozniak, J., & Guiod, P., 1997. Speech motor control: Acoustic goals, saturation feedback and internal model effects, auditory. *SPEECH CoriflMma-Speech Communication*, *22*, 227–250.

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., Taylor, M. J., 2000. Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology*, *37*, 127–152.

Pisoni, D. B., Aslin, R. N., Percy, A. J., & Hennessy, B. L., 1982. Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(2), 297–314. <https://doi.org/10.3851/IMP2701.Changes>

Polich, J., & Hoffman, L. D., 1998. P300 and handedness: On the possible contribution of corpus callosal size to ERPs. *Psychophysiology*, *35*(5), 497–507. <https://doi.org/10.1017/S0048577298970792>

Quené, H., 1992. Durational cues for word segmentation in Dutch. *Journal of Phonetics*, *20*, 331–350.

Rato, A., & Rauber, A. S., 2015. The Effects of Perceptual Training on the Production of English vowel contrasts by Portuguese Learners. *In The Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK: the University of Glasgow.*

Rogers, T. T., Rakison, D. H., & McClelland, J. L., 2004. U-Shaped Curves in Development: A PDP Approach. *Journal of Cognition and Development*, *5*(1), 137–145. <https://doi.org/10.1207/s15327647jcd0501_14>

Strange, W., & Dittmann, S., 1984. Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics*, *36*(2), 131–145. <https://doi.org/10.3758/BF03202673>

Suomi, K., Toivanen, J., & Ylitalo, R., 2008. *Finnish Sound Structure. Phonetics, phonology, phonotactics and prosody*. Oulu: Oulu University Press.

Taimi, L., Alku, P., Kujala, T., Näätänen, R., & Peltola, M. S., 2014. The effect of production training on non-native speech sound perception and discrimination in school-aged children: an MMN and behavioural study. *Linguistica Lettica*, *22*, 114–129.

Taimi, L., Jähi, K., Alku, P., & Peltola, M. S., 2014. Children Learning a Non-native Vowel – The Effect of a Two-day Production Training. *Journal of Language Teaching and Research*, *5*(6), 1229–1235. <https://doi.org/10.4304/jltr.5.6.1229-1235>

Tajima, K., Kato, H., Rothwell, A., Akahane-Yamada, R., & Munhall, K. G., 2008. Training English listeners to perceive phonemic length contrasts in Japanese. *The Journal of the Acoustical Society of America*, *123*(1), 397–413. <https://doi.org/10.1121/1.2804942>

Tamminen, H., & Peltola, M. S., 2015. Non-native memory traces can be further strengthened by short term phonetic training. *In The Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK: the University of Glasgow.*

Tamminen, H., Peltola, M. S., Kujala, T., & Näätänen, R., 2015. Phonetic training and non-native speech perception — New memory traces evolve in just three days as indexed by the mismatch negativity (MMN) and behavioural measures. *International Journal of Psychophysiology*, *97*(1), 23–29. <https://doi.org/http://dx.doi.org/10.1016/j.ijpsycho.2015.04.020>

Tremblay, K., Kraus, N., McGee, T., Ponton, C., & Otis, B., 2001. Central auditory plasticity: changes in the N1-P2 complex after speech-sound training. *Ear And Hearing*, 22(2), 79–90. <https://doi.org/10.1097/00003446-200104000-00001>

Tremblay, K., Kraus, N., & McGee, T., 1998. The time course of auditory perceptual learning: neurophysiological changes during speech-sound training. *Neuroreport*, *9*(16), 3557–3560.

Tremblay, K., Kraus, N., Carrell, T. D., & McGee, T., 1997. Central auditory system plasticity: Generalization to novel stimuli following listening training. *Journal of the Acoustical Society of America*, 102(6), 3762–3773. <https://doi.org/10.1121/1.420139>

Vannasing, P., Florea, O., González-Frankenberger, B., Tremblay, J., Paquette, N., Safi, D., Gallagher, A., 2016. Distinct hemispheric specializations for native and non-native languages in one-day-old newborns identified by fNIRS. *Neuropsychologia*, *84*, 63–69. <https://doi.org/10.1016/j.neuropsychologia.2016.01.038>

Wang, X., & Munro, M. J., 1999. The perception of English tense-lax vowel by native Mandarin speakers: The effect of training on attention to temporal and spectral cues. *Paper Presented at the 14th International Congress of Phonetic Sciences*, 125–128.

Wang, Y., Jongman, A., & Sereno, J. A., 2003. Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, *113*(2), 1033. <https://doi.org/10.1121/1.1531176>

Werker, J. F., & Tees, R. C., 1984. Cross-language speech perception: evidence for perceptual reorganisation during the first year of life. *Infant Behavior and Development*, *7*, 49–63. <https://doi.org/10.1016/S0163-6383(02)00113-3>

White, L., & Turk, A. E., 2010. English words on the Procrustean bed: Polysyllabic shortening reconsidered. *Journal of Phonetics*, *38*(3), 459–471. <https://doi.org/10.1016/j.wocn.2010.05.002>

Wiik, K. (1965). *Finnish and English Vowels*. PhD Thesis. University of Turku.

Ylinen, S., Huotilainen, M., & Näätänen, R., 2005. Phoneme quality and quantity are processed independently in the human brain. *Neuroreport*, *16*(16), 1857–1860.

Ylinen, S., Shestakova, A., Alku, P., & Huotilainen, M., 2005. The Perception of Phonological Quantity Based on Durational Cues by Native Speakers, Second-Language Users and Nonspeakers of Finnish. *Language and Speech*, *48*(3), 313–338. <https://doi.org/10.1177/00238309050480030401>

Ylinen, S., Shestakova, A., Huotilainen, M., Alku, P., & Näätänen, R., 2006. Mismatch negativity (MMN) elicited by changes in phoneme length: A cross-linguistic study. *Brain Research*, *1072*(1), 175–185. <https://doi.org/http://dx.doi.org/10.1016/j.brainres.2005.12.004>

Ylinen, S., Uther, M., Latvala, A., Vepsäläinen, S., Iverson, P., Akahane-Yamada, R., & Näätänen, R., 2010. Training the Brain to Weight Speech Cues Differently: A Study of Finnish Second-language Users of English. *Journal of Cognitive Neuroscience*, *22*(6), 1319. <https://doi.org/10.1162/jocn.2009.21272>

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2022

ISBN 978-951-29-8973-7 (PRINT)
ISBN 978-951-29-8974-4 (PDF)
ISSN 0082-6987 (Print)
ISSN 2343-3191 (Online)